Bayesian Approaches to Shrinkage and Sparse Estimation

A guide for applied econometricians

Dimitris Korobilis
University of Glasgow

Kenichi Shimizu
University of Glasgow

November 25, 2021

Abstract

In all areas of human knowledge, datasets are increasing in both size and complexity, creating the need for richer statistical models. This trend is also true for economic data, where high-dimensional and nonlinear/noparametric inference is the norm in several fields of applied econometric work. The purpose of this paper is to introduce the reader to the realm of Bayesian model determination, by surveying modern shrinkage and variable selection algorithms and methodologies. Bayesian inference is a natural probabilistic framework for quantifying uncertainty and learning about model parameters, and this feature is particularly important for inference in modern models of high dimensions and increased complexity.

We begin with a linear regression setting in order to introduce various classes of priors that lead to shrinkage/sparse estimators of comparable value to popular penalized likelihood estimators (e.g. ridge, lasso). We explore various methods of exact and approximate inference, and discuss their pros and cons. Finally, we explore how priors developed for the simple regression setting can be extended in a straightforward way to various classes of interesting econometric models. In particular, the following case-studies are considered, that demonstrate application of Bayesian shrinkage and variable selection strategies to popular econometric contexts: i) vector autoregressive models; ii) factor models; iii) time-varying parameter regressions; iv) confounder selection in treatment effects models; and v) quantile regression models. A MATLAB package and an accompanying technical manual allow the reader to replicate many of the algorithms described in this review.

Contents

1 Introduction

In all areas of human knowledge, datasets are increasing in both size and complexity, creating the need for richer models. This trend is also true for economic data, where high-dimensional and nonlinear/noparametric inference is the norm in several fields of applied econometric work. The purpose of this survey is to introduce the reader to Bayesian inference using shrinkage and variable selection priors. In particular we intend to demonstrate that the benefits of a Bayesian approach to high-dimensional estimation are manifold. Bayesian inference allows for a more accurate quantification of uncertainty. Parameters are treated as random variables that have their own probability density (or mass) functions. The use of a prior distribution provides a natural ground for enhancing possibly weak information in the likelihood.[1] Our first aim is to explore in this review classes of priors that can recover popular penalized regression estimators, such as the lasso of Tibshirani (1996). Next, we want to demonstrate how the Bayesian paradigm becomes a natural framework for combining prior forms in order to capture more complicated patterns of shrinkage and/or sparsity in the data. For example, Ročková and George (2018) extend the lasso with ideas from the Bayesian variable selection literature in order to obtain a "spike and slab lasso" estimator that is empirically superior to shrinkage or variable selection alone, and has desirable theoretical guarantees. Finally, we aim to illustrate that the Bayesian framework is ideal for applied economists who want to use shrinkage or sparsity in more complex or unconventional settings. Economists might be interested in combining data-rigorous statistical variable selection with economic restrictions on certain parameters[2], or use a shrinkage estimator in a model with breaks, stochastic volatility, missing data or other complexities. Penalized and constrained maximum likelihood frameworks can deal with such cases, but computation is non-trivial because it relies on optimizing complex functions. We demonstrate emphatically in this survey paper that Bayesian computation provides numerous tools and algorithms for shrinkage and sparsity that can be incorporated in very complex statistical models with the same ease they are used in univariate linear regression settings.

Even though the notions of sparsity and shrinkage estimation are ubiquitous since the explosion of Big Data in all fields of science (e.g. we doubt there are many economists these days who haven't heard about the lasso), we want to clarify these terms before proceeding with our formal definitions. Sparsity refers to finding parameter estimates that have more zeros than not (where zeros in estimation means absence of some effect or relationship). Shrinkage means estimation where many parameter elements are suppressed towards zero, but they are not

[1] Note that our interest here is in "wide" data (e.g. a linear regression model with more predictors than observations) where unrestricted estimation based only on the likelihood is either unreliable or impossible. In cases with "tall" data (many observations) the Bayesian posterior will tend to concentrate towards a point mass, i.e. uncertainty is small.

[2] For example, instead of the typical statistical shrinkage towards zero that indicates whether an effect is important or not, economists might want to shrink a parameter towards a calibrated value or a sign restriction provided by the solution of an economic model.

necessarily zero. While many readers might be familiar with these concepts, interpretation from a Bayesian point of view is slightly different from frequentist approaches. Sparsity is not identical for the simple reason that parameters in the Bayesian paradigm are (continuous, in many cases) random variables. Similarly, shrinkage estimation is embedded in Bayesian inference since any non-diffusing (non-flat) prior will tend to bias the likelihood; the frequentist statistician can only achieve shrinkage if they specify the estimation problem using an explicit penalized likelihood approach.

We explain these differences, and many more concepts, in this detailed review. We build our discussion gradually by introducing in this section basic components of Bayesian decision theory and estimation, and the principles of Bayesian model determination using the marginal likelihood. In Section 2 we introduce the concept of hierarchical priors and present the basic properties of a large class of hierarchical representations of Bayesian sparsity and shrinkage estimators. In Section 3 we focus on computation using hierarchical priors, and strategies for making inference in high-dimension computationally feasible. Section 4 demonstrates how the hierarchical priors and computational tools discussed in the previous sections, can be readily applied to a wide class of models that are important in economics and finance, as well as other fields of science. Section 5 concludes this review.

Throughout this review we make the assumption that the reader has a broad understanding of the concept of a prior distribution. If this is not the case, novice readers are advised to begin reading about the basics of Bayesian inference in subsection 1.2 and then move to subsection 1.1. More experienced readers, can move directly to section 2, skipping the material in this section.

1.1 Bayesian decision theory and estimation

In order to motivate shrinkage and sparsity, we first introduce the concept of loss-based estimation using a Bayesian decision theoretic approach. Detailed introductions can be found in Fourdrinier et al. (2018) and Robert (2007). Assume we have data $X \in \mathfrak{X}$ where $\mathfrak{X}$ (the sample space) is a measurable set of $\mathbb{R}^n$, and parameters $\theta \in \Theta$ where Θ (the parameter space) is a measurable set of $\mathbb{R}^p$. We define two probability density functions (p.d.f.) that are measurable on $\mathfrak{X}$ and Θ: a the likelihood function $p(X|\theta)$, and a prior function $\pi(\theta)$. Denote with $\widehat{\theta}(X)$ an estimator of θ, that is, a measurable function of data X that maps from $\mathbb{R}^n$ to $\mathbb{R}^p$.

Under these definitions we can now specify what is the loss and risk associated with the estimator $\widehat{\theta}(X)$. First, we can define loss functions of the form $L\left(\widehat{\theta}(X), \theta\right) = \rho\left(\widehat{\theta}(X), \theta\right)$ where $\rho(\bullet)$ can be a symmetric loss function (the quadratic being the most popular) or any asymmetric loss function that measures how close $\widehat{\theta}(X)$ is to the true θ. The Bayes risk associated with "decision" $\widehat{\theta}$ is defined as (see also Fourdrinier et al., 2018)

$$r\left(\pi, \widehat{\theta}\right) = \int_{\Theta} E_{\theta}\left(L\left(\widehat{\theta}(X), \theta\right)\right) d\pi(\theta). \tag{1}$$

The quantity $\mathscr{R}(\theta, \hat{\theta}) = E_\theta\left(L\left(\widehat{\theta}(X), \theta\right)\right)$ is the frequentist risk of $\widehat{\theta}$, which is defined as the expected value of the loss function over the data realization for a fixed θ. In contrast, the Bayes risk in Equation 1 is the average of frequentist risk $\mathscr{R}$ with respect to the prior distribution $\pi(\theta)$. Frequentist decision theory aims at making the expected loss $\mathscr{R}(\theta, \hat{\theta})$ small, while Bayesian decision theory aims at finding the minimum of $r\left(\pi, \widehat{\theta}\right)$. In particular, the quantity

$$r(\pi) = \inf_{\widehat{\theta}} r\left(\pi, \widehat{\theta}\right), \tag{2}$$

is the Bayes risk of the prior distribution π. Given a prior π, an associated Bayes estimator $\hat{\theta}_\pi$ is a minimizer in the sense that $r(\pi, \hat{\theta}_\pi) = r(\pi)$.

We can now define the concepts of minimaxity and admissibility. A decision rule (estimator) is *admissible* with respect to the loss function L if and only if no other rule dominates it. That is, iff $r\left(\pi, \widetilde{\theta}\right) < r\left(\pi, \widehat{\theta}\right)$ then $\widetilde{\theta}$ is admissible. An estimator is $\hat{\theta}_0$ is *minimax* for a given loss function L if

$$\sup_\theta \mathscr{R}(\theta, \hat{\theta}_0) = \inf_{\hat{\theta}} \sup_\theta \mathscr{R}(\theta, \hat{\theta}); \tag{3}$$

that is, it is the minimizer of the worst-case frequentist risk. For a given prior π, define an associated Bayes estimator $\hat{\theta}_\pi$. If $\sup_\theta \mathscr{R}(\theta, \hat{\theta}_\pi) = r(\pi, \hat{\theta}_\pi)$, then $\hat{\theta}_\pi$ can be shown to be minimax. In this case, the prior π is least favorable in the sense that $r(\pi', \hat{\theta}_\pi) \leq r(\pi, \hat{\theta}_\pi)$ for all other priors π'. That is, $\hat{\theta}_\pi$ is the best with respect to the least favorable prior distribution $\pi(\theta)$. Minimaxity is a desirable feature for comparing estimators but, of course, it can still become a misleading measure of comparison; see a counterexample and further discussion in Robert (2007). Finally, note that if a minimax estimator is a unique (Bayes) estimator, then this is also admissible.

Why is it important to think in terms of optimality of an estimator with respect to a loss function? To answer this question, consider the expected value of the squared error loss of a *scalar, point* estimator $\widehat{\theta} = \widehat{\theta}(X)$, which is also known as the mean squared error:

$$\begin{aligned} MSE(\widehat{\theta}) &= E\left[L\left(\widehat{\theta}, \theta\right)\right] = E\left[\left(\widehat{\theta} - \theta\right)^2\right], & (4)\\ &= E\left[\left(\widehat{\theta} - E\left\{\widehat{\theta}\right\} + E\left\{\widehat{\theta}\right\} - \theta\right)^2\right], & (5)\\ &= E\left[\left(\widehat{\theta} - E\left\{\widehat{\theta}\right\}\right)^2\right] + \left(E\left\{\widehat{\theta}\right\} - \theta\right)^2. & (6) \end{aligned}$$

The first term in the last equation above is the variance of $\widehat{\theta}$, and the second term is the square of its bias. The least squares estimator, which in many simple linear settings coincides with the maximum likelihood estimator, has zero bias (unbiased) and is the "best" meaning that it has narrowest sampling distribution (minimum variance) among all unbiased estimators. Despite these two desirable properties, it is not necessarily the case that OLS will always have the lowest mean squared error. Indeed, in high-dimensional cases with fat data (p large relative to n) the

sample variance of the OLS will tend to become very large. In cases with more parameters than observations ($p > n$), the OLS estimator has infinite solutions and infinite variance. In such cases, there exist biased estimators that achieve much lower variance compared to the unbiased estimator, to the extend that this reduction in variance compensates for any increase in the square of the bias (making the total MSE of the biased estimator lower). Specifically in the case of out-of-sample prediction the MSE of our modeled variable will be larger if the estimation MSE in Equation 6 is high, showing that evaluating estimation loss might be more important than looking only at (minimum variance) unbiasedness.

A well-known illustration of this concept, that changed dramatically the way statisticians think about estimators, is the example of the James-Stein estimator. Assume our likelihood is $X \sim N_p\left(\theta, \underline{\sigma}^2 I_p\right)$ where $\theta \in \mathbb{R}^p$ is the unknown parameter and $\underline{\sigma}^2$ is assumed to be known. Stein (1956) proved that the maximum likelihood estimator $\widehat{\theta}^{mle} = X$ is the minimum risk equivariant estimator under various loss functions, it is minimax, and it is admissible for $p = 1, 2$. However, for $p \geq 3$ the maximum likelihood estimator is inadmissible under a square loss function, and the James-Stein estimator

$$\widehat{\theta}^{JS} = \left(1 - \frac{(p-2)\underline{\sigma}^2}{\sum_{i=1}^{n} X_i}\right) X, \tag{7}$$

has lower risk than the MLE, that is, $\mathscr{R}(\widehat{\theta}^{JS}) < \mathscr{R}(\widehat{\theta}^{mle})$. Efron and Morris (1973) showed that the James-Stein estimator is a special case of an empirical Bayes estimator of θ, that is, an estimator that places a Gaussian prior on θ and sets its prior variance to be a certain function of the data X. Stein's estimator minimizes the *total* quadratic risk of θ, but there may be elements $\widehat{\theta}_i^{JS}$, $i \in [1, p]$, which have higher risk than the MLE. For that reason, Efron and Morris (1973) also propose a *limited translation empirical Bayes estimator*, which offers a compromise between Stein's estimator and the MLE.

Bayesian estimators are by default biased towards the prior expectation, which is a result of doing inference by using the information in both the likelihood and prior functions. Similarly, penalized likelihood estimators, such as the popular lasso of Tibshirani (1996), constrain the likelihood function with a penalty that intends to introduce a similar bias. The purpose of this subsection is to introduce an alternative view to traditional econometric inference with small parameter space, where unbiasedness is the holy grail. In high-dimensional settings some estimation bias may be desirable, especially when the purpose is prediction in which case richly parameterized specifications are not welcome. In many instances, in-sample parameter estimation accuracy (instead of out-of-sample prediction) is of primary importance, for example, when the quantity of interest is an elasticity or a causal effect that can inform policy decisions. We show later in this survey that even in such cases Bayesian and frequentist penalized regression estimators can be desirable.

1.2 Principles of Bayesian Model Choice: A regression perspective

According to Gelman et al. (2013) the process of Bayesian data analysis involves three steps

1. Setting up a full probability model. This doesn't only involve specifying a likelihood for our data (observables), but we need to specify a joint distribution for both observables and unobservables (parameters, or other unobserved data/variables)

2. Conditioning on the observed data in order to calculate posterior probabilities of all unobservables

3. Assessing model fit, for example, understanding limitations of the chosen likelihood and prior for recovering interpretable and useful parameters estimates, and addressing sensitivity of the results to these choices

In the first part of this review, we use a simple linear regression setting as the basis for developing shrinkage and sparsity priors (step 1), for discussing posterior computation (step 2) and assessing model fit (step 3). By doing so we aim to offer the same level playing field for presenting various hierarchical prior formulations. The final section presents several extensions of shrinkage and sparsity priors in more complex settings, such as factor models, time-varying parameter regression, and cofounder selection in treatment effect estimation.

The regression model we build upon has the form

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, ..., n, \tag{8}$$

where n is the number of observations, y_i is a scalar dependent variable, $\boldsymbol{X}_i$ is a $1 \times p$ vector of covariates (or *regressors* or *predictors*) that can possibly include an intercept, dummies, exogenous variables or other effects (e.g. trend in a time-series setting), $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\varepsilon_i \sim N(0, \sigma^2)$ is a Gaussian disturbance term with zero mean and scalar variance parameter σ^2. Within this setting our interest lies in obtaining "good" estimates of $\boldsymbol{\beta}$ and σ^2, specifically in settings with many covariates ("large p, small n" regression).

The linear regression formulation implies a certain Gaussian likelihood function $\mathscr{L}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X})$ that is proportional to the sampling density $p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$. These two quantities are not identical because the likelihood is not a true density function.[3] The Bayesian needs to specify a joint prior distribution of the parameters, in the form $p(\boldsymbol{\beta}, \sigma^2)$. Bayes Theorem postulates that

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y})}, \tag{9}$$

but for the purpose of parameter estimation, in particular, it is easier to ignore $p(\boldsymbol{y})$ since it is a normalizing constant (i.e. not a function of the parameters of interest $\boldsymbol{\beta}$, σ^2) and work instead with the formula

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2). \tag{10}$$

[3] The likelihood is a product of densities that lacks a normalizing constant.

A default prior setting in Bayesian inference is the natural conjugate prior which is defined as

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &= p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) && (11)\\ &= N(\mathbf{0}, \sigma^2\boldsymbol{D}) \times Inv-Gamma\left(\frac{v_0}{2}, \frac{s_0^2}{2}\right), && (12)\\ &\propto \left(\sigma^2\right)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}\right\} && (13)\\ &\quad \times \left(\sigma^2\right)^{-v_0/2-1} \exp\left\{-\frac{s_0^2/2}{\sigma^2}\right\}, && (14) \end{aligned}$$

where $(\boldsymbol{D}, v_0, s_0)$ are prior hyperparameters chosen by the researcher. Due to the fact that the likelihood has a similar structure to this prior, it is trivial to prove (see the accompanying Technical Document) that the posterior is of the form

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = N\left(\boldsymbol{V}\left(\boldsymbol{X}'\boldsymbol{y}\right), \sigma^2\boldsymbol{V}\right) \times Inv-Gamma\left(\frac{v}{2}, \frac{s^2}{2}\right), \quad (15)$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1}\right)^{-1}$, $v = v_0 + n + p$, $s^2 = s_0^2 + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}$
$\boldsymbol{X} = [\boldsymbol{X}_1', ..., \boldsymbol{X}_n']'$ and $\boldsymbol{y} = (y_1, ..., y_n)'$.

1.2.1 Goodness of fit measures: Marginal likelihood and information criteria

While Equation 10 is required for the derivation of parameter posterior distributions, the quantity $p(\boldsymbol{y})$ in Equation 9 is of paramount importance for Bayesian model determination. This is the *prior predictive likelihood*, more commonly known as the *marginal likelihood*, that is, the evidence in data $\boldsymbol{y}$ after we integrate out the effect of all possible values that the "random variables" $\boldsymbol{\beta}, \sigma^2$ can admit through their prior distribution. This can be proven via solving for $p(\boldsymbol{y})$ in Equation 9:

$$\begin{aligned} p(\boldsymbol{y})p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) &= p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2) && (16)\\ \Rightarrow \int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{y})p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})d\boldsymbol{\beta}d\sigma^2 &= \int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)d\boldsymbol{\beta}d\sigma^2 && (17)\\ \Rightarrow p(\boldsymbol{y})\int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{\beta}, \sigma^2|y)d\boldsymbol{\beta}d\sigma^2 &= \int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)d\boldsymbol{\beta}d\sigma^2 && (18)\\ \Rightarrow p(\boldsymbol{y}) &= \int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)d\boldsymbol{\beta}d\sigma^2, && (19) \end{aligned}$$

where $\int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{\beta}, \sigma^2|y)d\boldsymbol{\beta}d\sigma^2 = 1$ because this is a proper density. The marginal likelihood is the expected value of the likelihood where the expectation is taken with respect to the prior. Put differently, it is the prior mean of the likelihood function. An important characteristic of the marginal likelihood is that the integral in Equation 19 can only be calculated when the prior is a proper density, that is, if $p(\boldsymbol{\beta}, \sigma^2)$ integrates to one. The benchmark Uniform (Jeffrey's) prior on $\boldsymbol{\beta}$ and $\log(\sigma^2)$ is a key example where this condition fails and the marginal likelihood does not exist.

Assume we want to predict a new (future) observation y_{n+1} given $\boldsymbol{X}_{n+1}$ using the prediction (out-of-sample) model $p(y_{n+1}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y})$ which, in turn, is based on the in-sample estimated model $p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$. We can then define the *posterior predictive likelihood*

$$p(y_{n+1}|\boldsymbol{y}) = \int_{-\infty}^{\infty} \int_{0}^{\infty} p(y_{n+1}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\boldsymbol{\beta} d\sigma^2, \tag{20}$$

which is the distribution of the out-of-sample data point marginalized over the posterior distribution of the model parameters.

Both quantities – prior and posterior predictive distributions – are fundamental for model assessment in Bayesian inference. In the benchmark case of the linear regression with the natural conjugate prior, the marginal likelihood can be derived analytically and is of the form

$$\begin{aligned} p(\boldsymbol{y}) &= \frac{p(y|\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})} && (21) \\ &= \frac{\Gamma\left(\frac{v_0}{2}\right)^{-1} (s_0/2)^{\frac{v_0}{2}}}{(2\pi)^{\frac{n}{2}} \Gamma\left(\frac{v}{2}\right)^{-1} (s/2)^{\frac{v}{2}}} \frac{|\boldsymbol{D}|^{-\frac{1}{2}}}{|\boldsymbol{V}|^{-\frac{1}{2}}} && (22) \\ &\times \left[\frac{1}{2}\left(s_0 + \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\mu}^* \boldsymbol{V}^{-1} \boldsymbol{\mu}^*\right)\right] && (23) \\ & && (24) \end{aligned}$$

where $v_0, s_0, \boldsymbol{D}$ are parameters of the prior distribution (chosen by the researcher), and $v, s, \boldsymbol{V}$ are parameters of the posterior distribution whose values are provided in Equation 15 and $\boldsymbol{\mu}^* = \boldsymbol{V}(\boldsymbol{X}'\boldsymbol{y})$.

The predictive likelihood is also available analytically and it is of the form

$$y_{n+1}|\boldsymbol{y} \sim t_1\left(y_{n+1}; \boldsymbol{X}_{n+1}\boldsymbol{V}\left(\boldsymbol{X}'\boldsymbol{y}\right), \frac{s}{v}\left(1 + \boldsymbol{X}_{n+1}\boldsymbol{V}\boldsymbol{X}'_{n+1}\right), v\right) \tag{25}$$

where we define the p-dimensional t-density with location $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, and degrees of freedom d as

$$t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, d) = \frac{\Gamma\left(\frac{d+p}{2}\right)}{\Gamma\left(\frac{d}{2}\right) d^{p/2} \pi^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \left[1 + \frac{1}{d}(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right] \tag{26}$$

The marginal likelihood is rarely available analytically, and in most cases the integral in Equation 19 has to be approximated using Monte Carlo or numerical methods.[4] In cases of either a complex model or a complex prior structure, or both, evaluating the marginal likelihood can become challenging, if not impossible. In such cases it might be easier to calculate the posterior predictive likelihood in Equation 20 using a procedure called *leave one out cross-validation* (LOO-CV). This would involve fitting the model in training data and then using a

[4] Two early examples are Gelfand and Dey (1994) and Chib (1995); see also Chib and Jeliazkov (2001) for a review.

hold-out sample to evaluate the posterior predictive likelihood. Notice that if MCMC samples from the parameter posterior are available, evaluation of Equation 20 is straightforward using Monte Carlo integration.[5]

When marginal or posterior predictive likelihoods are difficult to obtain, a (computationally) straightforward alternative strategy is to rely on information criteria. For example, the Bayesian information criterion (BIC), is a first-order approximation to the marginal likelihood. Performing a Taylor expansion around the posterior mode[6] $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)$ for the logarithm of the term $p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{\beta}, \sigma^2\right)$ in Equation 19, we can write the log-marginal likelihood as

$$\begin{aligned} \log p\left(\boldsymbol{y}\right) &= \log p\left(\boldsymbol{y}|\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2\right) + \log p\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2\right) + \tfrac{p}{2}\log(2\pi) \\ &\quad -\tfrac{p}{2}\log n - \tfrac{1}{2}\log\left|J_n\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2\right)\right| + O\left(n^{-1}\right), \end{aligned} \tag{27}$$

where $J_n\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2\right)$ is the expected Fisher information matrix of $p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{\beta}, \sigma^2\right)$ evaluated at the posterior mode $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)$. In large samples, the posterior mode coincides with the MLE $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$. Considering this approximation and removing from Equation 27 any terms of order $O\left(1\right)$ or less, we obtain

$$\log p\left(\boldsymbol{y}\right) = \log p\left(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2\right) - \frac{p}{2}\log n + O\left(1\right). \tag{28}$$

The approximation above provides the basis for defining the Bayesian information criterion

$$BIC = -2\log \mathscr{L}\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2|\boldsymbol{y}, \boldsymbol{X}\right) + p\log n, \tag{29}$$

where $\mathscr{L}\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2|\boldsymbol{y}, \boldsymbol{X}\right)$ is the likelihood function evaluated at the MLE.

The BIC is only a crude approximation to the marginal likelihood and it is based on a point estimate. An alternative popular criterion is the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) which is of the form

$$DIC = -4E_{p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})}\left[\log p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right)\right] + 2\log p\left(\boldsymbol{y}|\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2\right). \tag{30}$$

The first term is the expectation of the data density with respect to the posterior[7] which can be evaluated numerically from the MCMC output by taking the mean of $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$ over all MCMC samples of the parameters. The second term is the value of the data density evaluated at the posterior mode $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)$. For more information on the DIC see also Chan and Grant

[5] Recognizing the numerical and computational shortcomings of model choice based on marginal likelihoods, there are several early studies that propose model choice criteria that are based on variants of the posterior predictive distribution, see Davison (1986), Gelfand and Ghosh (1998), Gelman et al. (1996), Laud and Ibrahim (1995), Ibrahim and Laud (1994) and Martini and Spezzaferri (1984).

[6] The posterior mode is chosen such that the first derivative of the posterior is zero, which simplifies terms when taking the Taylor expansion; see Raftery (1995) for a detailed proof.

[7] For that reason, the DIC is related to the posterior predictive likelihood, i.e. the integral in Equation 20, rather than the marginal likelihood.

(2016), Spiegelhalter et al. (2014) and van der Linde (2005).

Chen and Chen (2008) propose a modification to the Bayesian information criterion for high-dimensional spaces, which they call the extended Bayesian information criterion (EBIC). In the context of a proportional hazards model, Volinsky and Raftery (2000) propose a modification of the BIC penalty term that is consistent with a conjugate unit-information prior under this model. Foster and George (1994) propose the risk inflation criterion (RIC) while George and Foster (2000) present empirical Bayes selection criteria. Watanabe (2010, 2013) derives the widely applicable information criterion (WAIC), also known as the Watanabe-Akaike information criterion since this criterion can be considered to be a Bayesian variant of the popular Akaike information criterion. Gelman et al. (2014) and Vehtari et al. (2017) perform informative comparisons of the properties of BIC, DIC, WAIC and LOO-CV in a Bayesian context.

1.2.2 Testing hypotheses: Bayes factors

Consider now the case of two competing models, model one (denoted as M_1) and model two (denoted as M_2). For example, a key scenario that fits this setting, is that of testing hypotheses of the form $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, for some $j = 1, ..., p$. Evidence in favor of either H_0 or H_1, corresponds to how good is the fit of two corresponding nested regression models (M_1 is unrestricted, and M_2 has the restriction $\beta_j = 0$ imposed). In this setting it is convenient to condition parameter posteriors and marginal likelihoods for each model on the random variable M_i, $i = 1, 2$, that indexes each of the two models. For example, $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, M_1)$ and $p(\boldsymbol{y}|M_1)$ denote the parameter posterior and marginal likelihood, respectively, of regression model 1. Consequently, the quantity

$$BF_{12} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}, \tag{31}$$

is the *Bayes Factor* between models 1 and 2. The quantity

$$PO_{12} \equiv \frac{p\left(M_1|\boldsymbol{y}\right)}{p\left(M_2|\boldsymbol{y}\right)} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)} \times \frac{p\left(M_1\right)}{p\left(M_2\right)} \tag{32}$$

is the *posterior odds* between models 1 and 2. It is defined as the product of the Bayes factor and the prior odds. If we assign equal model probabilities a-priori, then $p\left(M_1\right) = p\left(M_2\right) = \frac{1}{2}$ and the Bayes factor is identical to the posterior odds ratio. The Bayes factor above is a primary tool for assessing evidence in favor of a statistical model versus a competing model.

Kass and Raftery (1995) provide a rule-of-thumb on how to interpret the statistical evidence against model 2 based on ranges of values of BF_{12}: for values higher than three the evidence is substantial, for values higher than 10 it is strong, and for values higher than 100 it is decisive. Given that marginal likelihoods are not available with improper priors (even if the posterior is proper), there has been plenty of interest in calculating Bayes factors when such priors are used. Aitkin (1991) proposes to calculate Bayes factors based on integrating the likelihood

with the posterior – this is equivalent to replacing $p(\boldsymbol{\beta}, \sigma^2)$ with $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$ in Equation 19. This formulation allows to calculate "posterior" Bayes factors regardless of the prior structure of each model, and at the same time it avoids Lindley's paradox (Aitkin, 1991). Berger and Pericchi (1996, 1998) suggest the use of the *intrinsic* Bayes factor. Their suggestion involves splitting the data into n subsets, such that one can obtain the marginal likelihood of the i^{th} subset conditional on all other subsets. Subsequently, either the arithmetic or geometric average of the Bayes factors estimated in all n subsets of the data can be used as the final estimate.

For nested model comparisons, Verdinelli and Wasserman (1995) show that Bayes factors can be calculated using the Savage-Dickey density ratio (SDDR) approach. Consider two regression models as in Equation 8 but for notational simplicity set $p = 1$, that is, only a single covariate is available. The first model, M_1, is an unrestricted model while model M_2 imposes the restriction $\beta = \beta^\star$ for some scalar value $\beta^\star$ (the previous example of testing of $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ fits this setting). In this case the Bayes factor can be written as

$$
\begin{aligned}
BF_{12} &= \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)} && (33)\\
&= \frac{\int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{y}|\beta, \sigma^2, M_1)p(\beta, \sigma^2|M_1)d\beta d\sigma^2}{\int_{0}^{\infty} p(\boldsymbol{y}|\beta^\star, \sigma^2, M_2)p(\beta^\star, \sigma^2|M_2)d\sigma^2} && (34)\\
&= \frac{\int_{0}^{\infty} p\left(\beta^\star, \sigma^2|\boldsymbol{y}, M_2\right) d\sigma^2}{\int_{0}^{\infty} p\left(\beta^\star, \sigma^2|M_2\right) d\sigma^2}, && (35)
\end{aligned}
$$

that is, SSDR is the ratio of the marginal posterior and prior of β under model M_2, evaluated at the point $\beta = \beta^\star$. In general it will be easy to evaluate these two distributions, especially when the Gibbs sampler is used for approximating the posterior distribution. This is because evaluation of the numerator using Monte Carlo integration would be fairly straightforward. Additionally, in the case of an independent prior of the form $p(\beta, \sigma^2) = p(\beta)p(\sigma^2)$ the denominator above becomes $\int_{0}^{\infty} p\left(\beta^\star, \sigma^2|M_2\right) d\sigma^2 = p\left(\beta^\star|M_2\right) \int_{0}^{\infty} p\left(\sigma^2|M_2\right) d\sigma^2 = p\left(\beta^\star|M_2\right)$, i.e. we only need to evaluate the (Gaussian) prior of β at the point $\beta^\star$.

There are of course numerous other ways of obtaining approximations to the Bayes factors that do not explicitly involve calculating ratios of marginal likelihoods. Goutis and Robert (1998) propose an alternative procedure for testing nested models based on the Kullback-Leibler divergence. The idea is to compute the projection of the unrestricted model to the restricted parameter space, and use the corresponding minimum distance to judge whether or not the restricted model is appropriate. The same way we used the BIC to obtain a first-order approximation to the marginal likelihood, we can also use the BIC to obtain approximations to Bayes factors – this approach is illustrated in Raftery (1995). Notable early studies on the topic of Bayes factors include Kass and Wasserman (1995), De Santis and Spezzaferri (1997), O'Hagan (1995), Berger and Pericchi (2001), Berger and Mortera (1999), Lewis and Raftery (1997), Raftery (1996) and DiCiccio et al. (1997). A systematic review of methods for

calculating Bayes factors can be found in Kadane and Lazar (2004).

Finally, it is worth noting that in the case of nested hypothesis testing we can derive an optimal Bayesian point estimate by minimizing expected loss averaged over the two hypotheses, using posterior model probabilities as weights. That is, considering again the simple case with $p = 1$ and ignoring the variance parameter σ^2 for simplicity, we aim to find point estimate $\widehat{\beta}$ such that the joint expected loss under the two models/hypotheses

$$E\left(L\left(\beta,\widehat{\beta}\right)\right) = \left[p(M_1|\boldsymbol{y})E\left(L\left(\beta,\widehat{\beta}\right)|M_1\right) + \tag{36}$$

$$p(M_2|\boldsymbol{y})E\left(L\left(\beta,\widehat{\beta}\right)|M_2\right)\right], \tag{37}$$

achieves a minimum. Under a quadratic loss function $L\left(\beta,\widehat{\beta}\right)$, the posterior means are optimal meaning that the optimal estimator is

$$\widehat{\beta}^{BPE} = p(M_1|\boldsymbol{y})E\left(p\left(\beta|\boldsymbol{y},M_1\right)\right) + p(M_2|\boldsymbol{y})E\left(p\left(\beta|\boldsymbol{y},M_1\right)\right). \tag{38}$$

This estimator can be considered a *Bayesian pre-test estimator*, hence the acronym BPE in the equation above; see Judge et al. (1985) for a detailed discussion. In the next section we will generalize this result to the case of K models, in order to motivate model choice in the presence of many models.

1.2.3 Model choice with many models: Bayesian model averaging

Model choice can have many forms, but the benchmark scenario that will motivate later in this paper to focus on shrinkage and sparse estimation, is that of model determination among many nested models. In particular, consider the problem of deciding which of p variables in the covariate matrix $\boldsymbol{X}$ should be in the "optimal" regression model. Each covariate can have two outcomes, either it is included in a model or it is excluded, meaning that the model space in the presence of p covariates is 2^p. We denote the model set as $\mathscr{M} = \{M_r : r = 1, \ldots, 2^p\}$. The covariates that pertain to model M_r are denoted in this subsection as $\boldsymbol{X}_r$ and their associated coefficients as $\boldsymbol{\beta}_r$. That is, $\boldsymbol{X}_r$ is a matrix that is constructed using only a subset of the columns in $\boldsymbol{X}$. Therefore, we denote regression model M_r as[8]

$$M_r : \boldsymbol{y} = \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}. \tag{39}$$

where $\boldsymbol{X}_r$ is $n \times p_r$ and $\boldsymbol{\beta}_r$ is $p_r \times 1$ with $p_r \in \{1, ..., p\}$. Now with 2^p models, even for small p, pairwise model comparison based on Bayes factors is impractical and alternative computational methods are needed. Most importantly, in the presence of many models the researcher might

[8]For simplicity we do not explicitly allow for an intercept. If an intercept is present in all competing models, then it is important to remove the sample mean from all covariates $\boldsymbol{X}$ (and, as a result, in all subsets $\boldsymbol{X}_r$) in order to ensure that the estimated intercept has exactly the same interpretation in all models. With demeaned covariates and the use of a flat prior, the intercept term becomes identical to the sample mean of $\boldsymbol{y}$ in all 2^p competing models.

not want to give the same weight to each and every model. For example, she might want to give more weight on parsimonious models or models that include a certain predictor suggested by some theory or common sense. For that reason we define prior model probabilities $p(M_r)$ with $\sum_{r=1}^{2^p} p(M_r) = 1$. Based on Bayes theorem, prior model probabilities combined with marginal likelihoods $p(\boldsymbol{y}|M_r)$ give posterior model probabilities

$$p(M_r|\boldsymbol{y}) \propto p(\boldsymbol{y}|M_r)p(M_r). \tag{40}$$

Bayesian model selection (BMS) corresponds to selecting the best model, that is, the model M_r with the highest $p(M_r|\boldsymbol{y})$. *Bayesian model averaging (BMA)* involves averaging over many models using $p(M_r|\boldsymbol{y})$ as weights. That is, for a quantity of interest Δ (e.g. an out-of-sample observation y_{n+1} of $\boldsymbol{y}$) BMA is constructed as the following weighted average

$$p(\Delta|\boldsymbol{y}) = \sum_{r=1}^{2^p} p(\Delta|\boldsymbol{y}, M_r)p(M_r|\boldsymbol{y}). \tag{41}$$

For small model spaces, typically when $p < 30$ posterior model probabilities can be calculated analytically such that we can enumerate and estimate all 2^p available models. For $p > 30$ it is impossible to enumerate and estimate all models in a deterministic way. In such cases, one can rely on Markov chain Monte Carlo algorithms which are able to "visit" in each iteration, in a stochastic way, the most probable models. Hoeting et al. (1999) and Fragoso et al. (2018) provide two systematic reviews on the topic.

While model selection and model averaging with an arbitrary number of models are straightforward extensions of the case with only two models, prior elicitation in multi-parameter and multi-model settings is anything but straightforward. In order to explain the intuition behind why this is the case, consider the natural conjugate prior defined previously, which in the case of model M_r can be written as

$$p(\boldsymbol{\beta}_r, \sigma^2|M_r) = N_{p_r}(\mathbf{0}_{p_r}, \sigma^2\boldsymbol{D}_r) \times Inv - Gamma\left(\frac{v_0}{2}, \frac{s_0^2}{2}\right). \tag{42}$$

Prior elicitation involves choice of $\boldsymbol{D}_r, v_0, s_0$. The hyperparameters v_0, s_0 are scalar in all regression models can be simply set to a small value close to zero, implying a Jeffrey's (diffuse) prior on σ^2. However, $\boldsymbol{D}_r$ is a matrix that changes size based on the number of predictors in model M_r. Assume for simplicity we define $\boldsymbol{D}_r = \tau\boldsymbol{I}_{p_r}$, with $\boldsymbol{I}_{p_r}$ the $p_r \times p_r$ identity matrix. In this case, prior elicitation breaks down to choosing a single hyperparameter τ. We can't use the diffuse choice $\tau \to \infty$ because the marginal likelihood in Equation 24 will become infinite, hence, τ should be finite in the multi-model case. However, using the same finite value of τ in all models, doesn't mean that the effect of this prior is identical (that is, "objective") for each model. Consider for instance two models, one with two predictors $\boldsymbol{X}_2 = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ and a restricted model with only the first predictor $\boldsymbol{X}_1 = \boldsymbol{x}_1$. The posterior

variance is $\boldsymbol{V}_r = \sigma^2 \left(\boldsymbol{X}_r' \boldsymbol{X}_r + (\tau \boldsymbol{I}_{p_r})^{-1} \right)^{-1}$ for each model $r = 1, 2$, so that the impact of τ on the common predictor in the two models will be identical only if $\boldsymbol{x}_1$ is not correlated with $\boldsymbol{x}_2$ and $\boldsymbol{X}_2' \boldsymbol{X}_2$ becomes diagonal. If this is not the case, the correlation between the two predictors will imply that the effect of τ on the regression coefficient of $\boldsymbol{x}_1$ will not be the same in the two models. This issue complicates prior elicitation further when considering $p \gg 2$ correlated covariates, that also potentially have different units of measurement.[9].

For that reason, many researchers have proposed empirical Bayes priors, in the spirit of the empirical Bayes formulation of Stein's estimation rule; see equation Equation 7 and discussion of Efron and Morris (1973). Empirical Bayes procedures allow to choose prior hyperparameters as a function of the data observations, sometimes also chosen to optimize some criterion (e.g. maximum marginal likelihood). A default prior for multi-model settings is the *g-prior* due to Zellner (1986). The g-prior for model M_r takes the form

$$\boldsymbol{\beta}_r | \sigma^2, M_r \sim N_{p_r} \left(\mathbf{0}_{p_r}, \frac{1}{g} \sigma^2 \left(\boldsymbol{X}_r' \boldsymbol{X}_r \right)^{-1} \right), \tag{43}$$

where $\sigma^2 \left(\boldsymbol{X}_r' \boldsymbol{X}_r \right)^{-1}$ is essentially the covariance matrix associated with the OLS estimator $\widehat{\boldsymbol{\beta}}_r$ and g a scalar tuning parameter. Under this prior, the posterior variance of $\boldsymbol{\beta}$ conditional on σ^2 becomes $\boldsymbol{V}_r = \frac{1}{1+g} \times \sigma^2 \left(\boldsymbol{X}_r' \boldsymbol{X}_r \right)^{-1}$, such that the posterior variance is uniformly affected by selection of g. Consequently, the posterior mean/mode is

$$\boldsymbol{\beta}_r^{\star} = \frac{1}{1+g} \widehat{\boldsymbol{\beta}}_r. \tag{44}$$

When $g \to 0$ the posterior mean tends to the OLS estimate of model M_r ($\widehat{\boldsymbol{\beta}}_r$) while when $g \to \infty$ the posterior contracts towards zero. While the effect of the prior now depends in a straightforward, transparent way[10] on a single hyperparameter, choice of this hypeparameter is very important for determining marginal likelihoods and model probabilities.

Fernández et al. (2001a,b) propose default values of g in the context of Bayesian model averaging and Eicher et al. (2011) expand this discussion by considering further values of g. A benchmark suggestion of Fernández et al. (2001b) is to set $g \equiv g_r = p_r / n$, that is, a value of g that is the ratio of the number of coefficients in each model r over the total number of observations. Wide models with many covariates models will have larger g, thus, tending to shrink their posterior towards zero more aggressively. Put differently, the prior variance is getting smaller meaning that the information in the prior increases relative to the information

[9]The scaling issue in $\boldsymbol{X}$ can be dealt with by standardizing the data, that is, dividing each column with its sample standard deviation. High correlation in columns of $\boldsymbol{X}$ can also be dealt with by orthogonalizing this matrix. While standardization is easy to apply and is recommended in all model averaging and variable selection algorithms, orthogonalization of the columns of $\boldsymbol{X}$ is only feasible when $n > p$. Therefore this latter procedure is not available in the high-dimensional case ($p > n$), which is exactly where there is higher chance of encountering many correlated predictors!

[10]We avoid using the term "objective", first, because as Gelman and Hennig (2017) argue it is counterproductive to do so and, second, because the g-prior is not in any way an objective prior.

in the likelihood. This is a basic principle of shrinkage and variable selection estimators: when p is large and especially when $p > n$, the information in the likelihood is not sufficient to estimate all p coefficients and the prior becomes increasingly important for determining posterior outcomes. That is, for both Bayesian and non-Bayesian approaches, the concepts of shrinkage and sparsity amount to the prior expectation that increasingly many coefficients a priori will be zero or close to zero.

Of course, there are more rigorous ways of selecting g. A key contribution is that of Liang et al. (2008) who put hyper-priors on g, treating it as a random variable. Such hierarchical approaches are the topic of close examination of the next section, so we won't expand on it here. Krishna et al. (2009) extend the g-prior into an *adaptive powered correlation prior* of the form

$$\boldsymbol{\beta}_r|\sigma^2, M_r \sim N_{p_r}\left(\mathbf{0}_{p_r}, \frac{1}{g}\sigma^2\left(\boldsymbol{X}'_r\boldsymbol{X}_r\right)^{\lambda}\right), \tag{45}$$

where $\lambda \in \mathbb{R}$ controls the prior's response to collinearity in predictors. $\lambda = -1$ gives the original prior proposed by Arnold Zellner, while $\lambda = 0$ gives the ridge regression prior.

While the g-prior addresses the issue of setting a prior on different regression models that might be nested and have correlated covariates, another important issue is how to define a prior on model space. For both conceptual and computational reasons Bayesians prefer to index all possible 2^p models using dummy variables $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)'$. When $\gamma_j = 0$ a covariate is excluded from a model and when $\gamma_j = 1$ it is included. Therefore, the model with no predictors is indexed as $\boldsymbol{\gamma} = (0, ..., 0)'$ and the model with all predictors is indexed as $\boldsymbol{\gamma} = (1, ..., 1)'$. All intermediate models are indexed by vectors $\boldsymbol{\gamma}$ that are sequences of zeros and ones. Instead of placing priors on the model space, we can now explicitly consider priors on $\boldsymbol{\gamma}$, and the binomial distribution is a good candidate for a parameter that takes 0/1 values. The binomial prior can become both uniform but also more informative when this is desirable (e.g. in high-dimensional spaces, where our prior is that only a small number of predictors will be important).

This setting that combines the g-prior on regression coefficients with a binomial prior on model space, is the major workhorse model for implementing Bayesian variable selection. While its theoretical underpinnings are well-understood (see Hoeting et al., 1999 for a thorough description), it provides the ground for some of the most interesting Bayesian work on computation in high-dimensional settings.[11] At the same time this setting possesses implicitly the benefits of a hierarchical prior approach. Therefore, we use this brief discussion of BMA as a stepping stone for introducing in the next the concept of full-Bayes/hierarchical Bayes priors that result in shrinkage and sparse estimators.

[11] See for example, Bottolo and Richardson (2010), Clyde et al. (2011), Dellaportas et al. (2002), Hans et al. (2007), Ji and Schmidler (2013), Madigan et al. (1995), Nott and Kohn (2005) and Peltola et al. (2012).

2 Hierarchical (full Bayes) priors

When interest lies in models with many parameters, simple priors such as the benchmark natural conjugate prior presented in the previous section, are inadequate for learning interesting features about our parameters and for quantifying uncertainty. In statistics, the concept of hierarchical or multi-level modeling refers to the process of enhancing a simpler model with a richer specification that allows for learning interesting features of a multi-parameter vector, such as groupings or sparsity and shrinkage towards zero, where the latter being the main focus of this review. The Bayesian interpretation of hierarchical modeling involves specifying prior distributions for the prior hyperparameters of regression coefficients, especially when p is large. A simple hierarchical specification for the regression coefficients $\boldsymbol{\beta}$,[12] takes the form

$$\begin{aligned} p\left(y_i|\boldsymbol{\beta},\sigma^2\right) &\sim N\left(x_i\boldsymbol{\beta},\sigma^2\right), \quad i=1,...,n, && (46)\\ p\left(\beta_j|\mu,\tau^2\right) &\sim N\left(0,\tau^2\right), \quad j=1,...,p, && (47)\\ p\left(\tau^2\right) &\sim F(a,b), && (48) \end{aligned}$$

where $F(a,b)$ denotes some distribution function with hyperparameters (a,b). Due to the fact that choice of τ^2 is so crucial for the posterior outcome of β_j, the idea behind this hierarchical specification is to treat the hyperparameter τ^2 as a random variable and learn about it from the data, via Bayes Theorem. For that reason, a prior such as the one in equations (47) - (48) is many times referred to as a *full-Bayes prior*, as it allows for full quantification of uncertainty around parameters of interest. While the example above pertains to linear regressions with Gaussian likelihood and prior distributions, Section 4 demonstrates that the concept of hierarchical priors is much more powerful and can be applied to numerous multivariate, non-Gaussian, nonlinear or other settings. Additionally, adaptive hierarchies can be defined in which β_j depends on hyperparameters specific to this j-th element (τ_j^2) that have their individual hyperprior distributions. Finally, if needed, further layers of the hierarchy can be defined: for instance, if choice of the hyperperameter a of τ^2 is not straightforward, we can define another level for the prior distribution of a, or we could introduce two variance parameters for β_j in Equation 47.[13]

An important feature of the hierarchical prior in equations (47) - (48) is that, while the

[12] Ignore estimation uncertainty of σ^2 for the moment, e.g. assume it is known and fixed.

[13] For example, a powerful class of hierarchical priors called *global-local shrinkage priors* (Polson and Scott, 2010) provides an excellent benchmark for specifying appropriate hierarchical priors. Such priors are of the form

$$\begin{aligned} p\left(\beta_j|\tau^2,\lambda_j^2\right) &\sim N\left(0,\tau^2\lambda_j^2\right), \quad j=1,...,p, && (49)\\ p\left(\tau^2\right) &\sim F_\tau(a,b), && (50)\\ p\left(\lambda_j^2\right) &\sim F_\lambda(c,d), && (51) \end{aligned}$$

where τ^2 is a global shrinkage parameter (applying the same shrinkage to the whole parameter vector $\boldsymbol{\beta}$) and λ_j is a local shrinkage parameter (applying shrinkage only to β_j). As we see next, such priors will typically have at least three hierarchical layers, but in practical situations they tend to have many more (e.g. by putting priors on some or all of the hyperparameters a,b,c,d).

conditional prior $p\left(\beta_j|\tau^2\right)$ is Gaussian, unconditionally the prior for β_j is non-Gaussian. Indeed the marginal prior for β_j becomes

$$p\left(\beta_j\right) = \int N\left(\beta_j; 0, \tau^2\right) p(\tau^2) d\tau^2, \tag{52}$$

that is, a *scale mixture of normals* representation that allows to approximate very complex prior shapes for β_j.[14] Mixtures have the benefit of allowing for classification and grouping of parameters. In the case of identifying sparsity and shrinkage, we can think of the mixture prior as grouping parameters into "important" and "non-important". Therefore, it is this implied mixture representation of hierarchical modeling with prior distributions that allows to extract interesting features in a multi-parameter setting. Finally, the posterior mode of $\boldsymbol{\beta}$ under a hierarchical prior has a penalized likelihood representation. For the linear regression model, penalized likelihood problems admit the following regularized least squares form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2 + g\left(\boldsymbol{\beta}, \lambda\right), \tag{53}$$

where the first term gives the solution to the usual least squares problem and the term $g\left(\boldsymbol{\beta}, \lambda\right)$ defines the penalty as a function of the regression parameters $\boldsymbol{\beta}$ and a scalar (or possibly vector) tuning parameter λ. Numerous penalized estimators, such as ridge (Tikhonov regularization), lasso, and elastic net fall under the general form in Equation 53, and Bayesian modal estimators under suitable hierarchical priors can fully recover all of them.

In order to understand the ability of hierarchical priors to classify parameters as important and non-important (or non-penalized and penalized), we plot in Figure 1 a normal prior with fixed variance vs three cases of a normal prior with variance parameter distributed as χ^2 with one degree of freedom, exponential with rate parameter $\lambda = 0.5$, and binomial with one trial and probability $\pi = 0.9$ (that is, a Bernoulli distribution). The simple normal prior provides more probability at the origin (zero) relative to its tails, however, it is fairly flat (diffuse) in a small area around zero. What the three mixture priors are introducing, is a more pronounced peak at zero such that when a parameter is in the region of zero it can be shrunk at a faster rate. At the same time, all three mixture distributions have fat tails, providing positive probability to parameter values that are far from zero. That is, these shapes allow for a clearer separation and classification of a parameter as being zero or non-zero. The extreme case of the Bernoulli prior on τ^2 (bottom right panel of Figure 1) creates a distribution that looks normal but also has a point mass at zero with high probability. Therefore, all three examples of hierarchical priors provide sharper inference in favor or against the groups of interest (important and non-important parameters).

[14] It is trivial to show that if τ^2 is not a fixed parameter, then unconditionally the prior for β_j always has excess kurtosis higher than zero, thus, being a leptokurtic distribution with tails thicker than the normal distribution.

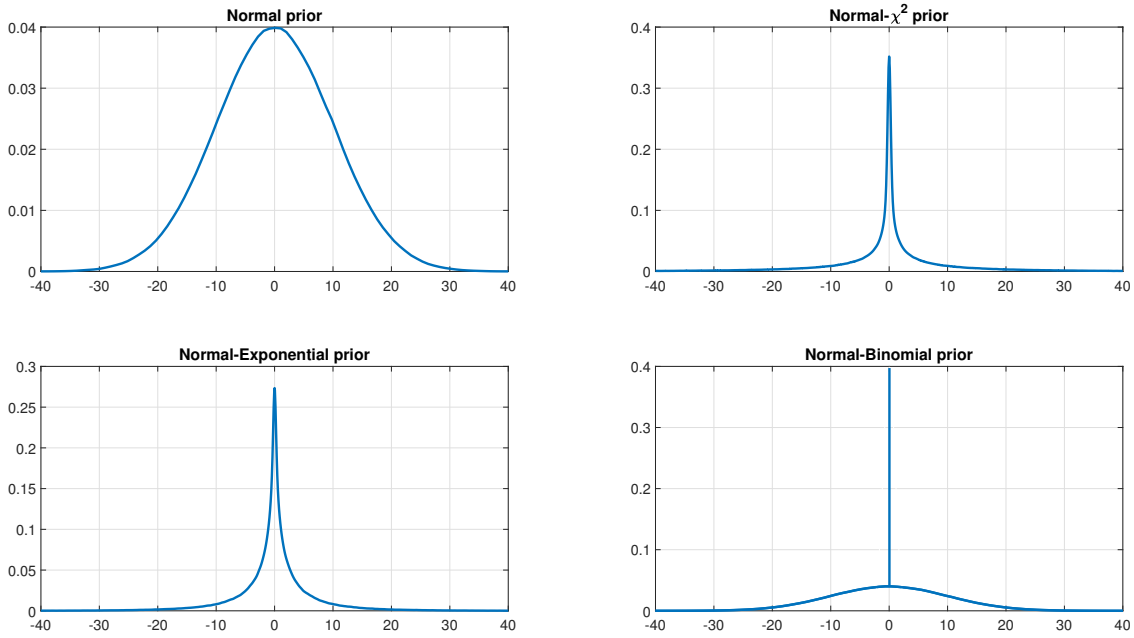

Figure 1: *Hierarchical priors for a scalar parameter. In all four panels the base distribution is* $\beta \sim N(0, 100 \times \tau^2)$. *In the top left panel* $\tau^2 = 1$ *is a fixed hyperparameter, while in the remaining panels it follows a* $\chi^2_{(1)}$, $Exp(0.5)$ *and* $Bernoulli(0.9)$ *priors.*

Computation with hierarchical priors is reviewed in detail in the next section. For now it suffices to note that because of the conditional structure of hierarchical priors, conditional posteriors are typically easy to derive even if the joint parameter posterior is intractable. Sampling from these conditional posteriors using Markov chain Monte Carlo (the *Gibbs sampler*, in particular) is equivalent to taking samples from the intractable joint posterior. Additionally, several approximate methodologies such as variational Bayes and maximum a-posterior (MAP) estimation rely on similar conditional distributions. Therefore, in our discussion in this section we present various hierarchical priors, explain their properties and focus on deriving conditional posteriors. In the next section we discuss in more detail how to use these conditional posteriors to estimate the desired parameters.[15]

2.1 Diffusing hierarchical prior

A natural choice for the variance parameter τ^2 in the hierarchical model of equations (46) - (48) is a prior distribution that is diffuse. Similar to Jeffrey's prior for the regression variance σ^2, the choice $\tau^2 \sim U(0, \infty)$ equivalently $p(\tau^2) \propto \tau^{-2}$ can be thought as a default prior choice that reflects our lack of information about sparsity patterns in the data. We might want to also allow for each β_j to be determined adaptively, in which case a Jeffrey's prior on hyperparameters τ_j^2, $j = 1, ..., p$ can be defined. Therefore, the full hierarchical prior specification for the regression

[15] Additional derivations and computational details can be found in the accompanying Technical Document.

model is of the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p &\sim N_p(\mathbf{0}, \boldsymbol{D}_\tau), && (54)\\
\tau_j^2 &\sim \frac{1}{\tau_j^2}, \quad \text{for } j = 1, ..., p, && (55)\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && (56)
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$. While a Jeffrey's prior on τ_j^2 is a first natural attempt towards hierarchical prior modeling, as Lindley (1983) notes, "a prior for τ^2 that behaves like τ^{-2} will cause trouble" meaning it will lead to an improper posterior. Gelman (2006) examines this issue in more detail and explains why a $Uniform(-\infty, \infty)$ prior on $\log\left(\tau^2\right)$ would also not work. However, as Kahn and Raftery (1992) and Gelman (2006) note, under certain conditions, Jeffrey's prior on τ^2 yields a limiting proper posterior density. Note that the same improper density can be obtained from the prior $\tau^2 \sim Inv - Gamma(\epsilon, \epsilon)$ for $\epsilon \to 0$ (see also subsection 2.2 below). Gelman (2006) argues that the $Inv - Gamma(\epsilon, \epsilon)$ prior does not have any proper limiting posterior distribution, such that inference becomes sensitive to the choice of ϵ – simply setting ϵ to any "small" value is not a reliable solution.

Figueiredo (2003) and Bae and Mallick (2004) are examples of empirical studies that rely on shrinkage using a uniform hyperprior distribution. Tipping (2001) specifies an inverse gamma prior on τ^2 (and calls the resulting hierarchical structure a *sparse Bayesian learning* prior) and adopts the limiting case $\epsilon = 10^{-4}$ as the default hyperparmeter choice. Diffusing priors should not be the first choice in empirical settings especially in high-dimensional and ultra-high-dimensional settings. There are numerous other hyperprior distributions that are interpretable and have better theoretical guarantees (Gelman, 2006).

2.2 Student-t shrinkage

While we just argued that it is not desirable to use the inverse gamma distribution as a way of imposing a diffusing prior on τ^2, informative inverse gamma priors provide flexible parametric shrinkage. Following the specification of the normal-inverse gamma prior in Armagan and Zaretzki (2010), we write this prior using the following form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p &\sim N_p(\mathbf{0}, \boldsymbol{D}_\tau), && (57)\\
\frac{1}{\tau_j^2} &\sim Gamma\left(\rho, \xi\right), \quad \text{for } j = 1, ..., p, && (58)\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && (59)
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$. This is a scale mixture of normals representation of the fat-tailed and leptokurtic Student-t distribution. Similar to our arguments in Figure 1 the excess kurtosis of the Student-t results in shrinkage towards zero at a faster rate than the simple

normal distribution. At the same time the fatter tails accommodate values of τ^2 that can be far from zero. In Figure 2 we illustrate the shape of the marginal distribution of β_j for various values of the parameters ρ, ξ.

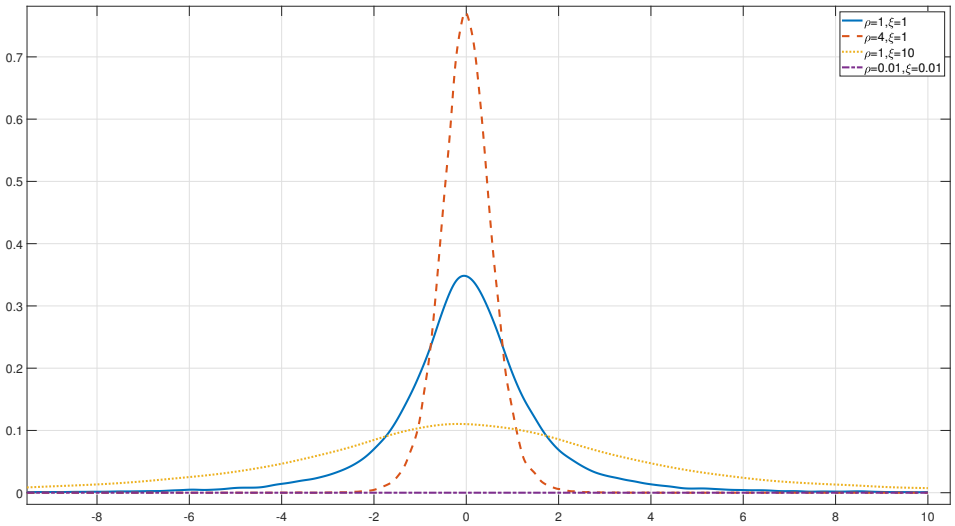

Figure 2: *Marginal distribution of β_j for the Student-t prior.*

Similar to an inverse gamma prior for the variance parameter σ^2, the conjugacy of this distribution allows for numerous methods of inference using this prior. For example, Tipping (2001) uses *type-II maximum likelihood* methods (Berger, 1985), but (as we discuss in the following section) variational Bayes and other approximate algorithms are also trivial to derive. Armagan and Zaretzki (2010) show that conditional posteriors are of the form

$$\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \sigma^2, \boldsymbol{y} \sim N_p\left(\boldsymbol{V} \times \left(\boldsymbol{X}'\boldsymbol{y}\right), \boldsymbol{V}\right), \tag{60}$$

$$\frac{1}{\tau_j^2}\Big|\beta_j, \boldsymbol{y} \sim Gamma\left(\rho + \frac{1}{2}, \xi + \frac{\beta_j^2}{2}\right), \quad j = 1, ..., p, \tag{61}$$

$$\frac{1}{\sigma^2}\Big|\boldsymbol{\beta}, \boldsymbol{y} \sim Gamma\left(\frac{n}{2}, \frac{\Psi}{2}\right) \tag{62}$$

where $\boldsymbol{V} = \left(\sigma^{-2}\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)^{-1}$ and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. The Gibbs sampler can be used to sample sequentially from these conditional posteriors, as these samples are guaranteed to be samples from the desired joint parameter posterior.

For the conditional posterior (61) of the prior precisions, $\frac{1}{\tau_j^2}$, we have

$$p\left(\frac{1}{\tau_1^2},\ldots,\frac{1}{\tau_p^2}\middle|\boldsymbol{\beta},\boldsymbol{y}\right) \propto \prod_{j=1}^{p}(2\pi\tau_j^2)^{-1/2}\exp\left[-\frac{1}{2\tau_j^2}\beta_j^2\right]\left(\frac{1}{\tau_j^2}\right)^{\rho-1}\exp\left[-\frac{\xi}{\tau_j^2}\right]$$
$$\propto \prod_{j=1}^{p}\left(\frac{1}{\tau_j^2}\right)^{\left(\rho+\frac{1}{2}\right)-1}\exp\left[-\frac{1}{\tau_j^2}\left(\xi+\frac{\beta_j^2}{2}\right)\right]$$

where the proportional sign is with respect to $\frac{1}{\tau_j^2}$'s. It can be seen that the conditional posterior of $\frac{1}{\tau_j^2}$'s is independent across j and that it has the form in Equation 61.

2.3 Normal-gamma priors

Caron and Doucet (2008) proposed the normal-gamma family of hierarchical priors, and Griffin and Brown (2010, 2017) established further results and their excellent properties. This prior takes the following hierarchical form

$$\boldsymbol{\beta} \sim N_p(\mathbf{0},\boldsymbol{D}_\tau), \tag{63}$$
$$\tau_j^2 \sim Gamma(\lambda,\gamma^2/2), \tag{64}$$
$$\sigma^2 \sim \frac{1}{\sigma^2}, \tag{65}$$

where again $\boldsymbol{D}_\tau = diag(\tau_1^2,...,\tau_p^2)$. The pdf of τ_j is

$$p\left(\tau^2\right) = \frac{\left(\frac{\gamma^2}{2}\right)^\lambda}{\Gamma\left(\lambda\right)}\left(\tau_j^2\right)^{\lambda-1}\exp\left(-\frac{\gamma^2}{2}\tau_j^2\right), \tag{66}$$

such that the marginal pdf of β_j is

$$p(\beta_j) = \frac{\gamma^{(\lambda+1/2)}}{2^{(\lambda-1/2)}\sqrt{\pi}\Gamma\left(\lambda\right)}|\beta_j|^{(\lambda-1/2)}\mathcal{K}_{(\lambda-1/2)}\left(\gamma|\beta_j|\right) \tag{67}$$

where $\mathcal{K}_v$ is the modified Bessel function of the second kind, and the tails of this distribution decrease in $|\beta_j|^{(\lambda-1)}\exp\left(\gamma|\beta_j|\right)$.

Due to the connection of the gamma distribution with a wide array of other distributions (e.g. inverse gamma, inverse Gaussian, χ^2, etc) choice of the hyperparameters λ and γ^2 can result in various shapes for the unconditional distribution of $\boldsymbol{\beta}$ that have different shrinkage properties. This prior becomes diffusing when $\lambda,\gamma^2\to 0$, however, this choice falls under the same critique of Gelman (2006) for the diffusing inverse gamma prior. This is due to the fact that when $\lambda < 1/2$ the normal-gamma prior places infinite mass in the vicinity of zero, that is, $\lim_{\beta_j\to 0}p(\beta_j)=\infty$.

2.4 LASSO prior and extensions

The least absolute shrinkage and selection operator (lasso) of Tibshirani (1996) has been established as a key workhorse of scientists in all fields working with high-dimensional settings. The estimator takes the form

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| < t, \tag{68}$$

where t is a prespecified free parameter that determines the degree of regularization. The Lagrangian form of this program is

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2 + \lambda||\boldsymbol{\beta}||_1, \tag{69}$$

where $||x||_1 = \sum |x_i|$ is the ℓ_1 norm and $||x||_2 = \sqrt{\sum x_i^2}$ is the ℓ_2 norm. λ is a tuning parameter related to t, controlling for how strongly shrinkage is exercised. As $\lambda \to 0$ the penalty term vanishes and the lasso becomes indistinguishable from the least squares problem. This optimization formula is related to *basis pursuit denoising*, which is the preferred term for the lasso among researchers in computer science and signal processing.

Tibshirani (1996, Section 5) first noted that the lasso estimate can be derived as a Bayes posterior mode under the following Laplace prior distribution

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp\left(-\lambda|\beta_j|\right) = \left(\frac{\lambda}{2}\right)^p \exp\left(-\lambda||\boldsymbol{\beta}||_1\right). \tag{70}$$

However, as Castillo et al. (2015) note the full posterior distribution under a Laplace prior does not contract at the same rate as its mode, making uncertainty quantification using the Bayesian lasso unreliable. The intuition behind this is that the λ coefficient above needs to be large enough to penalize coefficients β_j to zero, but not too large such that nonzero coefficients can be modeled. This issue is addressed by modifications such as the adaptive lasso (Alhamzawi and Ali (2018); see end of this section) and the spike and slab lasso (Ročková and George (2018); see section on spike and slab priors) and is related to the motivating arguments of Johnson and Rossell (2010) for proposing the non-local priors (see relevant section below).

The first application of the lasso prior stems from computing science and is due to Girolami (2001). While the joint parameter posterior under a Laplace prior is not of standard form, Girolami (2001) used variational Bayes inference (which at the time was not popular in mainstream statistics) to approximate the posterior mean and variance. Figueiredo (2003) used the fact that the Laplace prior admits a hierarchical representation in the form of a normal-exponential (double exponential) mixture. The hierarchical representation of this prior

is of the form

$$\begin{aligned} \beta_j|\tau_j &\sim N(0,\tau_j^2), && (71)\\ \tau_j^2|\lambda^2 &\sim \textit{Exponential}\left(\frac{\lambda^2}{2}\right), && (72) \end{aligned}$$

where the exponential distribution has the functional form $p(\tau^2|\lambda^2) = \left(\frac{\lambda^2}{2}\right)\exp\left(\frac{\lambda^2}{2}\tau_j^2\right)$. The marginal distribution for β conditional on λ^2 is of the form

$$p\left(\beta_j|\lambda^2\right) = \frac{\sqrt{\lambda^2}}{2}\exp\left(\sqrt{\lambda^2}|\beta_j|\right) \equiv \frac{\lambda}{2}\exp\left(\lambda|\beta_j|\right), \quad (73)$$

which is the desired Laplace distribution for β_j. Figueiredo (2003) derived an EM algorithm for obtaining the posterior mode (MAP estimator).

A formal Bayesian treatment of the Bayesian lasso using MCMC can be found in Park and Casella (2008). These authors choose to specify the Bayesian lasso as a normal-exponential mixture but conditional on the regression variance σ^2. This is because a hierarchical prior on β_j that is independent of σ^2 results in a multimodal posterior for β_j. The Park and Casella (2008) Laplace prior takes the form

$$\begin{aligned} \boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p,\sigma^2 &\sim N_p(\mathbf{0},\sigma^2\boldsymbol{D}_\tau), && (74)\\ \tau_j^2|\lambda^2 &\sim \textit{Exponential}\left(\frac{\lambda^2}{2}\right), \quad \text{for } j=1,...,p, && (75)\\ \lambda^2 &\sim \textit{Gamma}(r,\delta) && (76)\\ \sigma^2 &\sim \frac{1}{\sigma^2}, && (77) \end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2,...,\tau_p^2)$. Conditional posteriors under this hierarchical representation are trivial to derive and more details can be found in the accompanying Technical Document.

The approach in Park and Casella (2008) is probably the most widely used but it is not the only one available. Hans (2009) specified the lasso in terms of the normal orthant distribution. Let $\mathfrak{Z} = \{-1,1\}^p$ represent the set of all 2^p possible vectors of length p whose elements are ± 1. For any realization $z \in \mathfrak{Z}$ define the orthant $\mathcal{O}_z \subset \mathbb{R}^p$. If $\boldsymbol{\beta} \in \mathcal{O}_z$, then $\beta_j \geq 0$ if $z = 1$ and $\beta_j < 0$ if $z = -1$. Then $\boldsymbol{\beta}$ follows the normal-orthant distribution with mean m and covariance S, which is of the form

$$\boldsymbol{\beta} \sim N^{[z]}(\boldsymbol{m},\boldsymbol{S}) \equiv \Phi(\boldsymbol{m},\boldsymbol{S})\, N_p(\boldsymbol{m},\boldsymbol{S})\, I(\in \mathcal{O}_z). \quad (78)$$

The Hans (2009) prior takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\lambda,\sigma &\sim \left(\frac{\lambda}{2\sqrt{\sigma^2}}\right)^p \exp\left(-\lambda\sum_{j=1}^{p}|\beta_j|/\sqrt{\sigma^2}\right), && (79)\\
\lambda &\sim Gamma(r,\delta), && (80)\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && (81)
\end{aligned}$$

and, using the definition of the normal orthant distribution, conditional posteriors are of the form

$$\begin{aligned}
\beta_j|\beta_{-j},\lambda,\sigma^2,\boldsymbol{y} &\sim \phi_j N^{[+]}\left(\mu_j^+,\omega_{jj}^{-1}\right)+(1-\phi_j)N^{[-]}\left(\mu_j^-,\omega_{jj}^{-1}\right), && (82)\\
\lambda|\boldsymbol{y} &\sim Gamma\left(p+r,\frac{\sum_{j=1}^{p}|\beta|}{\sqrt{\sigma^2}}+\delta\right), && (83)\\
\sigma|\boldsymbol{\beta},\boldsymbol{y} &\propto (\sigma^2)^{-(\frac{n+p}{2}+1)}\exp\left(\frac{\Psi}{2\sigma^2}-\frac{\lambda\sum_{j=1}^{p}|\beta|}{\sqrt{\sigma^2}}\right), && (84)
\end{aligned}$$

where:

- $N^{[-]}$ and $N^{[+]}$ correspond to the $N^{[z]}$ distribution for $z=-1$ and $z=1$, respectively;
- $\mu_j^+ = \widehat{\beta}_j^{OLS} + \left\{\sum_{i=1,i\neq j}^{p}\left(\widehat{\beta}_i^{OLS}-\beta_i\right)(\omega_{ij}/\omega_{jj})\right\} + \left(-\frac{\lambda}{\sqrt{\sigma^2}\omega_{jj}}\right)$;
- ω_{ij} is the ij element of the matrix $\Omega=\Sigma^{-1}=\left(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)^{-1}$;
- $\phi_j = \dfrac{\Phi\left(\frac{\mu_j^+}{\sqrt{\omega_{jj}}}\right)/N\left(0|\mu_j^+,\omega_{jj}^{-1}\right)}{\Phi\left(\frac{\mu_j^+}{\sqrt{\omega_{jj}}}\right)/N\left(0|\mu_j^+,\omega_{jj}^{-1}\right)+\Phi\left(-\frac{\mu_j^-}{\sqrt{\omega_{jj}}}\right)/N\left(0|\mu_j^-,\omega_{jj}^{-1}\right)}$;
- $\Psi=(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$.

The conditional posterior of σ^2 is not of a standard form and, therefore, cannot be sampled directly. Hans (2009) suggests a simple accept/reject step within the Gibbs sampler that allows to obtain approximate samples from the posterior of σ^2. Finally, Mallick and Yi (2014) propose a third hierarchical representation of the Laplace prior, this time as a mixture of Uniform distributions (see our Technical Document for details of this algorithm).

There are numerous extensions to the basic lasso that come in various forms. For example, the elastic net combines the benefits of ridge regression (ℓ_2 penalization) and the lasso (ℓ_1 penalization) by solving the problem

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2+\lambda_1||\boldsymbol{\beta}||_1+\lambda_2||\boldsymbol{\beta}||_2, \quad (85)$$

where now λ_1 and λ_2 are tuning parameters. The Bayesian prior that provides the solution to

the elastic net estimation problem is of the form

$$\boldsymbol{\beta}|\sigma^2 \sim \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2\right)\right\}. \tag{86}$$

Li and Lin (2010) start from this prior and derive a mixture approximation and a Gibbs sampler that has the minor disadvantage that requires an accept-reject step for obtaining samples from the conditional posterior of σ^2 (similar to the sampler of Hans (2009) for the lasso). The formulation of the elastic net prior in Kyung et al. (2010) is slightly different to the one above, but they manage to derive a slightly different mixture representation and a slightly more straightforward Gibbs sampler.

Other popular extensions to the lasso include the group lasso that allows for group shrinkage; the fused lasso that allows for spatial or temporal relationships between neighbouring parameters; and the adaptive lasso that fixes some variable selection consistency issues with the regular lasso. All these extensions have straightforward hierarchical forms, and we refer the reader to discussions in Kyung et al. (2010), Griffin and Brown (2011), Leng et al. (2014) and Alhamzawi and Ali (2018), among several other studies. Our Technical document provides details of posterior inference using the elastic net, group lasso, fused lasso and adaptive lasso.

2.5 Generalized double Pareto shrinkage

Armagan et al. (2013) propose the following generalized double Pareto (GDP) prior on $\boldsymbol{\beta}$

$$\boldsymbol{\beta}|\sigma \sim \prod_{j=1}^{p}\frac{1}{2\sigma\delta/r}\left(1+\frac{1}{r}\frac{|\beta_j|}{\sigma\delta/r}\right)^{-(r+1)}. \tag{87}$$

This distribution can be represented using the familiar, from the Bayesian lasso, normal-exponential-gamma mixture. The only difference is that, while the Exponential component has the same rate parameter for all $j = 1, ..., p$, in the representation of the GDP mixture this parameter is adaptive. The generalized double Pareto distribution has a spike at zero with Student's t-like heavy tails.

The generalized double Pareto prior takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j\}_{j=1}^{p}, \sigma^2 &\sim N_p\left(\mathbf{0}, \sigma^2\boldsymbol{D}_\tau\right), && (88)\\
\tau_j^2|\lambda_j &\sim Exponential\left(\frac{\lambda_j^2}{2}\right), \quad \text{for } j = 1, ..., p, && (89)\\
\lambda_j &\sim Gamma(r, \delta), \quad \text{for } j = 1, ..., p, && (90)\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && (91)
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \sigma^2, \boldsymbol{y} &\sim N_p\left(\boldsymbol{V} \times \left(\boldsymbol{X}'\boldsymbol{y}\right), \sigma^2\boldsymbol{V}\right), && (92)\\
\frac{1}{\tau_j^2}|\beta_j, \lambda_j^2, \boldsymbol{y} &\sim IG\left(\sqrt{\frac{\lambda_j^2\sigma^2}{\beta_j^2}}, \lambda^2\right), \quad \text{for } j = 1, ..., p, && (93)\\
\lambda_j^2|\boldsymbol{y} &\sim Gamma\left(r+1, \sqrt{\frac{\beta_j^2}{\sigma^2}} + \delta\right), && (94)\\
\frac{1}{\sigma^2}|\boldsymbol{\beta}, \boldsymbol{y} &\sim Gamma\left(\frac{n-1+p}{2}, \frac{\Psi}{2} + \frac{\boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right), && (95)
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)^{-1}$, $\boldsymbol{D}_\tau^{-1} = diag(\tau_1^{-2}, ..., \tau_p^{-2})$ and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Pal et al. (2017) show, both theoretically and numerically, that the above "three-block" Gibbs sampler is less efficient than a modified two-block Gibbs sampler they propose.

2.6 Dirichlet-Laplace

The Dirichlet-Laplace prior was introduced in Bhattacharya et al. (2015), and Zhang and Bondell (2018) studied its posterior consistency as well as consistency in variable selection in the context of a linear regression model. The Dirichlet-Laplace hierarchical prior, which is a generalization of the Laplace prior, takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j\}_{j=1}^p, \{\psi_j\}_{j=1}^p, \lambda, \sigma^2 &\sim N_p\left(\boldsymbol{0}, \sigma^2\boldsymbol{D}_{\lambda,\tau,\psi}\right), && (96)\\
\tau_j^2 &\sim Exponential(1/2), \quad j = 1, ..., p, && (97)\\
\psi_j &\sim Dirichlet(\alpha), \quad j = 1, ..., p, && (98)\\
\lambda &\sim Gamma(n\alpha, 1/2), && (99)\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && (100)
\end{aligned}$$

where $\boldsymbol{D}_{\lambda,\tau,\psi} = diag(\lambda^2\tau_1^2\psi_1^2, ..., \lambda^2\tau_p^2\psi_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \{\psi_j\}_{j=1}^p, \lambda, \sigma^2, \boldsymbol{y} &\sim N_p\left(\boldsymbol{V} \times \left(\boldsymbol{X}'\boldsymbol{y}\right), \sigma^2 \boldsymbol{V}\right), && (101)\\
\frac{1}{\tau_j^2}|\lambda^2, \sigma^2, \boldsymbol{y} &\sim IG\left(c^*, 1\right), \quad j = 1, ..., p, && (102)\\
\lambda|\boldsymbol{\beta}, \boldsymbol{y} &\sim GIG\left(2\frac{\sum_{j=1}^p |\beta_j|}{\psi_j \sigma}, 1, p(\alpha - 1)\right), && (103)\\
\psi_j|\boldsymbol{\beta}, \boldsymbol{y} &= \frac{T_j}{\sum_{j=1}^p T_j}, \quad j = 1, ..., p, && (104)\\
\text{where } T_j &\sim GIG\left(2\sqrt{\frac{\beta_j^2}{\sigma^2}}, 1, \alpha - 1\right) && (105)\\
\frac{1}{\sigma^2}|\boldsymbol{\beta}, \boldsymbol{y} &\sim Gamma\left(a^*, b^*\right), && (106)
\end{aligned}$$

where $a^* = (n+p)/2$, $b^* = (\Psi + \boldsymbol{\beta}'\boldsymbol{D}_{\tau,\lambda,\psi}^{-1}\boldsymbol{\beta})/2$, $c^* = \sqrt{\lambda^2\psi_j^2\sigma^2/\beta_j^2}$, $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{\tau,\lambda,\psi}^{-1}\right)^{-1}$, and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. IG is the two-parameter inverse Gaussian distribution, and GIG is the three-parameter generalized inverse Gaussian.

2.7 Horseshoe prior

The horseshoe prior was first introduced by Carvalho et al. (2010) and it since its inception has been the most popular and influential hierarchical prior in Bayesian inference. The survey paper by Bhadra et al. (2020) provides a thorough review of the applications of this prior in numerous inference problems in statistics and machine learning, including nonlinear models and neural networks. The Horseshoe is a prime representative of the class of global-local shrinkage priors (see Footnote 13) and it can be represented as a scale mixture of normals with half-Cauchy mixing distributions. That is, the prior has the following formulation

$$\begin{aligned}
\boldsymbol{\beta}|\{\lambda_j\}_{j=1}^p, \tau &\sim N_p\left(\boldsymbol{0}, \sigma^2\tau^2\boldsymbol{\Lambda}\right), && (107)\\
\lambda_j|\tau &\sim C^+(0,1), \quad \text{for } j = 1, ..., p, && (108)\\
\tau &\sim C^+(0,1), && (109)
\end{aligned}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1^2, ..., \lambda_p^2)$, and $C^+(0, \alpha)$ is the half-Cauchy distribution on the positive reals with scale parameter α. That is, λ_j has conditional prior density

$$\lambda_j|\tau = \frac{2}{\pi\tau\left(1 + (\lambda_j/\tau)^2\right)}. \tag{110}$$

Under this hierarchical specification, the marginal prior for each β_j is unbounded at the origin and has tails that decay polynomially.

There are numerous theoretical results established for this prior, most notably Datta and

Ghosh (2013) and van der Pas et al. (2014), and the reader is referred to Bhadra et al. (2020) for a more detailed discussion. There are also various computational approaches to the Horseshoe (see the accompanying Technical Document for details), but the most straightforward is the one proposed by Makalic and Schmidt (2016). These authors note that the half-Cauchy distribution can be written as a mixture of inverse-gamma distributions. In particular, if

$$x^2|z \sim Inv-Gamma(1/2, 1/z), \qquad z \sim Inv-Gamma(1/2, 1/\alpha^2), \tag{111}$$

then $x \sim C^+(0, \alpha)$. Therefore, the Makalic and Schmidt (2016) prior takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\lambda_j\}_{j=1}^p, \tau, \sigma^2 &\sim N\left(\mathbf{0}, \sigma^2\tau^2\boldsymbol{\Lambda}\right), && (112)\\
\lambda_j^2|v_j &\sim Inv-Gamma(1/2, 1/v_j), \quad j = 1, ..., p, && (113)\\
v_j &\sim Inv-Gamma(1/2, 1), \quad j = 1, ..., p, && (114)\\
\tau^2|\xi &\sim Inv-Gamma(1/2, 1/\xi), && (115)\\
\xi &\sim Inv-Gamma(1/2, 1), && (116)\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && (117)
\end{aligned}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1^2, ..., \lambda_p^2)$.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\{\lambda_j\}_{j=1}^p, \tau^2, \sigma^2, \boldsymbol{y} &\sim N_p\left(\boldsymbol{V} \times \left(\boldsymbol{X}'\boldsymbol{y}\right), \sigma^2\boldsymbol{V}\right), && (118)\\
\lambda_j^2|\boldsymbol{\beta}, v_j, \tau^2, \sigma^2\boldsymbol{y} &\sim Inv-Gamma\left(1, \frac{1}{v_j} + \frac{\beta_j^2}{2\tau^2\sigma^2}\right), \quad j = 1, ..., p, && (119)\\
v_j|\lambda_j, \boldsymbol{y} &\sim Inv-Gamma\left(1, 1 + \frac{1}{\lambda_j^2}\right), \quad j = 1, ..., p, && (120)\\
\tau^2|\boldsymbol{\beta}, \xi, \{\lambda_j\}_{j=1}^p, \sigma^2, \boldsymbol{y} &\sim Inv-Gamma\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{1}{2\sigma^2}\sum_{j=1}^p \frac{\beta_j^2}{\lambda_j^2}\right) && (121)\\
\xi|\tau^2, \boldsymbol{y} &\sim Inv-Gamma\left(1, 1 + \frac{1}{\tau^2}\right), && (122)\\
\sigma^2|\boldsymbol{\beta}, \boldsymbol{y} &\sim Inv-Gamma\left(\frac{n+p}{2}, \frac{\Psi}{2} + \frac{\boldsymbol{\beta}'\boldsymbol{D}_{\tau,\lambda}^{-1}\boldsymbol{\beta}}{2}\right), && (123)
\end{aligned}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_{\tau,\lambda}^{-1}\right)^{-1}$, $\boldsymbol{D}_{\tau,\lambda} = diag(\tau^2\lambda_1^2, ..., \tau^2\lambda_p^2) = \tau^2\boldsymbol{\Lambda}$ and $\Psi = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$.

2.8 Generalized Beta mixtures of Gaussians

Armagan et al. (2011) motivate the use of a three-parameter beta (TPB) distribution for the prior variance parameter, as a flexible class of shrinkage priors. The TPB distribution takes

the form

$$p(x|a,b,\varphi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\varphi^b x^{b-1}(1-x)^{a-1}\left[1+(\varphi-1)x\right]^{-(a+b)}, \tag{124}$$

for $0 < x < 1$, $a, b, \varphi > 0$. The TPB normal scale mixture representation for the distribution of random variable β_j is given by

$$\beta_j \sim N(0, 1/\rho_j - 1), \qquad \rho_j \sim TBP(a,b,\varphi). \tag{125}$$

Proposition 1 in Armagan et al. (2011) shows that this distribution can either be written as normal-inverted beta mixture, or a normal-gamma-gamma mixture. The second choice gives a very straightforward Gibbs sampler scheme, and it can be seen as a special case of the normal-gamma class of priors (Griffin and Brown, 2017).

The Generalized Beta mixtures of Gaussians prior takes the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p &\sim N_p(\mathbf{0}, \boldsymbol{D}_\tau), && (126)\\
\tau_j^2|\lambda_j &\sim Gamma(a, \lambda_j), \quad \text{for } j = 1,...,p, && (127)\\
\lambda_j|\varphi &\sim Gamma(b, \varphi), \quad \text{for } j = 1,...,p, && (128)\\
\varphi &\sim Gamma\left(\frac{1}{2}, \omega\right), && (129)\\
\omega &\sim Gamma\left(\frac{1}{2}, 1\right), && (130)\\
\sigma^2 &\sim \frac{1}{\sigma^2}, && (131)
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2,...,\tau_p^2)$. Note that setting $a = b = 1/2$ we can obtain the horseshoe prior of Carvalho et al. (2010). For other choices we can recover popular cases of shrinkage priors.

The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta}|\{\tau_j^2\}_{j=1}^p, \sigma^2, \boldsymbol{y} &\sim N_p\left(\boldsymbol{V} \times \left(\boldsymbol{X}'\boldsymbol{y}\right), \boldsymbol{V}\right), && (132)\\
\tau_j^2|\beta_j, \lambda_j^2, \boldsymbol{y} &\sim GIG\left(a - \frac{1}{2}, 2\lambda_j, \beta_j^2\right), \quad \text{for } j = 1,...,p, && (133)\\
\lambda_j|\boldsymbol{y} &\sim Gamma(a+b, \tau_j^2 + \varphi), \quad \text{for } j = 1,...,p, && (134)\\
\varphi|\{\lambda_j\}_{j=1}^p, \omega, \boldsymbol{y} &\sim Gamma\left(pb + \frac{1}{2}, \sum_{j=1}^p \lambda_j + \omega\right), && (135)\\
\omega|\varphi, \boldsymbol{y} &\sim Gamma(1, \varphi + 1), && (136)\\
\frac{1}{\sigma^2}|\boldsymbol{\beta}, \boldsymbol{y} &\sim Gamma\left(\frac{n+p}{2}, \frac{\Psi}{2}\right), && (137)
\end{aligned}$$

where $\boldsymbol{V} = \left(\sigma^{-2}\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)^{-1}$, $\boldsymbol{D}_\tau = diag(\tau_1^2,...,\tau_p^2)$ and $\Psi = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

The TPB normal mixture includes as special cases Strawderman-Berger and horseshoe priors.

2.9 Non-local priors

Non-local priors have been proposed by Johnson and Rossell (2010) in the context of hypothesis testing of the form $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. From a frequentist perspective, such testing procedures are used in order to find out how likely it would be for a set of observations to occur under the null hypothesis. However, in a Bayesian setting the data are assumed to be observed once, and parameters are continuous random variables. Traditional (local) priors put significant probability in both the null and alternative hypotheses, thus, making it harder for the (continuous) posterior distribution to detect-non zero coefficients asymptotically. Non-local densities place zero probability at zero, and this feature allows such priors to separate more clearly between the null and alternative hypotheses. That is, such priors do not place any prior probability under the null.[16]

Any distribution that "decreases to 0 near the boundaries between disjoint null and alternative parameter spaces might be considered" (Johnson and Rossell, 2010) to be a non-local prior density. Within the context of a linear regression setting similar to the one defined in Equation 8, Johnson and Rossell (2012) propose two specific classes of priors. The first class of prior densities for $\boldsymbol{\beta}$ consists of product moment (pMOM) densities, which are defined as

$$p\left(\boldsymbol{\beta}|\tau^2,\sigma^2,r\right) \propto (2\pi)^{-p/2}\left(\tau^2\sigma^2\right)^{-p(r+1/2)} \exp\left\{-\frac{1}{2\tau^2\sigma^2}\boldsymbol{\beta}'\boldsymbol{\beta}\right\}\prod_{j=1}^{p}\beta_j^{2r}. \tag{138}$$

Figure 3 plots the pMOM density for $\tau^2 = \sigma^2 = 1$ and for three values of r ($r = 1, 2, 3$). This graph clearly shows the shapes that this prior can achieve, especially with regards to the rate at which this prior decreases in the region of zero. The second class of prior densities consists of the product inverse moment (piMOM) densities, which are defined as

$$p\left(\boldsymbol{\beta}|\tau^2,\sigma^2,r\right) = \frac{\left(\tau^2\sigma^2\right)^{rp/2}}{\Gamma\left(r/2\right)^p} \exp\left\{-\tau^2\sigma^2\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\right\}\prod_{j=1}^{p}|\beta_j|^{-(r+1)}. \tag{139}$$

In both of these two priors, τ^2 is a scale parameter that determines dispersion of the prior around zero. Therefore, this parameter determines the size of the regression coefficients that will be shrunk to zero, and it is of prime importance. Johnson and Rossell (2012) and Shin et al. (2018) treat τ^2 to be fixed and show that high-dimensional model selection consistency is achieved under the pMOM prior, as long as τ^2 is of a larger order than $\log p$ and it increases subexponentially in n. However, fixing this parameter might not be desirable in most applied

[16] As Johnson and Rossell (2010) note:

> [...] to a large extent, we have ignored philosophical issues regarding the logical necessity to specify an alternative hypothesis that is distinct from the null hypothesis. In general, it is our view that one hypothesis (and a test statistic) is enough to obtain a p-value, but that two hypotheses are required to obtain a Bayes factor.

high-dimensional problems[17], and a hierarchical approach might be desirable. Cao et al. (2020) propose a hyperprior density for τ^2 of the form

$$p(\tau^2) = \frac{\left(\frac{n}{2}\right)^{1/2}}{\Gamma\left(\frac{1}{2}\right)} \tau^{-3} \exp\left(-\frac{n}{2\tau^2}\right). \tag{140}$$

The hierarchical pMOM (or "hyper-pMOM") prior they propose achieves strong model selection consistency when p increases at a polynomial rate with n. Unfortunately, neither the pMOM, hyper-pMOM or piMOM priors allows for a closed-form computation of joint, marginal or conditional posteriors. Therefore, Cao et al. (2020) rely on Laplace approximations.

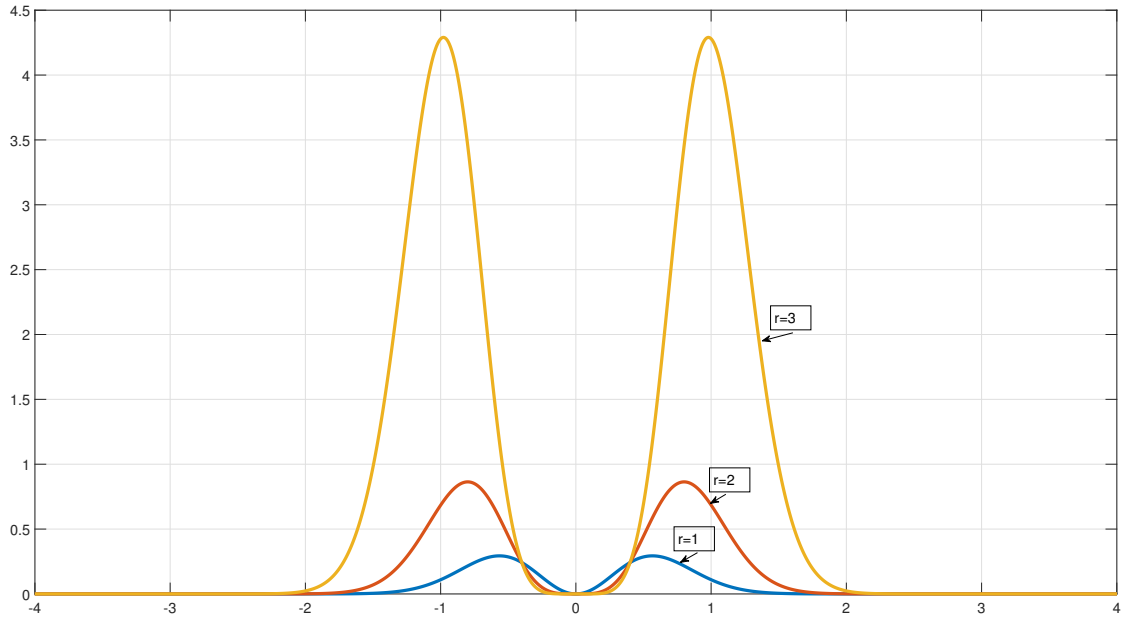

Figure 3: *Plot of pMOM density for* $r = 1, 2, 3$.

2.10 Spike and slab priors

Similar to non-local priors, spike and slab priors allow for variable selection and testing of the hypotheses $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. Unlike non-local priors, spike and slab prior densities place significant probability into both hypotheses. In a regression context, the spike and slab prior (Mitchell and Beauchamp, 1988) takes the form

$$\beta_j|\gamma_j \sim (1-\gamma_j)\delta_0(\beta_j) + \gamma_j N(0, \tau^2), \tag{141}$$

$$\gamma_j \sim Bernoulli(\pi_0), \tag{142}$$

[17]For example, Johnson and Rossell (2012) note that if the covariate matrix $\boldsymbol{X}$ is not standardized, then it would be important to define an adaptive shrinkage parameter τ_j^2 for each $j = 1, ..., p$. In such a case, choice of each individual τ_j^2 for large p becomes inconvenient, if not infeasible.

for each $j = 1, ..., p$, where $\delta_0(\beta_j)$ is the Dirac delta function placing point mass at zero and γ_j are 0/1 (dummy) variables indicating whether column j of $\boldsymbol{X}$ is included in the regression or not. The mechanism with which this prior classifies predictors as important or not, is simple: when $\gamma_j = 1$ the prior for β_j is $N(0, \tau^2)$, that is, estimation is not restricted by the prior for reasonably large values of τ^2; when $\gamma_j = 0$ the prior becomes a point mass function concentrated at zero and it dominates the likelihood such that the posterior is also concentrates its mass at zero. The concept of variable selection is fully determined by the indicator random variables γ_j's. Samples from the posterior of each γ_j will be sequences of zeros and ones, and the posterior mean denotes the *posterior inclusion probability* of each predictor in the best model. For example, if we sample MCMC 10,000 draws and find that 2,000 times $\gamma_j = 1$, then the posterior mean is simply $2000/10000 = 0.2$ which translates into 20% posterior inclusion probability of predictor j. Barbieri and Berger (2004) show that the median probability model, that is, the model where only variables with probabilities larger than 0.5 are selected/retained, is optimal for prediction. O'Hara and Sillanpää (2009) suggest that as a variable selection mechanism such variable selection priors should work well up to cases where p is 10-15 times larger than n, but of course this proportion is only a rule of thumb that is heavily determined by the informativeness of the data and modeling choices.

The spike and slab prior belongs to the general class of hierarchical full-Bayes priors introduced earlier in this section, since it can be written in the form

$$\beta_j|\gamma_j \quad \sim \quad N\left(0, \tau^2\gamma_j\right). \tag{143}$$

If, in addition, we introduce a hyperprior distribution on τ^2 (e.g. inverse-gamma, see Ishwaran and Rao 2003), then the spike and slab prior is not only a hierarchical prior, but also belongs to the class of local-global shrinkage priors with global shrinkage parameter τ^2 and local shrinkage parameters γ_j. In signal processing and similar fields, the spike and slab is known as a "normal-Bernoulli" or "Gaussian-Bernoulli" prior.

A third parametric formulation of this particular spike and slab prior is due to Kuo and Mallick (1998). In their formulation the regression model with variable selection prior is written as

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{X}_j \gamma_j \beta_j + \boldsymbol{\varepsilon}, \tag{144}$$

where β_j is the coefficient on predictor j and γ_j is a 0/1 variable indicating whether predictor j is included in the model. This formulation is equivalent to the previous two, but it implies that the vector of indicators $\boldsymbol{\gamma}$ enters only via the likelihood and not through the (hierarchical) prior for $\boldsymbol{\beta}$. In the Kuo and Mallick (1998) formulation each β_j will simply have a typical Gaussian prior with variance τ^2. Notice that when $\gamma_j = 1$, β_j will be sampled from its posterior, but when $\gamma_j = 0$, β_j is not identified. In this case what happens is – as is the case with any unidentified parameter in a Bayesian setting (e.g. mutlicollinearity) – that β_j is sampled from

its prior. This lack of identification of β_j is not a problem, as what we care about is the joint effect $\gamma_j \times \beta_j$ and the fact that predictor j simply has to be removed whenever $\gamma_j = 0$. This detail means that in variable selection a-la Kuo and Mallick (1998) the posterior of β_j with $\gamma_j = 0$ will be equal to its normal prior, while the posterior of the same parameter under the spike and slab prior of equation (141) is a point mass at zero. Other than this (possibly minor) difference, Bayesian variable selection using all three forms presented above is conceptually and empirically comparable.

The class of spike and slab priors and its theoretical properties have been studied extensively in the literature; see Johnstone and Silverman (2004), Ishwaran and Rao (2005), Jiang (2006), Bogdan et al. (2011) and Castillo and van der Vaart (2012). From an applied scientist's point of view, the spike and slab prior is very versatile and can take numerous useful forms.[18] We next briefly review possible formulations of the spike and slab prior, and their implications for modeling coefficients and selecting variables in a linear regression. We finish this section with a discussion of some key computational aspects of this class of priors.

Tuning of parameters in the spike and slab prior

In the formulation in Equation 141 one only has to choose the variance parameter τ^2. This cannot be zero because the slab will become identical to the spike component, and it cannot become infinity because it would also be impossible to separate the spike from the slab component (remember from the previous section that Bayes factors with diffuse priors do not exist). Therefore, τ^2 has to be quite different from zero and not too large (e.g. $\tau^2 = 4$ is a reasonable choice). Of course one can use any of the hyperprior distributions already explored in the previous sections, e.g. the choice $\tau^2 \sim exponential(\lambda^2/2)$ will convert the slab into a Laplace prior. However, one should be careful not to overshrink the slab (e.g. by setting λ too large in the Laplace prior) because then the spike and slab will be indistinguishable and posterior inclusion probabilities will be meaningless.

A computationally more efficient formulation of the spike and slab prior (at least within an MCMC setting) is the one proposed by George and McCulloch (1993, 1997), where both the spike and slab distributions are continuous

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_0^2) + \gamma_j N(0,\tau_1^2), \tag{145}$$

where τ_0^2 is a "small" variance parameter (corresponding to the spike) and τ_1^2 is a "large" variance parameter (corresponding to the slab). In the limit, when $\tau_0^2 = 0$, the spike becomes the Dirac delta at zero, but for any other values of τ_0^2 close but different from zero the spike distribution is unable to shrink β_j exactly to zero. That is, this version of the spike and slab is appropriate for testing $H_0 : \beta_j \approx 0$ vs $H_1 : \beta_j \neq 0$, that is, it provides a soft thresholding rule.

[18] For example, Koop and Korobilis (2016) specify a spike and slab prior that is able to search for homogeneities in panel data. That is, the spike and slab prior is modified in order to test the hypothesis of the form $H_0 : \beta_i = \beta_j$ vs $H_1 : \beta_i \neq \beta_j$.

Chipman et al. (2001) provide the threshold value above (below) which a regression coefficient is classified as belonging to the slab (spike) component and is not shrunk (shrunk) to zero:

$$\Delta = \sqrt{\frac{\log\left(\frac{\tau_1^2}{\tau_0^2}\right)}{\frac{1}{\tau_1^2} - \frac{1}{\tau_0^2}}}. \tag{146}$$

Therefore, elicitation of τ_0^2, τ_1^2 becomes very important for variable selection in the George and McCulloch (1993) prior. Narisetty and He (2014) show that fixing these two variance hyperparameters may result in variable selection inconsistency, and propose values that are functions of n and p that ensure good performance of the prior when the data dimensions increase. Ishwaran and Rao (2005) set $\tau_0^2 = \tau^2$ and $\tau_1^2 = c\tau^2$ where $c >> 1$ and $\tau^2 \sim Inv - Gamma$, although τ^2 could also follow any of the hierarchical distributions defined previously, e.g. Horseshoe or Laplace. Früwirth-Schnatter and Wagner (2010) go one step further by motivating a mix-and-match strategy where τ_0^2 has a Laplace prior, while τ_1^2 has an inverse-gamma prior. More recently, Ročková and George (2018) showed that, under mild conditions, a spike and slab lasso prior produces posterior distributions that concentrate asymptotically around the true regression coefficients at nearly the minimax rate. In their formulation both the spike and the slab are based on Laplace distributions (represented as normal-exponential mixtures), with the spike distribution shrunk more aggressively than the slab distribution.

An important feature of variable selection priors is the prior on γ_j. As in Equation 142 this is typically Bernoulli with prior probability π_0, or equivalently a binomial prior for the full vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)'$. Unfortunately, the choice $\pi_0 = 0.5$ in a binomial prior is not uniform as it implies a prior expectation that half of the p predictors will be included in the final model. Therefore, in high-dimensional settings it is customary to set this parameter to a value closer to zero, e.g. $\pi_0 = 0.1$. If desired, a prior can be placed on this parameter and a conjugate choice is the beta distribution, that is, $\pi_0 \sim Beta(1, \alpha_0)$. The choice $\alpha_0 = 1$ makes this prior uniform, but in high-dimensional cases it will be preferable to set α_0 to become proportional to the number of predictors p. Note that in the presence of a beta hyperprior on π_0, it is not necessary to use indicator variables γ_j. For example, following Dunson et al. (2008) we can specify a spike and slab of the form[19]

$$\beta_j|\pi_0 \sim (1-\pi_0)\delta_0(\beta_j) + \pi_0 N(0, \tau^2), \tag{147}$$

$$\pi_0 \sim Beta(1, \alpha_0), \tag{148}$$

that provides a smoother mixture of the two components. (We can, of course, specify an equivalent formulation for the George and McCulloch (1993) continuous spike and slab formulation.) Finally, Carvalho et al. (2008) turn this latter formulation into a sparsity

[19] See also Korobilis (2013a,b, 2016) for related priors applied to econometric contexts such as dynamic regressions and vector autoregresions.

inducing variable selection prior by replacing Equation 148 with

$$\pi_0|\rho \sim (1-\rho)\delta_0(\pi_0) + \rho Beta(1, \alpha_0), \tag{149}$$

that is, a spike and slab prior for π_0. Finally, Yuan and Lin (2005) propose a prior for $\boldsymbol{\gamma}$ that accounts for correlation in predictors, such that if two predictors are highly correlated only one is included in the selected model. In their formulation they multiply the standard binomial prior for $\boldsymbol{\gamma}$ with the determinant of the Gram matrix of predictors, that is, $|\boldsymbol{X}'\boldsymbol{X}|$. High-correlated predictors have small $|\boldsymbol{X}'\boldsymbol{X}|$ and are discouraged from being selected. Such enhancements of the base spike and slab prior are important for variable selection, because marginal inclusion probabilities may be poor under high correlation. In particular, highly correlated predictors may be jointly selected often but each predictor only a small number of times.

Computation with spike and slab priors

Computation with spike and slab priors is as straightforward as is the case with most other hierarchical priors. Conditional on γ_j being either zero or one, the prior for β_j is either a point mass at zero or normal (in the representation of Mitchell and Beauchamp, 1988) or it is one of two normal components (in the representation of George and McCulloch, 1993). Therefore, conditional on γ_j, results for the normal linear model can be used. The same holds in the case where the components of the spike and slab are non-normal, rather they are Student-t, Laplace etc: as long a hierarchical prior structure is used and the prior can be written in conditionally normal form, derivation of conditional posteriors is straightforward.

Regarding posterior computation of γ_j's this usually has to be done element-by-element, that is, we need to derive γ_j conditional on $\boldsymbol{\gamma}_{-j}$ (the set $\boldsymbol{\gamma}$ with the j-th element removed).[20] However, in the case of the Mitchell and Beauchamp (1988) prior of Equation 141, the conditional posterior $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y})$ cannot be used to obtain samples from the posterior of γ_j. Intuitively, this is because when we sample $\gamma_j = 0$ then the prior for β_j is the Dirac delta function that puts infinite mass at zero. Therefore, in the next iteration $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, data)$ will give $\gamma_j = 0$ with probability one, meaning that the sampler will get stuck in a loop where the only possible outcome is $\beta_j = \gamma_j = 0$. This is not an issue in the continuous spike and slab prior of George and McCulloch (1993), since the spike is a continuous normal distribution and allows samples of β_j to be slightly different from zero.

To see this, let's derive $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y})$ in the case of the spike and slab prior of equations (141) - (143), which we rewrite for convenience

$$\beta_j|\gamma_j \sim (1-\gamma_j)\delta_0(\beta_j) + \gamma_j N(0, \tau^2), \tag{150}$$

$$\gamma_j \sim Bernoulli(\pi_0). \tag{151}$$

[20] For that reason, when the Gibbs sampler is used to sample from the conditional posterior of γ_j given $\boldsymbol{\gamma}_{-j}$, it is advisable in each Gibbs iteration to sample in random order j to avoid high autocorrelation of samples.

For simplicity, we do not introduce prior distributions on τ^2 and π_0, so we assume these are fixed and chosen by the researcher. Using Bayes theorem, the posterior of $\gamma_j = 0$ is

$$p(\gamma_j = 0|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) \propto p(\boldsymbol{y}|\gamma_j = 0, \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2)p(\beta_j|\gamma_j = 0)p(\gamma_j = 0). \tag{152}$$

In this decomposition, the first term is provided by the likelihood where we set the j-th element of $\boldsymbol{\beta}$ equal to zero (since $\gamma_j = 0$), regardless of what the sampled value β_j is in the previous iteration of the Gibbs sampler. This is a normal distribution with mean $\boldsymbol{X}\boldsymbol{\beta}^\star$, where $\boldsymbol{\beta}^\star$ is equal to β with the j-th element equal to zero, and variance σ^2. The second term is the prior for β_j under the restriction $\gamma_j = 0$, that is, the Dirac delta density $\delta_0(\beta_j)$. The last term is given simply by the Bernoulli prior for γ_j and it is equal to $(1 - \pi_0)$. Therefore, this posterior is:

$$p(\gamma_j = 0|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) \propto N_n(\boldsymbol{X}\boldsymbol{\beta}^\star, \sigma^2\boldsymbol{I}_n)\delta_0(\beta_j)(1 - \pi_0). \tag{153}$$

Using similar arguments, we have that

$$p(\gamma_j = 1|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) \propto N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)N(0, \tau^2)\pi_0. \tag{154}$$

Therefore, the conditional posterior of γ_j is

$$p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) \sim Bernoulli\left(\frac{N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)N(0, \tau^2)\pi_0}{N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)N(0, \tau^2)\pi_0 + N_n(\boldsymbol{X}\boldsymbol{\beta}^\star, \sigma^2\boldsymbol{I}_n)\delta_0(\beta_j)(1 - \pi_0)}\right). \tag{155}$$

Notice how the Dirac delta $\delta_0(\beta_j)$ enters the denominator term. If in the sampling process it happens to sample $\gamma_j = 0$, then $\beta_j = 0$ and any subsequent γ_j's will also be zero for ever. This is because once a $\beta_j = 0$ is observed, $\delta_0(\beta_j)$ becomes infinite and the ratio in the Bernoulli posterior is zero.

The solution to this problem is integration. That is, we need to remove dependence to β_j, and instead of the posterior $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y})$ we compute $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \sigma^2, \boldsymbol{y})$, that is, we integrate out β_j and condition only on $\boldsymbol{\beta}_{-j}$. Intuitively, because γ_j depends only to β_j through the spike and slab prior (i.e. it is independent to $\boldsymbol{\beta}_{-j}$), the ratio in the Bernoulli posterior of Equation 155 will only involve the densities $p(\boldsymbol{y}|\gamma_j = 0, \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2)$, $p(\boldsymbol{y}|\gamma_j = 1, \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}, \sigma^2)$ and $p(\gamma_j = 0)$, $p(\gamma_j = 1)$. The accompanying Technical document provides details of conditional posteriors under various forms of spike and slab prior distributions, including cases with more complex hiearchical layers such as the spike and slab lasso of Ročková and George (2018).

2.11 Monte Carlo study: Specification of spike and slab priors for variable selection

Consider a George and McCulloch (1993, 1997) type spike and slab prior

$$\begin{aligned}
\boldsymbol{y}|\sigma^2, \boldsymbol{\beta} &\sim N_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right), \\
\beta_j|\tau_{0j}^2, \sigma^2, \gamma_j = 0 &\sim N\left(0, \sigma^2\tau_{0j}^2\right), \text{ for } j = 1, \ldots, p, \\
\beta_j|\tau_{1j}^2, \sigma^2, \gamma_j = 1 &\sim N\left(0, \sigma^2\tau_{1j}^2\right), \text{ for } j = 1, \ldots, p, \\
\sigma^2 &\sim Inv - Gamma(a, b), \\
P(\gamma_j = 1) &= \pi_0, \text{ for } j = 1, \ldots, p, \\
\pi_0 &\sim Beta(c, d)
\end{aligned}$$

The conditional posteriors of $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}$, and π_0 are

$$\begin{aligned}
\boldsymbol{\beta}|\bullet &\sim N_p\left(\boldsymbol{V}\boldsymbol{X}^{'}\boldsymbol{y}, \sigma^2\boldsymbol{V}\right), \text{ where } \boldsymbol{V} = (\boldsymbol{D}^{-1} + \boldsymbol{X}^{'}\boldsymbol{X})^{-1}, \\
\sigma^2|\bullet &\sim Inv - Gamma\left(a + \frac{n}{2} + \frac{p}{2}, b + \frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\boldsymbol{Q}^{-1}\boldsymbol{\beta}\right]\right), \\
\gamma_j|\bullet &\sim Bern\left(\frac{\phi\left(\beta_j|0, \sigma^2\tau_{1j}^2\right)\pi_0}{\phi\left(\beta_j|0, \sigma^2\tau_{1j}^2\right)\pi_0 + \phi\left(\beta_j|0, \sigma^2\tau_{0j}^2\right)(1 - \pi_0)}\right), \text{ for } j = 1, \ldots, p, \\
\pi_0|\bullet &\sim Beta\left(c + \sum_{j=1}^{p}\gamma_j, d + \sum_{j=1}^{p}(1 - \gamma_j)\right)
\end{aligned}$$

where $\phi(\cdot|m, v)$ is the normal density with mean m and variance v and $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $\{(1 - \gamma_j)\tau_{0j}^2 + \gamma_j\tau_{1j}^2\}_{j=1}^{p}$.

2.11.1 SSVS-Lasso

Suppose we employ a Laplace density for the slab component

$$\tau_{1j}^2|\lambda_1^2 \sim Exponential\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, ..., p$$

and consider three different ways of defining priors for the spike component that are commonly used in practice, which we define as SSVS-Lasso 1-3.

In SSVS-Lasso-1, τ_{0j}^2 is fixed i.e. $\tau_{0j}^2 = c_1$ for some small $c_1 > 0$ and in SSVS-Lasso-2, it is proportional to the prior variance for the slab component i.e. $\tau_{0j}^2 = c_2\tau_{1j}^2$ for some small $c_2 > 0$. In both SSVS-Lasso-1 and 2, with the prior $\lambda_1^2 \sim Gamma(r_1, \delta_1)$, the prior variance for the

slab is updated according to

$$\lambda_1^2|\bullet \sim Gamma\left(\sum_{j=1}^{p}\gamma_j + r_1, \sum_{j=1}^{p}\tau_{1j}^2\gamma_j/2 + \delta_1\right)$$
$$1/\tau_{1j}^2|\bullet \sim IG\left(\sqrt{\lambda_1^2\sigma^2/\beta_j^2}, \lambda_1^2\right), \quad \text{for } j = 1, ..., p$$

In SSVS-Lasso-3, we place two separate Laplace densities on the components i.e.

$$\tau_{0j}^2|\lambda_0^2 \sim Exponential\left(\frac{\lambda_0^2}{2}\right), \quad \text{for } j = 1, ..., p,$$
$$\tau_{1j}^2|\lambda_1^2 \sim Exponential\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, ..., p$$

with $\lambda_0 \gg \lambda_1$ so that the density for $N(0, \sigma^2\tau_{0j}^2)$ is the "spike" and $N(0, \sigma^2\tau_{1j}^2)$ is the "slab". This is similar to the spike-and-slab Lasso in Ročková and George (2014) and Bai et al. (2021) [21]. The prior variances are updated according to

$$1/\tau_{0j}^2|\bullet \sim IG\left(\sqrt{\lambda_0^2\sigma^2/\beta_j^2}, \lambda_0^2\right), \quad \text{for } j = 1, ..., p,$$
$$1/\tau_{1j}^2|\bullet \sim IG\left(\sqrt{\lambda_1^2\sigma^2/\beta_j^2}, \lambda_1^2\right), \quad \text{for } j = 1, ..., p$$

Which specifications are appropriate in applications? In order to investigate this question, we consider simulation by generating data from the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. We let $n = 100$ and $\sigma^2 = 3$. We construct the true vector of slope parameters $\boldsymbol{\beta} = c\tilde{\boldsymbol{\beta}}$ by assigning values $\{1.5, -1.5, 2, -2, 2.5, -2.5\}$ to the first 6 elements of $\tilde{\boldsymbol{\beta}}$ and setting others to zero. We choose a constant $c > 0$ to achieve a desired level of signal-to-noise ratio [22]. The data matrix $\boldsymbol{X}$ is generated from the multivariate normal distribution with mean zero and covariance matrix being an identity matrix. The covariates are standardized for estimation. We examine different values of the number of covariates $p \in \{50, 100, 300\}$ and the signal-to-noise ratio $R_{pop}^2 \in \{0.4, 0.8\}$.

The analysis was repeated 100 times with new covariates and responses generated each time. For each, the metrics recorded were: the bias and MSE of the first 6 elements of the coefficients vectors, the number of false negatives (FN), the number of false positives (FP), and the number of true positives (TP). Posterior means were used as point estimates of the slope coefficients and the error variance. We utilize the post-processing approach of Li and Pati (2017) in order to categorize the covariates into signals and noises.

We compare performance of SSVS-Lasso 1-3 under different data generating processes.

[21] They propose an EM algorithm for estimation.

[22] In a general linear regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the signal-to-noise ratio (SNR) is defined as $SNR = \frac{||\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\beta}||^2}{\sigma^2}$ where σ^2 is the error variance and $\boldsymbol{\Sigma}_X$ is a $p \times p$ covariance matrix of $\boldsymbol{X}$. $||\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\beta}||^2 = \boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}$ measures the overall signal strength. A related quantity is R_{pop}^2, the population value of R^2, defined as $\frac{SNR}{1+SNR}$.

We fix $c = d = 1$ so that the prior on the inclusion probability θ is uniform. We also let $a = b = 0.1$. For SSVS-Lasso 1 and 2, the hyperparameters for the prior on λ_1^2 are fixed as $r_1 = 1$ and $\delta_1 = 1$. We let $c_1 = 10^{-4}$ for SSVS-Lasso-1, $c_2 = 10^{-4}$ for SSVS-Lasso-2, and $\lambda_0 = 20, \lambda_1 = 1$ for SSVS-Lasso-3. We also show results of Narisetty and He (2014) which is a two component mixture of normals prior with fixed prior variances and Kuo and Mallick (1998) which can be seen as a spike-and-slab prior with the spike component being a point mass at zero. Table 1 summarizes results.

Under a relatively strong signal i.e. $R^2_{pop} = 0.8$, generally speaking, SSVS-Lasso-3 and Narisetty-He outperform others in all measures, and this tendency becomes more apparent in high-dimensional case i.e. $p = 300$. Kuo-Mallick tends to have larger FPs than others. When the signal-to-ratio is lower, i.e. $R^2_{pop} = 0.4$, SSVS-Lasso-3 outperforms others in terms of bias/MSE of the signals and shows reasonable performance in other metrics.

	Bias	MSE	FN	FP	TP
$R^2_{pop} = 0.8, p = 50$					
SSVS-Lasso-1	0.14	0.02	0.63	0	5.4
SSVS-Lasso-2	0.16	0.03	0.70	0	5.3
SSVS-Lasso-3	0.11	0.01	0.53	0	5.4
Narisetty-He	0.09	0.01	0.31	0	5.6
Kuo-Mallick	0.11	0.02	0.49	0.04	5.5
$R^2_{pop} = 0.8, p = 100$					
SSVS-Lasso-1	0.23	0.06	0.96	0	5.1
SSVS-Lasso-2	0.28	0.09	1.15	0	4.8
SSVS-Lasso-3	0.13	0.02	0.53	0.05	5.4
Narisetty-He	0.10	0.01	0.39	0	5.6
Kuo-Mallick	0.29	0.13	1.07	20.8	4.9
$R^2_{pop} = 0.8, p = 300$					
SSVS-Lasso-1	0.52	0.29	1.91	21.0	4.0
SSVS-Lasso-2	0.56	0.33	0.98	39.6	5.0
SSVS-Lasso-3	0.32	0.12	1.44	9.1	4.5
Narisetty-He	0.22	0.08	1.95	0	4.0
Kuo-Mallick	0.53	0.30	0.09	91.1	5.9

(a) $R^2_{pop} = 0.8$.

	Bias	MSE	FN	FP	TP
$R^2_{pop} = 0.4, p = 50$					
SSVS-Lasso-1	0.12	0.02	1.1	3.4	4.8
SSVS-Lasso-2	0.12	0.02	1.1	4.9	4.8
SSVS-Lasso-3	0.10	0.01	1.1	5.4	4.8
Narisetty-He	0.16	0.03	3.4	2.6	2.6
Kuo-Mallick	0.11	0.02	1.2	8.8	4.8
$R^2_{pop} = 0.4, p = 100$					
SSVS-Lasso-1	0.15	0.03	0.8	11.5	5.1
SSVS-Lasso-2	0.17	0.03	0.7	15.9	5.2
SSVS-Lasso-3	0.12	0.02	1.1	17.6	4.8
Narisetty-He	0.18	0.04	2.2	14.9	3.7
Kuo-Mallick	0.28	0.13	2.9	29.4	3.0
$R^2_{pop} = 0.4, p = 300$					
SSVS-Lasso-1	0.24	0.06	0.78	50.3	5.22
SSVS-Lasso-2	0.26	0.07	0.71	60.4	5.29
SSVS-Lasso-3	0.19	0.04	0.88	56.9	5.12
Narisetty-He	0.24	0.06	0.91	82.6	5.09
Kuo-Mallick	0.21	0.05	0.76	93.1	5.24

(b) $R^2_{pop} = 0.4$.

Table 1: Average metrics over 100 repetitions for each of the approaches. Estimated error variance, and bias and MSE of the first 6 elements of the slope vector, and the numbers of False Negatives (FN), False Positives (FP), and True Positives (TP). The posterior means were used as point estimates. The post-processing method of Li and Pati (2017) was used to distinguish signals from noises. $n = 100$.

3 Bayesian Computation with hierarchical priors

We have established that hierarchical priors obey conditional structures that make derivation of conditional posteriors a straightforward business. As a consequence, the Gibbs sampler is the primary computational tool for variable selection and shrinkage problems using hierarchical priors. However, exactly because such full Bayes shrinkage estimators are mostly needed in high and ultra-high dimensions, the Gibbs sampler and related Monte Carlo-based methods become computational costly. In such cases there are numerous other strategies that allow for faster computation. These strategies include approximate methods for computing marginal posterior distributions, or iterative, non-sampling methods that approximate the posterior mode or mean. Many of these algorithms originate in computing science, where data dimensions have always been larger than traditional economic data sets. Currently, Bayesian computation in high-dimensional spaces – especially in the presence of hierarchical priors – is the topic of an expanding research agenda in mainstream statistics as well as in the field of machine learning. In this section, we summarize this vibrant research, focusing on both MCMC and fast approximate algorithms.

3.1 Brute-force/analytical algorithms

Analytical algorithms for hierarchical priors, in general, do not exist apart from a few special cases that can be fairly restrictive. In the context of estimating a normal mean θ (see our discussion of Efron and Morris, 1973 in subsection 1.1), Kahn and Raftery (1992) put uniform hyperpriors on the mean and variance hyperparameters of a normal prior distribution of θ. In order to obtain the posteriors of these hyperparameters they need to integrate θ, something they are able to do numerically since in their case only univariate integrals are involved on the support $[0, 1]$. In the context of a regression with spike and slab prior, Clyde (1999) shows that if the design is orthogonal (that is if the Gram matrix $\boldsymbol{X}'\boldsymbol{X} \propto I$) and the regression variance σ^2 is known, variable selection indicators $\boldsymbol{\gamma}$ can be obtained without resorting to Monte Carlo methods (either Gibbs sampler or Monte Carlo). Papaspiliopoulos and Rossell (2017) also derive an efficient non-sampling algorithm for Bayesian model averaging in regressions with a block diagonal design.[23] Their methods involve calculation of model probabilities and parameter estimates using one-dimensional numerical integration.

An interesting case that allows for approximately analytical posterior inference using hierarchical priors is provided in van den Boom et al. (2015a) and van den Boom et al. (2015b). These authors use rotation matrices to partition the regression model into a component explained by predictor X_j and all remaining predictors $\boldsymbol{X}_{(-j)}$. In particular, they split the regression into two components

[23] Examples of modeling scenarios with block-diagonal matrix $\boldsymbol{X}'\boldsymbol{X}$ include time-series regressions with time-varying parameters, and vector autoregressions written in "seemingly unrelated regressions" form; see subsection 4.1 for more details.

- one partition that is a regression of $n-1$ observations of a rotation of $\boldsymbol{y}$ on $\boldsymbol{X}_{(-j)}$ (i.e. dependence on X_j is removed), and
- one partition that is a regression of the remaining one observation of a separate rotation of $\boldsymbol{y}$ on X_j, conditional on $\boldsymbol{X}_{(-j)}$.

These authors use a non-shrinking natural conjugate prior in the first part of the rotated regression in order to obtain analytically an estimate of $\boldsymbol{\beta}_{(-j)}$ and σ^2. Then conditional on these estimates, they introduce in the second part a hierarchical shrinkage prior on β_j and derive analytically its posterior, since the regression variance is known using its estimate from the first partition. This procedure requires to repeat the rotation and partition of the regression model for each predictor j, $j = 1, ..., p$. The outcome is an analytical derivation of the posterior of each element β_j of $\boldsymbol{\beta}$ under a hierarchical prior that would otherwise require posterior simulation. This algorithm is of course approximate because it requires to obtain the posterior of β_j by integrating out the influence of each $\boldsymbol{\beta}_{(-j)}$ using a natural conjugate prior, rather than the same hierarchical prior used on β_j. Korobilis and Pettenuzzo (2019) extend this idea to various several hierarchical priors, including normal-Jeffrey's, spike and slab, and normal-gamma.

3.2 Gibbs sampler

We have already established that complex distributions (e.g. Student-t, Laplace, normal-half Cauchy) can be written in a conditionally conjugate form by using hierarchical representations. Depending on whether we also condition on the regression variance parameter, or not, we obtain the following two hierarchical prior formulations

$$\begin{array}{cccc} \underline{\text{Natural conjugate prior}} & & & \\ \boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}^2 & \sim & N_p\left(\mathbf{0}, \sigma^2 \boldsymbol{D}_\tau\right), \\ \boldsymbol{\tau}^2 & \sim & p(\boldsymbol{\tau}^2), \\ \sigma^2 & \sim & \frac{1}{\sigma^2}, \end{array} \qquad \begin{array}{cccc} \underline{\text{Independent prior}} & & & \\ \boldsymbol{\beta}|\boldsymbol{\tau}^2 & \sim & N_p\left(\mathbf{0}, \boldsymbol{D}_\tau\right), \\ \boldsymbol{\tau}^2 & \sim & p(\boldsymbol{\tau}^2), \\ \sigma^2 & \sim & \frac{1}{\sigma^2}, \end{array} \tag{156}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$ and depending on the structure of the distribution $p(\boldsymbol{\tau}^2)$ (which itself can be a hierarchical mixture of several distributions) we obtained the various interesting cases we explored so far.

Because of the conditional structure of the prior, posterior conditionals are easy to derive. For example, consider the case of the independent prior, then the joint posterior is of the form

$$p\left(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2 | \boldsymbol{y}\right) \propto p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right) p(\boldsymbol{\beta}|\boldsymbol{\tau}^2) p(\boldsymbol{\tau}^2) p(\sigma^2). \tag{157}$$

We can derive the conditional posterior of $\boldsymbol{\beta}$ as

$$p\left(\boldsymbol{\beta}|\boldsymbol{\tau}^2, \sigma^2, \boldsymbol{y}\right) \propto p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right) p(\boldsymbol{\beta}|\boldsymbol{\tau}^2), \tag{158}$$

because $p(\boldsymbol{\tau}^2)$ and $p(\sigma^2)$ are constants when conditioning on $\boldsymbol{\tau}^2, \sigma^2$ (because they do not involve the random variable $\boldsymbol{\beta}$). The prior distribution $p(\boldsymbol{\beta}|\tau^2)$ is normal and, due to the modeling assumptions, $p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right)$ is also normal. As a result, the conditional posterior for $p\left(\boldsymbol{\beta}|\boldsymbol{\tau}^2, \sigma^2, \boldsymbol{y}\right)$ is identical to the conditional posterior under the non-hierarchical version of the same prior (see for example Gelman et al., 2013). Similarly, the conditional posterior for σ^2 becomes

$$p\left(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\tau}^2, \boldsymbol{y}\right) \propto p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right) p(\sigma^2), \tag{159}$$

which is also identical to the case of the non-hierarchical independent normal-inverse gamma prior. Finally, the conditional posterior for τ becomes

$$p\left(\boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}\right) \propto p(\underline{\boldsymbol{\beta}}|\boldsymbol{\tau}^2)p(\boldsymbol{\tau}^2). \tag{160}$$

The data density $p\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right)$ does not contain information about $\boldsymbol{\tau}$ so it becomes a constant. Instead the "model" for $\boldsymbol{\tau}^2$ is provided by the density $p(\boldsymbol{\beta}|\boldsymbol{\tau}^2)$ where $\boldsymbol{\beta}$ are observed "data" (fixed to their sampled values), since in this conditional posterior the only random variable is $\boldsymbol{\tau}^2$.

It becomes apparent that because of the conditional structure of hierarchical priors, for the vast majority of hierarchical priors we have common formulas for the conditional posteriors of $\boldsymbol{\beta}$ and σ^2, while the formulation of the conditional posterior of $\boldsymbol{\tau}^2$ will depend on how complicate its prior is. Under the natural conjugate prior the conditional posteriors are of the form

$$\begin{array}{rcc} \underline{\text{Conditional Posteriors (Natural conjugate prior)}} & & \\ \boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}^2, \sigma^2, \boldsymbol{y} & \sim & N_p\left(\boldsymbol{V}\boldsymbol{X}'y, \sigma^2\boldsymbol{V}\right), \\ \boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y} & \sim & p(\boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}), \\ \sigma^2|\boldsymbol{\beta}, \boldsymbol{y} & \sim & Inv-Gamma\left(\frac{n+p}{2}, \frac{1}{2}(\boldsymbol{\Psi} + \boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta})\right), \end{array} \tag{161}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)^{-1}$ and $\boldsymbol{\Psi} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Under the independent prior the posteriors are of the form

$$\begin{array}{rcc} \underline{\text{Conditional Posteriors (Independent prior)}} & & \\ \boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}^2, \sigma^2, \boldsymbol{y} & \sim & N_p\left(\boldsymbol{V}\boldsymbol{X}'y/\sigma^2, \boldsymbol{V}\right), \\ \boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y} & \sim & p(\boldsymbol{\tau}^2|\boldsymbol{\beta}, \boldsymbol{y}), \\ \sigma^2|\boldsymbol{\beta}, \boldsymbol{y} & \sim & Inv-Gamma\left(\frac{n}{2}, \frac{1}{2}\boldsymbol{\Psi}\right), \end{array} \tag{162}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \boldsymbol{D}_\tau^{-1}\right)^{-1}$ and $\boldsymbol{\Psi} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Notice how σ^2 affects the posterior of $\boldsymbol{\tau}^2$ in the conjugate prior case, while σ^2 doesn't show up in the posterior of $\boldsymbol{\tau}^2$. See the Technical Document for derivations.

A Gibbs sampler will cycle through equations (161) or equations (162), obtaining a sample of parameters conditional on all others, until a large enough sample from the posterior of each parameter is available. Results in Pal and Khare (2014) and Khare and Hobert (2013) establish,

for a large class of hyper-prior distributions $p(\boldsymbol{\tau}^2)$, that the above Gibbs sampler is ergodic and has the joint posterior $p\left(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2 | \boldsymbol{y}\right)$ as its stationary density. Despite its ergodicity, the basic Gibbs sampler for models with hierarchical priors may suffer from slow mixing and convergence to the desired posterior. The conditional structure of a hierarchical prior implies a long chain of dependence of $\boldsymbol{\beta}$ on $\boldsymbol{\tau}^2$ and its hyper-priors. For example, the representation of the Horseshoe prior suggested by Makalic and Schmidt (2016) as a hierarchical mixture of a normal distribution and four inverse gamma hyper-prior distributions (see subsection 2.7) is one example where slow mixing might become a serious issue. For that reason, in the case of the Horseshoe in particular, several authors propose to use more efficient slice sampling schemes, some of which we explore in detail in the accompanying Technical Document. Another disadvantage of the Gibbs sampler is the fact that sampling becomes cumbersome as the dimension p of covariates increases. In the next we explore various approaches for speeding up MCMC and for dealing with convergence issues, particularly in the case where p is large or even $p >> n$.[24]

Fast sampling from Normal posteriors

In high-dimensional settings with p large, the most cumbersome step in a Gibbs sampler for the linear regression model with hierarchical priors is sampling from the p-variate normal conditional posterior distribution of $\boldsymbol{\beta}$. This step involves an inversion of precision matrix $\boldsymbol{Q} = \boldsymbol{V}^{-1} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)$ (in the case of the natural conjugate prior) or $\boldsymbol{Q} = \boldsymbol{V}^{-1} = \left(\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \boldsymbol{D}_\tau^{-1}\right)$ (in the case of the independent prior) in order to obtain the posterior covariance matrix $\boldsymbol{V}$.[25] Next, the Cholesky decomposition of $\boldsymbol{V}$ is needed in order to sample from the desired normal distribution. While the inversion step can be sped up (e.g. by using Woodbury's identity), standard built-in algorithms (in various programming languages) for obtaining the Cholesky decomposition of a $p \times p$ matrix have a worst asymptotic complexity (measured in flops) of $\mathcal{O}(p^3)$. Therefore, simple sampling from a normal posterior is deemed to become computationally cumbersome, if not infeasible, as p increases.

Rue (2001) provides a precision-based sampler in order to obtain samples from a normal distribution efficiently when the precision matrix $\boldsymbol{Q} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)$ is known. Due to the fact that the Bayesian conditional posterior of $\boldsymbol{\beta}$ (ignoring σ^2, e.g. assume it is fixed to value 1) is of the form $\boldsymbol{\beta}|\bullet \sim N(\boldsymbol{V}\boldsymbol{X}'\boldsymbol{y}, \boldsymbol{V})$, the procedure proposed by Rue (2001) takes the following form

[24] All the approaches we explore propose novel ways of sampling from the parameter posterior under the Gibbs sampling scheme. However, given a specific algorithm, the ability of the programming language to handle large matrices is also important. This is illustrated, for example, in Matusevich et al. (2016), where in the context of Bayesian variable selection they combine array database management systems (DBMS) and R processing capabilities, allowing R to handle large matrices without running out of RAM.

[25] Note that in the case of the natural conjugate prior without hierarchical structure, the matrix $\boldsymbol{D}_\tau$ is known (calibrated by the researcher), as is the data information $\boldsymbol{X}'\boldsymbol{X}$. In this case, one could calculate and invert $\boldsymbol{Q}$ once, outside the loop of the Gibbs sampler. However, the presence of a hierarchical prior for $\boldsymbol{\tau}^2$ means that the matrix $\boldsymbol{D}_\tau$ changes values in each Gibbs iteration. Therefore, $\boldsymbol{V}^{-1}$ should be computed and inverted in each iteration (regardless of whether we used the independent prior on $\boldsymbol{\beta}$ or not).

- Compute the lower Cholesky factorization $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}'$
- Generate $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$
- Set $\boldsymbol{v} = \boldsymbol{L}^{-1}(\boldsymbol{X}'\boldsymbol{y})$
- Set $\boldsymbol{\mu} = \boldsymbol{L}'^{-1}\boldsymbol{v}$
- Set $\boldsymbol{u} = \boldsymbol{L}'^{-1}\boldsymbol{Z}$
- Set $\boldsymbol{\beta} = \boldsymbol{\mu} + \boldsymbol{u}$

It is trivial to show that $E(\boldsymbol{\beta}) = \boldsymbol{\mu} = (\boldsymbol{L}'^{-1}\boldsymbol{L}^{-1})\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{L}\boldsymbol{L}')^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{V}\boldsymbol{X}'\boldsymbol{y}$ and $cov(\boldsymbol{\beta}) = cov(\boldsymbol{\mu}+\boldsymbol{u}) = \boldsymbol{L}'^{-1}cov(\boldsymbol{Z})\boldsymbol{L}^{-1} = \boldsymbol{L}'^{-1}\boldsymbol{L}^{-1} = \boldsymbol{V}$, which means that the above procedure provides valid samples from the desired normal distribution. The main feature of this algorithm is that it requires to invert the Cholesky factor of $\boldsymbol{Q}$, instead of inverting $\boldsymbol{Q}$ itself to obtain $\boldsymbol{V}$. While the worst case complexity of this algorithm is also $\mathcal{O}(p^3)$, it provides high efficiency gains in certain classes of models, e.g. when the Gram matrix $\boldsymbol{X}'\boldsymbol{X}$ is block-diagonal (assuming the prior variance $\boldsymbol{D}_\tau$ is diagonal or at most block-diagonal of similar structure).

More recently, Bhattacharya et al. (2016) proposed an efficient algorithm that makes full use of Woodbury matrix inversion lemma in the context of generating normal variates from a distribution of the form $\boldsymbol{\beta}|\bullet \sim N(\boldsymbol{V}\boldsymbol{X}'\boldsymbol{y}, \boldsymbol{V})$ (again for simplicity, ignore σ^2). Their algorithm takes the following form

1. Sample $\boldsymbol{\eta} \sim N_p(\boldsymbol{0}, \boldsymbol{D}_\tau)$ and $\boldsymbol{\delta} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$
2. Set $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{\eta} + \boldsymbol{\delta}$
3. Set $\boldsymbol{w} = (\boldsymbol{X}\boldsymbol{D}_\tau\boldsymbol{X}' + \boldsymbol{I}_n)^{-1}[\boldsymbol{y} - \boldsymbol{v}]$
4. Set $\boldsymbol{\beta} = \boldsymbol{\eta} + \boldsymbol{D}\boldsymbol{X}'\boldsymbol{w}$

It is also easy to show that the sample of $\boldsymbol{\beta}$ comes from the desired normal distribution with mean $\boldsymbol{V}\boldsymbol{X}'\boldsymbol{y}$ and variance $\boldsymbol{V}$. Note that this algorithm requires inversion of the $n \times n$ matrix $(\boldsymbol{X}\boldsymbol{D}_\tau\boldsymbol{X}' + \boldsymbol{I}_n)^{-1}$, while sampling directly from the normal posterior requires inversion of the $p \times p$ matrix $\boldsymbol{Q}$. Therefore, step 3 in this algorithm only becomes efficient for $p >> n$. The main efficiency gains in this algorithm stem from the fact that it requires to sample from two normal distributions with diagonal covariance matrices (a p-variate distribution with covariance $\boldsymbol{D}_\tau$ and an n-variate distribution with identity covariance). Generating uncorrelated normal draws is much more efficient than sampling directly from the p-variate normal posterior of $\boldsymbol{\beta}$ using the full covariance matrix $\boldsymbol{V}$. In particular, the worst-case complexity (asymptotic upper bound) of this algorithm is $\mathcal{O}(n^2p)$, that is, it is only linear in p. Therefore, for $n > p$ this algorithm will perform worse than the algorithm of Rue (2001) since the term n^2 will dominate, but this algorithm shines in the $p > n$ case where it can offer some dramatic improvements in computation times.

Note that for very large p, computation of $\boldsymbol{X}\boldsymbol{D}_\tau\boldsymbol{X}'$ in step 3. above will become cumbersome. In such ultra high-dimensional cases, Johndrow et al. (2020) provide an approximate version of Bhattacharya et al. (2016). This involves removing "irrelevant" columns of $\boldsymbol{X}$ such that the above product can be computed using a significantly smaller number of algorithmic operations.

Scalable Gibbs

The standard form of the Gibbs sampler for the linear regression model with hierarchical prior contains three blocks as in Equation 161. There is one block for each set of parameters, namely $\boldsymbol{\beta}$, $\boldsymbol{\tau}^2$ and σ^2. In the context of hierarchical priors, Rajaratnam et al. (2019) propose to sample $(\boldsymbol{\beta}, \sigma^2)$ in one block. Their proposed Gibbs sampler is more efficient, as reducing the number of blocks to sample from, also reduces correlation among draws from parameter posteriors. When working with the natural conjugate form of a hierarchical prior, the *scalable Gibbs* algorithm requires only to sample from $p((\boldsymbol{\beta}, \sigma^2)|\boldsymbol{\tau}^2, \boldsymbol{y})$ and $p(\boldsymbol{\tau}^2|(\boldsymbol{\beta}, \sigma^2), \boldsymbol{y})$. The first joint distribution for $(\boldsymbol{\beta}, \sigma^2)$ can be approximated by first sampling from $p(\sigma^2|\boldsymbol{\tau}^2, \boldsymbol{y})$ (notice the lack of dependence on $\boldsymbol{\beta}$) and subsequently from $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}^2, \boldsymbol{y})$. The scalable Gibbs algorithm has the following form

$$\begin{cases} \sigma^{-2}|\boldsymbol{\tau}^2, \boldsymbol{y} & \sim & Gamma\left(\frac{n-1}{2}, \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{V}\boldsymbol{X}')\boldsymbol{y}/2\right) \\ \boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}^2, \boldsymbol{y} & \sim & N_p\left(\boldsymbol{V}\boldsymbol{X}'\boldsymbol{y}, \boldsymbol{V}\right) \\ \boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y} & \sim & p(\boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}), \end{cases}, \tag{163}$$

where $\boldsymbol{V}$ has the same definition as in Equation 161. The proof why the posterior for σ^2, after integrating out $\boldsymbol{\beta}$, has the form shown above, can be found in the Appendix of Rajaratnam et al. (2019); see also Pal et al. (2017).

Skinny Gibbs

In the context of the spike and slab prior with continuous spike and slab distributions, Narisetty et al. (2018) propose an efficient sampling scheme that also separates the posterior into a mixture distribution with independent components. We remind that the continuous spike and slab prior (George and McCulloch, 1993; Narisetty and He, 2014) can be written in matrix form as

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim (\boldsymbol{I} - \boldsymbol{\Gamma})N_p(\boldsymbol{0}, \tau_0^2\boldsymbol{I}) + \Gamma N_p(\boldsymbol{0}, \tau_1^2\boldsymbol{I}), \tag{164}$$

where $\boldsymbol{\Gamma} = diag(\boldsymbol{\gamma})$, or more compactly

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(\boldsymbol{0}, \boldsymbol{D}_\gamma), \tag{165}$$

where $\boldsymbol{D}_\gamma = diag\left((1-\gamma_1)^2\tau_0^2 + \gamma_1^2\tau_1^2, ..., (1-\gamma_p)^2\tau_0^2 + \gamma_p^2\tau_1^2\right)$. The conditional posterior under this prior is of the form

$$\boldsymbol{\beta}|\bullet \sim N_p(\boldsymbol{V}\boldsymbol{X}'\boldsymbol{y}/\sigma^2, \boldsymbol{V}), \tag{166}$$

where $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \boldsymbol{D}_\gamma^{-1}\right)^{-1}$ and $|\bullet$ denotes conditioning on other parameters in the model as well as the data. Computing $\boldsymbol{V}$ requires an inversion as well as obtaining the Cholesky decomposition in order to sample from the p-dimensional normal posterior. As we already saw, when p is large these operations can be extremely cumbersome.

The skinny Gibbs algorithm of Narisetty et al. (2018) solves this issue by splitting the conditional posterior into two independent components, an active (A) and an inactive (I) and sample $\boldsymbol{\beta}$ as the union of the following conditionals

$$\boldsymbol{\beta}_A|\bullet \sim N_{p_A}(\boldsymbol{V}_A\boldsymbol{X}_A'\boldsymbol{y}/\sigma^2, \boldsymbol{V}_A), \tag{167}$$

$$\boldsymbol{\beta}_I|\bullet \sim N_{p_I}(\boldsymbol{0}, \boldsymbol{V}_I), \tag{168}$$

where $\boldsymbol{\beta}_A$ is the p_A-dimensional vector of elements of $\boldsymbol{\beta}$ corresponding to $\gamma_j = 1$, and $\boldsymbol{\beta}_I$ are the remaining $p_I = p - p_A$ elements that correspond to $\gamma_j = 0$. In the first posterior the covariance matrix is $\boldsymbol{V}_A = \left(\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \frac{1}{\tau_1^2}\boldsymbol{I}\right)^{-1}$, while in the second posterior $\boldsymbol{V}_I = \left(n + \frac{1}{\tau_0^2}\right)^{-1}\boldsymbol{I}_{p_I}$. In sparse settings we would expect to find that $p_I >> p_A$, meaning that the bulk of the elements of $\boldsymbol{\beta}$ would be sampled as restricted elements $\boldsymbol{\beta}_I$. This means that we can sample very efficiently p_I coefficients from a normal posterior with diagonal convariance matrix, and sample the remaining p_A elements from a normal posterior with a full covariance matrix.

Of course, in practical situations it is expected that the two sets of coefficients, $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_I$, will be correlated with each other. As a result, sampling the full vector $\boldsymbol{\beta}$ from two independent conditional posteriors could leave us with a significant approximation error. For that reason, Narisetty et al. (2018) add a "compensation term" in the conditional posterior of $\boldsymbol{\gamma}$ that accounts for the approximation involved in sampling the coefficients $\boldsymbol{\beta}$. This term ensures that the skinny Gibbs converges to a stationary distribution, while keeping computational complexity minimal. Narisetty et al. (2018) show that the skinny Gibbs posterior possesses strong selection consistency property.

Orthogonal Data Augmentation

In the conext of variable selection using spike and slab priors, Ghosh and Clyde (2011) note that when the design matrix $\boldsymbol{X}$ is orthogonal, a stochastic search using MCMC can become very efficient when p is large. Similarly for penalized likelihood problems (whether Bayesian or not), orthogonal designs can be very efficient and result in consistent variable selection regardless of how large p is relative to n. The proposal of Ghosh and Clyde (2011) is to augment the $n \times p$ correlated design matrix $\boldsymbol{X}$ with an $n_a \times p$ matrix $\boldsymbol{X}_a$, such that the $(n + n_a) \times p$ "complete"

design matrix

$$\boldsymbol{X}_c = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}_a \end{bmatrix}, \tag{169}$$

has orthogonal columns, that is, $\boldsymbol{X}_c'\boldsymbol{X}_c = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_a'\boldsymbol{X}_a = \boldsymbol{W}$, where $\boldsymbol{W} = diag(w_1, ..., w_p)$ is a diagonal matrix with $w_j > 0$. Furthermore, we have the restriction that the augmented data matrix $\boldsymbol{X}_a$ has real entries, and $\boldsymbol{X}_a'\boldsymbol{X}_a$ must be a positive semidefinite symmetric matrix. Ghosh and Clyde (2011) select a diagonal matrix $\boldsymbol{W}$ which then implies the value of $\boldsymbol{X}_a$ from the orthogonality condition $\boldsymbol{A} \equiv \boldsymbol{X}_a'\boldsymbol{X}_a = \boldsymbol{W} - \boldsymbol{X}'\boldsymbol{X}$. $\boldsymbol{X}_a$ can be obtained as the symmetric matrix square root of $\boldsymbol{A}$, thus, ensuring that its entries are real. Ghosh and Clyde (2011) discuss in detail ways of choosing $\boldsymbol{W}$; for example, since in most variable selection settings the columns of $\boldsymbol{X}$ are typically standardized to have unit norm or variance, one can set $w_1 = ... = w_p = w$, such that choice of $\boldsymbol{W}$ collapses into a choice of a scalar w.

Once $\boldsymbol{X}_a$ has been specified, we can estimate the augmented orthogonal regression model

$$\boldsymbol{y}_c = \boldsymbol{X}_c\boldsymbol{\beta} + \boldsymbol{\varepsilon}_c, \tag{170}$$

where $y_c = [\boldsymbol{y}', \boldsymbol{y}_a']'$ with $\boldsymbol{y}_a$ latent data which we need to sample and $\boldsymbol{\varepsilon}_c \sim N_{(n+n_a)}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. In the most general case, Ghosh and Clyde (2011) consider a hierarchical variable selection prior, which combines a spike at zero with a component that is Student-t, obtained via normal-inverse-gamma mixture (see subsection 2.2)

$$\begin{aligned} \beta_j|\sigma^2, \gamma_j, \tau_j^2 &\sim N(0, \sigma^2\gamma_j\tau_j^2), && (171)\\ \tau_j^2 &\sim Gamma(\alpha/2, \alpha/2), && (172)\\ \gamma_j &\sim Bernoulli(\pi_0), && (173)\\ \sigma^2 &\sim \frac{1}{\sigma^2}, && (174) \end{aligned}$$

for all $j = 1, ..., p$. For $\alpha = 1$ this prior becomes a heavy-tailed Cauchy distribution of the form $\beta_j \sim C(0, \sigma^2\gamma_j)$. Ghosh and Clyde (2011) propose a Gibbs sampling scheme the iterates over the following conditional distributions

1. $p\left((\sigma^2, \boldsymbol{y}_a)\,|\boldsymbol{\gamma}, \boldsymbol{y}\right)$
2. $p(\gamma_j|\sigma^2, \boldsymbol{\tau}^2, \boldsymbol{y}_c)$ for $j = 1, ..., p$
3. $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\gamma}\boldsymbol{y}_c)$
4. $p(\boldsymbol{\tau}^2|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y}_c)$

All the conditional distributions have standard forms, and details can be found in Ghosh and Clyde (2011).

3.3 Approximate computation with hierarchical priors

Approximate inference methods typically involve optimization algorithms for approximating posterior moments (typically the mean or the mode, and the variance) instead of sampling from the full posterior distribution. Then one can proceed their analysis using only these moments, treating the Bayesian estimator similar to a frequentist point estimator. This approach to Bayesian inference has been popularized in computing science, e.g. in estimation of high-dimensional Bayes networks, where large datasets is the norm and MCMC inference is extremely costly. An obvious critique of approximate Bayesian inference of this sort, is that we can't fully take into account parameter uncertainty by characterizing the full parameter posterior distribution. However, it is in high-dimensional and the so-called ultra high-dimensional models (see Shin et al., 2018), that shrinkage and sparsity via a hierarchical prior is necessary. Since analytical results for most classes of hierarhical priors are not available, and Monte Carlo sampling is costly in very high dimensions, it is not surprising that approximate methods have become very popular. Additionally, at the conceptual level, we saw that only Bayesian posterior medians/modes correspond to penalized likelihood estimators, but there are not always good theoretical guarantees for the tails of the posterior.[26] Finally, the concept of sparsity is indeed more interpretable in a setting with point estimates of coefficients, that is, it is more straightforward to test and interpret $H_0 : \beta_j = 0$ when β_j is approximated with a point estimate rather than when we have thousands of samples from the full posterior of β_j.

All these reasons lead us to review some of the most popular approximate methods for posterior inference. While many of these methods have been popularized and used extensively in computing science often without theoretical justifications, investigation of their theoretical/asymptotic properties is currently a topic of vivid research in mainstream statistics.

3.3.1 Variational Bayes

Variational Bayes is probably the most prominent of algorithms, at least when it comes to high-dimensional inference using hierarchical priors. The idea behind this class of algorithms is rather simple, and under certain assumptions (what we will call *mean-field approximation* later) variational Bayes algorithms can be fairly simple to implement by practitioners who are familiar with the Gibbs sampler. For notational simplicity, assume we have a vector of parameters $\boldsymbol{\theta}$ with support Θ and data $\boldsymbol{D}$, resulting to the posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{D})}. \tag{175}$$

Assume that this posterior is intractable because the data density $p(\boldsymbol{D}|\boldsymbol{\theta})$ is complex (e.g. it is a highly nonlinear function, or it has unidentified parameters), or because the prior $p(\boldsymbol{\theta})$ is complex (non-conjugate), or because θ is high-dimensional (in which case the posterior is a

[26] See for example our discussion of the results of Castillo et al. (2015) in the Bayesian lasso prior.

high-dimensional function), or due to combinations of the above cases. In such settings, MCMC is not only computationally costly, but it can also become numerically unstable/unreliable.[27]

The idea behind variational Bayes is to introduce a family of simpler, approximate densities over the parameters $\boldsymbol{\theta}$ which is denoted by the set $\mathbb{Q}$. The objective is to find a member of the family $q(\boldsymbol{\theta}) \in \mathbb{Q}$ that is as close as possible to the true posterior. "Closeness" is measured by the Kullback-Leibler (KL) divergence, and the optimal density $q^{\star}(\boldsymbol{\theta})$, among all densities $q(\boldsymbol{\theta})$, is the one that minimizes this criterion:

$$q^{\star}(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in\mathbb{Q}}{\arg\min}\, KL\left(q\left(\boldsymbol{\theta}\right)||p\left(\boldsymbol{\theta}|\boldsymbol{D}\right)\right). \tag{176}$$

In the above formula the KL measure on the RHS is defined as

$$\begin{aligned} KL &= E_{q(\boldsymbol{\theta})}\left(\log(q(\boldsymbol{\theta}))\right) - E_{q(\boldsymbol{\theta})}\left(\log(p(\boldsymbol{\theta}|\boldsymbol{D}))\right) && (177)\\ &= E_{q(\boldsymbol{\theta})}\left(\log(q(\boldsymbol{\theta}))\right) - E_{q(\boldsymbol{\theta})}\left(\log\left(\frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{D})}\right)\right) && (178)\\ &= E_{q(\boldsymbol{\theta})}\left(\log(q(\boldsymbol{\theta}))\right) + \log(p(\boldsymbol{D})) - E_{q(\boldsymbol{\theta})}\left(\log\left(p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)\right), && (179) \end{aligned}$$

where
all expectations are w.r.t $q(\boldsymbol{\theta})$, for example, $E_{q(\boldsymbol{\theta})}\left(\log(p(\boldsymbol{\theta}|\boldsymbol{D}))\right) = \int_{\boldsymbol{\theta}\in\Theta} q(\boldsymbol{\theta})\log(p(\boldsymbol{\theta}|\boldsymbol{D}))d\boldsymbol{\theta}$. In the last equation we have used the fact that $E_{q(\boldsymbol{\theta})}\left(\log(p(\boldsymbol{D}))\right) = \log(p(\boldsymbol{D}))$ since $p(\boldsymbol{D})$ does not involve $\boldsymbol{\theta}$. For the same reason the variational Bayes minimization problem is equal to minimizing the difference between the first and the third terms in equation (179). We can also solve this equation for the log marginal likelihood $\log(p(\boldsymbol{D}))$ to show that

$$\log(p(\boldsymbol{D})) = KL\left(q\left(\boldsymbol{\theta}\right)||p\left(\boldsymbol{\theta}|\boldsymbol{D}\right)\right) + ELBO, \tag{180}$$

where we define evidence lower bound (ELBO) to be the quantity $ELBO = +E_{q(\boldsymbol{\theta})}\left(\log\left(p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)\right) - E_{q(\boldsymbol{\theta})}\left(\log(q(\boldsymbol{\theta}))\right)$. The ELBO has this name exactly because it is a lower bound for the log evidence (marginal data density). This is because in the equation above the KL divergence term is non-negative, such that $\log(p(\boldsymbol{D})) \geq ELBO$. Therefore, the optimal $q^{\star}(\boldsymbol{\theta})$ can be found by equivalently maximizing the ELBO criterion function.

The CAVI algorithm

When latent variables are present, optimizations such as maximizing the ELBO criterion can be implemented using the popular expectation-maximization (EM) algorithm, where the complete log likelihood is computed (E-step) and then it is maximized (M-step). However, in Bayesian inference all parameters are latent (random) variables and as the optimization problem above involves optimizing over the functional $q(\boldsymbol{\theta})$ and not $\boldsymbol{\theta}$ itself, the EM algorithm is not appropriate. Variational inference instead requires to choose the variational family of

[27] For example, consider the case of a high-dimensional nonlinear regression with highly correlated predictors. In this example, unless modifications are introduced such as adaptive tuning, mixing and convergence of standard Gibbs sampler algorithms with hierarchical priors will tend to be slow.

distributions $\mathbb{Q}$ and then maximize the ELBO. In most cases this can be done iteratively, with certain schemes that resemble the EM algorithm (but are not identical to EM), and convergence is guaranteed to a local maximum and if the likelihood is log-concave then to a global maximum. The simplest algorithm for maximizing the ELBO is called Coordinate Ascent Variational Inference (CAVI). Its simplicity comes at the cost of certain simplifying assumptions. The first one is that $\mathbb{Q}$ must strictly belong to the exponential family of distributions (e.g. the normal satisfies this condition, but the Student's t does not). The second restriction is the use of the mean-field approximation that postulates that the proposed posterior distribution $q(\boldsymbol{\theta})$ can be decomposed into M independent groups of the form

$$q(\boldsymbol{\theta}) = \prod_{m=1}^{M} q_m(\boldsymbol{\theta}_m), \tag{181}$$

where the groups could either have $\boldsymbol{\theta}_m$ being a scalar or a vector. The estimated variational posteriors will be independent, meaning that the mean-field approximation/factorization implies that $\boldsymbol{\theta}_m$ will be a-posteriori uncorrelated with $\boldsymbol{\theta}_k$, for $k \neq m$ and $k, m = 1, ..., M$. This assumption in several modeling settings can be harmless, but in several others it can become harmful – we discuss this issue in detail later when we examine variational Bayes inference in a linear regression with variable selection prior.

Under the assumption of the mean field approximation it can be shown (Blei et al., 2017) that the optimal densities $q_m(\boldsymbol{\theta}_m)$ satisfy

$$\log\left(q_m(\boldsymbol{\theta}_m)\right) \propto E_{q_{(-m)}(\boldsymbol{\theta}_{(-m)})}\left(\log\left(p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)\right), \tag{182}$$

where $E_{q_{(-m)}(\boldsymbol{\theta}_{(-m)})}()$ means that the expectation is w.r.t all variational densities except $q_m(\boldsymbol{\theta}_m)$. Broadly speaking this formula says that in order to optimize w.r.t. $q_m(\boldsymbol{\theta}_m)$ we need to evaluate the posterior under the assumption that all other parameters $\boldsymbol{\theta}_{(-m)}$ are fixed to their posterior expectation (posterior mean). We would obviously need to iterate through Equation 182 for each $m = 1, ..., M$ keeping all other parameters fixed to their posterior means, but it can be shown that such iteration results in increasing the ELBO criterion. If the ELBO hasn't changed from one iteration to the next, the algorithm has converged. This criterion resembles the EM algorithm that converges when the value of the likelihood in subsequent iterations is approximately similar. The fact that $\boldsymbol{\theta}_m$ is updated conditional on fixing all other parameters $\boldsymbol{\theta}_{(-m)}$ makes variational Bayes resemble Gibbs sampling inference – despite the fact that there is no sampling involved. We next derive a CAVI algorithm for a linear regression model with variable selection prior, in order to clearly demonstrate how the mean-field approximation is applied and how the functions in Equation 182 look like.

A variational Bayes approach to variable selection

This subsection follows closely the analysis of Ormerod et al. (2017), and the reader should consult this paper for extensive discussion and proofs. Consider the Kuo and Mallick (1998)

regression we explored in subsection 2.10 and is of the form

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, && (183)\\
\boldsymbol{\varepsilon} &\sim N(\mathbf{0}_{n\times 1}, \sigma^2 \boldsymbol{I}_n), && (184)\\
\boldsymbol{\beta} &\sim N_p(\mathbf{0}_{p\times 1}, \boldsymbol{D}), && (185)\\
\sigma^2 &\sim Inv - Gamma(a_0, b_0), && (186)\\
\gamma_j &\sim Bernoulli(\pi_0), \quad j = 1, ..., p, && (187)
\end{aligned}$$

where $\boldsymbol{\Gamma} = diag(\gamma_1, ..., \gamma_p)$ and $\boldsymbol{D}$ is a diagonal prior covariance matrix, e.g. $\boldsymbol{D} = c \times \boldsymbol{I}_p$ for some constant c. Therefore, according to the information above the joint prior is decomposed into $p\left(\boldsymbol{\beta}, \{\gamma_j\}_{j=1}^p, \sigma^2\right) = p(\sigma^2)\prod_{j=1}^p p(\beta_j) \times p(\gamma_j)$ meaning that all parameters are a-priori uncorrelated. Such choices are both conceptually and practically fine, first because we don't have prior information on how parameters are correlated, and second because we can construct very powerful shrinkage and variable selection algorithms based on these forms. It is important for the posterior to allow the parameters to be correlated, as this posterior correlation will come from information in the data likelihood. We discussed previously that the mean-field factorization implies that some groups of parameters will be uncorrelated a-posteriori. Ormerod et al. (2017) look into three different ways of applying the mean field factorization, based on how we want to define the groups $m = 1, ..., M$, and their implications for posterior inference. These factorizations are the following

$$\begin{aligned}
\text{(A):}\quad q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) &= q(\boldsymbol{\beta}, \boldsymbol{\gamma})q(\sigma^2), && (188)\\
\text{(B):}\quad q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) &= q(\sigma^2)\prod_{j=1}^{p} q(\beta_j, \gamma_j), && (189)\\
\text{(C):}\quad q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) &= q(\boldsymbol{\beta})q(\sigma^2)\prod_{j=1}^{p} q(\gamma_j). && (190)
\end{aligned}$$

The factorization (A) means that application of formula Equation 182 to the set of parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ given σ^2 gives

$$\log(q(\boldsymbol{\beta}, \boldsymbol{\gamma})) \propto \lambda\boldsymbol{\gamma} - \frac{1}{2}\boldsymbol{\beta}'\left(\kappa\boldsymbol{\Gamma}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\beta} + \boldsymbol{D}^{-1}\right)\boldsymbol{\beta} + \kappa\boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{X}'\boldsymbol{y}, \tag{191}$$

or, similarly, that

$$q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \exp\left\{\lambda\boldsymbol{\gamma} - \frac{1}{2}\boldsymbol{\beta}'\left(\kappa\boldsymbol{\Gamma}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\beta} + \boldsymbol{D}^{-1}\right)\boldsymbol{\beta} + \kappa\boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{X}'\boldsymbol{y}\right\}, \tag{192}$$

where $\lambda = \log\left(\frac{\pi_0}{1-\pi_0}\right)$ and $\kappa = E_q\left(1/\sigma^2\right)$. Therefore, we can easily obtain the conditional

variational density of $\boldsymbol{\beta}$ and the (marginal) variational density of $\boldsymbol{\gamma}$ as

$$q\left(\boldsymbol{\beta}|\boldsymbol{\gamma}\right) \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{V}_{\boldsymbol{\gamma}}\right), \tag{193}$$

$$q\left(\boldsymbol{\gamma}\right) \propto \int q(\boldsymbol{\beta}, \boldsymbol{\gamma}) d\boldsymbol{\beta} = |\boldsymbol{V}_{\boldsymbol{\gamma}}|^{1/2} \exp\left[\lambda\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\gamma}}' \boldsymbol{V}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\gamma}}\right] \tag{194}$$

where $\boldsymbol{V}_{\boldsymbol{\gamma}} = \left(\kappa \boldsymbol{\Gamma} \boldsymbol{X}' \boldsymbol{X} \boldsymbol{\Gamma} \boldsymbol{\beta} + \boldsymbol{D}^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{V}_{\boldsymbol{\gamma}} \boldsymbol{\Gamma} \boldsymbol{X}' \boldsymbol{y}$. In the second equation we only have the kernel of $q\left(\boldsymbol{\gamma}\right)$, but we can easily normalize this to integrate to one by dividing with the sum of the density of all possible combinations of $\boldsymbol{\gamma}$ (which is 2^p in the regression with p covariates). In order to derive the marginal variational posterior density of $\boldsymbol{\beta}$ we need to integrate out the $\boldsymbol{\gamma}$, which is easily done as these are binary indicators. This marginal density is of the form

$$q(\boldsymbol{\beta}) = \sum_{\boldsymbol{\gamma} \in 0,1^p} q(\boldsymbol{\gamma}) N\left(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{V}_{\boldsymbol{\gamma}}\right), \tag{195}$$

which is a combinatorial sum over all 2^p outcomes for the vector $\boldsymbol{\gamma}$. This sum can be evaluated in finite time only for small p. However, for small p there are other numerous analytical algorithms that can be used, for example, we can use a g-prior and obtain marginal likelihoods analytically for all 2^p models, in which case there is no point in using variational Bayes. Therefore, this mean-field factorization is not useful.

The mean-field factorization/approximation (B), which was used in Carbonetto and Stephens (2012), provides a scalable variational Bayes algorithm where the $q(\beta_j)$ can be estimated independently and efficiently (by means of parallelization) for each $j = 1, ..., p$. However, posterior variances of these regression coefficients will tend to be underestimated exactly because of this assumption of posterior independence. Unless the predictors in $\boldsymbol{X}$ are uncorrelated (which is not realistic for economic data, and for large-p settings), the bias in posterior variances can be substantial.

The most reasonable case, which is the choice of Ormerod et al. (2017), is case (C). Under this factorization, one full implementation of the iterations in Equation 182 looks like this:

1. $q(\boldsymbol{\beta}) = N\left(\boldsymbol{\mu}, \boldsymbol{V}\right)$
   where $\boldsymbol{V} = \left(\kappa(\boldsymbol{X}'\boldsymbol{X}) \odot \boldsymbol{\Omega} + \boldsymbol{D}^{-1}\right)^{-1}$ and $\boldsymbol{\mu} = \kappa \boldsymbol{V}\left(\boldsymbol{\Pi} \boldsymbol{X}' \boldsymbol{y}\right)$.

2. $q(\sigma^2) = Inv - Gamma\left(a, b\right)$
   where $b = b_0 + \frac{1}{2}\left[||\boldsymbol{y}||^2 - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\Pi}\boldsymbol{\mu} + tr\left\{\left(\boldsymbol{X}'\boldsymbol{X} \odot \boldsymbol{\Omega}\right)\left(\boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{V}\right)\right\}\right]$ and $a = a_0 + n/2$.
   The posterior mean of σ^{-2} is, thus, $\kappa = \frac{a}{b}$.

3. $q(\gamma_j) = Bernoulli(\pi_j)$,
   where $\pi_j = \frac{exp(\eta_j)}{1+exp(\eta_j)}$ with $\eta_j = \log\left(\frac{\pi_0}{1-\pi_0}\right) - \frac{\kappa}{2}\left(\mu_j^2 + V_{j,j}\right)||\boldsymbol{X}_j||^2 + \kappa\left[\mu_j \boldsymbol{X}_j' \boldsymbol{y} - \boldsymbol{X}_j' \boldsymbol{X}_{(-j)} \boldsymbol{\Pi}_{(-j)}\left(\boldsymbol{\mu}_{(-j)}\mu_j + \boldsymbol{V}_{(-j),j}\right)\right]$.

In the equations above we have used some matrices vectors/matrices that are based on π_j, namely $\boldsymbol{\pi} = (\pi_1, ..., \pi_p)'$, $\boldsymbol{\Pi} = diag(\boldsymbol{\pi})$ and $\boldsymbol{\Omega} = \boldsymbol{\pi}\boldsymbol{\pi}' + \boldsymbol{\Pi}(\boldsymbol{I} - \boldsymbol{\Pi})$. The symbol $\odot$ denotes the

Hadamard product. We used the notations that for a general matrix $\boldsymbol{A}$, $\boldsymbol{A}_j$ is the jth column of $\boldsymbol{A}$, $\boldsymbol{A}_{(-j)}$ is $\boldsymbol{A}$ with the jth column removed, $A_{i,j}$ is the (i,j)th entry of $\boldsymbol{A}$, $\boldsymbol{A}_{(-i),j}$ is the vector corresponding to the jth column of $\boldsymbol{A}$ with the ith component removed.

The formulas look quite similar to the conditional posteriors in the Kuo and Mallick (1998) Gibbs samplers, but there is no sampling involved. Instead, when the variational posterior mean and variance of $\boldsymbol{\beta}$ are calculated, σ^{-2} is fixed to its posterior mean κ and the same for γ (the vector $\boldsymbol{\pi}$ and its variants, i.e. $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$). Given that in the very first iteration κ and $\boldsymbol{\pi}$ will be initialized to some random values, a convergence period is required until we end up with final estimates of the posterior moments of all parameters.

Further readings

There are numerous papers on variational Bayes inference in computing science problems, for example, natural language processing (text analytics) and Bayesian networks. In statistics and machine learning there has been a consistent effort to establish consistency and other properties of variational Bayes estimates; see for example Giordano et al. (2018) andWang and Blei (2019). With regards to high-dimensional regression and hierarchical priors, the contributions of Carbonetto and Stephens (2012), Ormerod et al. (2017) and Neville et al. (2014) are an excellent starting point. Koop and Korobilis (2018) provide a variational Bayes algorithm for a dynamic spike and slab prior for models featuring time-varying parameters and stochastic volatility (see also next section).

3.3.2 EM algorithm

We discussed previously how the expectation-maximization (EM) algorithm is not appropriate for variational Bayes inference, since there we are looking to find the "best" density function of $\boldsymbol{\theta}$, rather than a point estimate of our parameters $\boldsymbol{\beta}$. However, the EM algorithm can be used to find the posterior mode of $p(\boldsymbol{\theta}|\boldsymbol{D})$, an inference method known as maximum a-posteriori (MAP) inference. The mode of the posterior under a diffusing (flat) prior distribution is identical to the maximum likelihood estimate, while the MAP estimate under a hierarchical prior corresponds to a penalized likelihood estimator. Therefore, MAP inference – which was also popularized in computing science – can be thought of as a bridge between Bayesian and maximum/penalized likelihood inferences where it combines the strengths of both approaches.

There are numerous implementations of MAP inference using the EM algorithm but, unlike the Gibbs sampler, in many cases algorithms are model-specific and cannot generalize easily. With regards to variable selection and shrinkage we indicatively mention the key contributions of Caron and Doucet (2008), Figueiredo (2003) and Griffin and Brown (2011). A notable recent contribution is the EM variable selection (EMVS) of Ročková and George (2014). These authors adopt a setting (likelihood and prior) that is identical to George and McCulloch (1993) but they use the EM algorithm as a means of lowering the computational burden of Markov-Chain Monte Carlo methods when estimating posterior distributions over subsets of potential predictors.

3.3.3 Other approximate algorithms

There are several other algorithms for approximate high-dimensional inference. These include parallel MCMC, Hamiltonian Monte Carlo, Approximate Bayesian Computation (ABC), Expectation propagation, and Message Passing. A review of all these classes of algorithms can be found in Korobilis and Pettenuzzo (2020). A few representative works relying on such algorithms are Dehaene and Barthelmé (2018), Kim and Wand (2016), Korobilis (2021), Liu et al. (2019), Wainwright and Jordan (2008) and Zou et al. (2016).

3.4 Monte Carlo exercise: Conjugate vs independent hierarchical priors

Should we be using a conditional or unconditional hierarchical prior (see Equation 156)? In order to investigate this question, we consider simulation by generating data from the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. We let $n = 100$ and $\sigma^2 = 3$. We construct the true vector of slope parameters $\boldsymbol{\beta} = c\tilde{\boldsymbol{\beta}}$ by assigning values $\{1.5, -1.5, 2, -2, 2.5, -2.5\}$ to the first 6 elements of $\tilde{\boldsymbol{\beta}}$ and setting others to zero. We choose a constant $c > 0$ to achieve a desired level of signal-to-noise ratio [28]. The data matrix $\boldsymbol{X}$ is generated from the multivariate normal distribution with mean zero and covariance matrix being an identity matrix. The covariates are standardized for estimation. We examine different values of the number of covariates $p \in \{50, 100, 300\}$ and the signal-to-noise ratio $R^2_{pop} \in \{0.4, 0.8\}$.

We consider three shrinkage priors (1) student-t, (2) lasso, and (3) horseshoe and compare performance under the conditional and independent priors. The analysis was repeated 100 times with new covariates and responses generated each time. For each, the metrics recorded were: the estimated value of σ^2, the bias and MSE of the first 6 elements of the coefficients vectors, the number of false negatives (FN), the number of false positives (FP), and the number of true positives (TP). Posterior means were used as point estimates of the slope coefficients and the error variance. We utilize the post-processing approach of Li and Pati (2017) in order to categorize the covariates into signals and noises.

Table 2 summarizes the results. Panel (a) shows results when the signal is relatively strong i.e. $R^2_{pop} = 0.8$. Both conjugate and independent priors do well in terms of TPs and FNs. However, there are some notable differences. First, the error variance tends to be underestimated when conjugate priors are used. This was in fact pointed out by Moran et al. (2019). Intuitively, conjugate priors implicitly add p "pseudo-observations" to the posterior (compare Equation 161 with Equation 162) which can result in underestimations of σ^2 when $\boldsymbol{\beta}$ is sparse. Second, the independent priors tend to have larger bias and MSE of the signals under the high-dimensional case (i.e. $p = 300$). Park and Casella (2008) point out that independent priors can induce bi-modality of the posterior on the slope coefficients. This can make the posterior distributions for $\boldsymbol{\beta}$ more spread than in the conjugate case. We also

[28] In a general linear regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the signal-to-noise ratio (SNR) is defined as $SNR = \frac{||\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\beta}||^2}{\sigma^2}$ where σ^2 is the error variance and $\boldsymbol{\Sigma}_X$ is a $p \times p$ covariance matrix of $\boldsymbol{X}$. $||\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{\beta}||^2 = \boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}$ measures the overall signal strength. A related quantity is R^2_{pop}, the population value of R^2, defined as $\frac{SNR}{1+SNR}$.

see that independent priors have larger FPs when $p = 300$, which could be a result of this. Panel (b) shows results under relatively weak signal i.e. $R^2_{pop} = 0.4$. We see that all methods face difficulty with distinguishing signals with noise (see FNs, FPs, and TPs) and have large bias and MSEs, compared to the case with $R^2_{pop} = 0.8$. However, the general findings on the difference between conjugate and independent priors are the same: conjugate priors tend to underestimate σ^2 while independent priors tend to have higher bias and MSE of the signals when p is large compared to n. We encourage researchers to be aware of these issues when choosing priors and to conduct sensitivity checks.

	$\hat{\sigma}^2$	Bias	MSE	FN	FP	TP
$R^2_{pop} = 0.8, p = 50$						
Student-t (conjugate)	1.77	0.19	0.05	0.10	0	5.9
Bayesian Lasso (conjugate)	1.96	0.18	0.05	0.06	0	5.9
Horseshoe (conjugate)	2.35	0.19	0.05	0.07	0	5.9
Student-t (independent)	2.99	0.19	0.05	0.09	0	5.9
Bayesian Lasso (independent)	2.86	0.18	0.05	0.05	0	6.0
Horseshoe (independent)	2.93	0.19	0.05	0.08	0	5.9
$R^2_{pop} = 0.8, p = 100$						
Student-t (conjugate)	0.42	0.34	0.18	0.43	0.01	5.5
Bayesian Lasso (conjugate)	0.69	0.28	0.12	0.18	0	5.8
Horseshoe (conjugate)	1.05	0.26	0.11	0.19	0	5.8
Student-t (independent)	2.01	0.30	0.15	0.35	0.01	5.6
Bayesian Lasso (independent)	2.02	0.25	0.10	0.15	0	5.8
Horseshoe (independent)	2.28	0.26	0.11	0.19	0	5.8
$R^2_{pop} = 0.8, p = 300$						
Student-t (conjugate)	0.35	0.35	0.24	0.41	2.79	5.6
Bayesian Lasso (conjugate)	0.66	0.78	0.71	0.74	2.86	5.2
Horseshoe (conjugate)	0.92	0.66	0.63	0.46	12.0	5.5
Student-t (independent)	3.88	1.52	2.43	0.03	68.6	5.9
Bayesian Lasso (independent)	2.59	1.42	2.12	0.01	68.9	5.9
Horseshoe (independent)	3.27	1.43	2.17	0.01	61.1	5.9

(a) $R^2_{pop} = 0.8$.

	$\hat{\sigma}^2$	Bias	MSE	FN	FP	TP
$R^2_{pop} = 0.4, p = 50$						
Student-t (conjugate)	1.65	0.19	0.06	0.77	0.82	5.2
Bayesian Lasso (conjugate)	1.78	0.19	0.06	0.72	0.54	5.3
Horseshoe (conjugate)	2.16	0.19	0.06	0.72	0.62	5.3
Student-t (independent)	2.99	0.19	0.06	0.61	1.22	5.4
Bayesian Lasso (independent)	2.86	0.19	0.06	0.67	0.59	5.4
Horseshoe (independent)	2.96	0.20	0.06	0.76	0.49	5.2
$R^2_{pop} = 0.4, p = 100$						
Student-t (conjugate)	0.35	0.35	0.20	1.41	18.7	4.6
Bayesian Lasso (conjugate)	0.52	0.29	0.14	1.11	15.5	4.7
Horseshoe (conjugate)	0.90	0.27	0.12	0.92	14.5	5.1
Student-t (independent)	1.85	0.30	0.14	0.81	22.8	5.2
Bayesian Lasso (independent)	1.89	0.26	0.11	0.56	17.1	5.4
Horseshoe (independent)	2.20	0.27	0.11	0.42	17.0	5.6
$R^2_{pop} = 0.4, p = 300$						
Student-t (conjugate)	0.28	0.46	0.26	0.99	35.6	5.0
Bayesian Lasso (conjugate)	0.47	0.50	0.29	0.53	49.6	5.4
Horseshoe (conjugate)	0.75	0.55	0.34	0.33	53.6	5.6
Student-t (independent)	3.50	0.65	0.46	0.34	70.6	5.6
Bayesian Lasso (independent)	2.28	0.65	0.45	0.21	71.1	5.7
Horseshoe (independent)	2.79	0.65	0.45	0.23	66.1	5.7

(b) $R^2_{pop} = 0.4$.

Table 2: Average metrics over 100 repetitions for each of the approaches. Estimated error variance, and bias and MSE of the first 6 elements of the slope vector, and the numbers of False Negatives (FN), False Positives (FP), and True Positives (TP). The posterior means were used as point estimates. The post-processing method of Li and Pati (2017) was used to distinguish signals from noises. $n = 100$.

4 Bayesian shrinkage and variable selection beyond linear regression

So far we explored variable selection in the high-dimensional linear regression, also known as "large p, small n" regression. This setting is already flexible enough, as there are various cases of generalized linear models (GLMs) that have conditionally linear forms. The purpose of this section is to demonstrate that there is an even larger list of models where hierarchical priors have immediate applicability. In particular, we explore key applications of hierarchical shrinkage and variable selection priors in vector autoregressions, factor models, time-varying parameter models, high-dimensional confounder selection in models for treatment effects, and Bayesian quantile regression. This list is far from exhaustive[29] and its only purpose is to illustrate how Bayesian computation simplifies high-dimensional inference in unconventional settings.

4.1 Vector autoregressions

The most popular working model for economists using time series variables is the vector autoregression (VAR). VARs are used for the joint modeling of the dynamics of many macroeconomic and financial time series, $\boldsymbol{Y}$, allowing analysts and policy-makers to answer questions regarding dynamic responses of variable $\boldsymbol{Y}_i$ to a shock in some other variable $\boldsymbol{Y}_j$, $i \neq j$. This is a very important tool especially when variable $\boldsymbol{Y}_j$ is controlled by the policy-maker. For example, the central bank controls the short-term interest rate as well as other quantities related to monetary policy, while government controls taxes and fiscal policy in general. Due to the fact that availability of time series observations for macroeconomic and (low-frequency) financial data is limited[30], estimation of macroeconomic VARs by and large relies on Bayesian shrinkage priors. Additionally, parameters in VAR models proliferate at a polynomial rate as the dimensions of the model increases. In univariate linear regression settings, a model with twice as many exogenous predictors has twice as many parameters to estimate. There is not such an analogy in VARs where all variables are endogenous and each variable (and its lagged terms) affects all other variables in the system.

Unlike our previous notation, consider time series observations $t = 1, ..., T$ and an n-dimensional vector of variables $\boldsymbol{Y}_t$, that is, n denotes the number of variables of interest (and not the number of observations anymore) with $\boldsymbol{Y} = [\boldsymbol{Y}_1', ..., \boldsymbol{Y}_T']'$ is a $T \times n$ data matrix. The

[29] For example, one application of Bayesian shrinkage and selection that we do not cover in this section, but is of importance in statistics and in finance, is high-dimensional covariance matrix estimation and selection, see Wang and Pillai (2013) as an indicative example. Another topic we won't cover here, but is becoming increasingly very important in statistics and econometrics, is Bayesian additive regression trees (BART). For an up-to-date review of the topic see Hill et al. (2020).

[30] Especially in countries other than the US, where statistical agencies might have available only a handful of decades of data; e.g. euro area time series typically begin in 1995 or 1999.

VAR model for $\boldsymbol{Y}_t$ with p lags, also denoted as VAR(p), is of the form

$$\boldsymbol{Y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{Y}_{t-1} + \boldsymbol{A}_2 \boldsymbol{Y}_{t-2} + ... + \boldsymbol{A}_p \boldsymbol{Y}_{t-p} + \boldsymbol{E}_t, \tag{196}$$

where $\boldsymbol{c}$ is an $n \times 1$ vector of intercepts, $\boldsymbol{A}_i$ are $n \times n$ matrices of lagged terms for each $i = 1, ..., p$, and $\boldsymbol{E}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ an $n \times n$ symmetric, semi-positive definite covariance matrix. The VAR is a heavily parametrized model: it has $(1 + np)n$ coefficients $\boldsymbol{B} = [\boldsymbol{c}, \boldsymbol{A}_1, ..., \boldsymbol{A}_p]$, plus another $n(n+1)/2$ unique elements in the covariance matrix $\boldsymbol{\Sigma}$. For example, the largest VAR model specified in Koop et al. (2019) has $n = 129$ and $p = 13$ which implies that the total number of parameters is in excess of $200,000$.

Vector autoregressions are effectively linear regression models with parameter matrix $\boldsymbol{B} = [\boldsymbol{c}, \boldsymbol{A}_1, ..., \boldsymbol{A}_p]$ and data matrix $\boldsymbol{X}_t = [\boldsymbol{1}, \boldsymbol{Y}_{t-1}, ..., \boldsymbol{Y}_{t-p}]$, such that application of hierarchical priors for shrinkage and variable selection is fairly straightforward. The task of sampling from the conditional posterior of the regression coefficients $\boldsymbol{B}$ can be further simplified if the VAR is written in *seemingly unrelated regressions (SUR)*[31] form

$$vec(\boldsymbol{Y}) = (\boldsymbol{I} \otimes \boldsymbol{X}) vec(\boldsymbol{B}) + vec(\boldsymbol{\varepsilon}), \tag{197}$$

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{198}$$

where $vec(\bullet)$ is the operator that stacks the columns of a matrix into a single column vector. That way, $\boldsymbol{y} = vec(\boldsymbol{Y})$ is a $Tn \times 1$ vector where the first T elements are the observations of the first variable, the next T rows correspond to observations of the second variable, and so on up to variable n. The measurement matrix $\boldsymbol{Z} = (\boldsymbol{I} \otimes \boldsymbol{X})$ is a block-diagonal matrix with the $T \times (1 + np)$ matrix $\boldsymbol{X}$ repeating on its diagonal n times. The formulation above is observationally identical to the one in Equation 196, since there are no new parameters or data introduced, but it has the benefit that VAR parameters show up as the $(1 + np)n \times 1$ vector $\boldsymbol{b} = vec(\boldsymbol{B})$. Therefore, the SUR form in Equation 198 is identical to a univariate regression model, even though this model has both many covariates but also many observations ($\boldsymbol{y}$ and $\boldsymbol{Z}$ both have Tn rows, instead of T rows in a univariate regression). Therefore, it is straightforward to define any hierarchical prior we desire for the vector of VAR parameters $\boldsymbol{b} = vec(\boldsymbol{B})$ and derive conditional posteriors, despite the fact that Gibbs sampling might become quite cumbersome as the dimension n of the VAR increases.[32]

In large n cases and when shrinkage on the covariance matrix $\boldsymbol{\Sigma}$ is needed, Carriero et al. (2019) and Koop et al. (2019) propose to estimate the VAR equation-by-equation. For example, in the formulation of Koop et al. (2019) one can write Equation 196 as

$$\boldsymbol{Y}_t = \boldsymbol{B}\boldsymbol{X}_t + \boldsymbol{P}\boldsymbol{V}_t, \tag{199}$$

[31] For a thorough and accessible introduction to Bayesian inference in VARs see Koop and Korobilis (2010).
[32] See for example, Korobilis (2013b, 2016) and Koop and Korobilis (2016).

where $\boldsymbol{P}$ is a lower triangular matrix with ones on its main diagonal (unitriangular) that satisfies the LDL-decomposition $\boldsymbol{\Sigma} = \boldsymbol{PDP'}$. In this case $V_t \sim N(\boldsymbol{0}, \boldsymbol{D})$ where $\boldsymbol{D}$ is a diagonal matrix with variance elements d_{ii}^2 on its main diagonal, $i = 1, ..., n$. This formulation is equivalent to Equation 196 because $\boldsymbol{PV}_t \sim N(\boldsymbol{0}, \boldsymbol{PDP'}) = N(\boldsymbol{0}, \boldsymbol{\Sigma}) \overset{d}{=} \boldsymbol{E}_t$. Since by construction $\boldsymbol{P}$ is invertible, with $\boldsymbol{P}^{-1}$ also a unitriangular matrix, we can write

$$\begin{aligned} \boldsymbol{P}^{-1}\boldsymbol{Y}_t &= \boldsymbol{P}^{-1}\boldsymbol{B}\boldsymbol{X}_t + \boldsymbol{V}_t \Rightarrow & (200) \\ (\boldsymbol{I} + \widetilde{\boldsymbol{P}}^{-1})\boldsymbol{Y}_t &= \boldsymbol{\Gamma}\boldsymbol{X}_t + \boldsymbol{V}_t \Rightarrow & (201) \\ \boldsymbol{Y}_t &= \boldsymbol{\Gamma}\boldsymbol{X}_t - \widetilde{\boldsymbol{P}}^{-1}\boldsymbol{Y}_t + \boldsymbol{V}_t, & (202) \end{aligned}$$

where $\boldsymbol{\Gamma} = \boldsymbol{P}^{-1}\boldsymbol{B}$, and we have split $\boldsymbol{P}^{-1}$ into an identity matrix and a lower triangular matrix $\widetilde{\boldsymbol{P}}^{-1}$ by means of the equation $\boldsymbol{P}^{-1} = \boldsymbol{I} + \widetilde{\boldsymbol{P}}^{-1}$. In Equation 202 we have a VAR on $\boldsymbol{Y}_t$ where the covariance matrix elements in the original covariance matrix show up as contemporaneous elements of $\boldsymbol{Y}_t$ itself on the right-hand side in the term $-\widetilde{\boldsymbol{P}}^{-1}\boldsymbol{Y}_t$. Notice that in matrix form this is a nonlinear system as $\boldsymbol{Y}_t$ shows up both on the left hand side and the right hand side of Equation 202. However, exactly because $\widetilde{\boldsymbol{P}}^{-1}$ is lower triangular and $\boldsymbol{V}_t$ has a diagonal covariance matrix $\boldsymbol{D}$, equation-by-equation estimation is feasible. In particular, each VAR equation i depends on lags of all other equations and contemporaneous terms in the previous $i - 1$ equations. This means that estimation of the VAR collapses to estimation of n independent univariate models, such that specification of hierarchical priors and MCMC estimation are also straightforward. The additional benefit from this approach is that hierarchical priors can be specified to the elements of $\widetilde{\boldsymbol{P}}^{-1}$, thus leading to shrinkage or sparse estimation of the original VAR covariance matrix $\boldsymbol{\Sigma}$, similar to the methodology of Smith and Kohn (2002). More details of this approach can be found in Koop et al. (2019), while variants of this approach have been proposed in Baumeister et al. (2020), Carriero et al. (2019), and Korobilis and Pettenuzzo (2019).

4.2 Factor model shrinkage and selection

Factor models have a long history in econometrics, and an even longer history in statistics and psychology/psychometrics. For that reason, while there are some popular formulations across different literatures, there are also different variations based on the data and applications. We first establish some key results for the specific case of the so-called static factor model, and subsequently we review some of the most popular factor models used in economics and finance. We end our discussion with strategies for Bayesian shrinkage and variable selection in this class of models.

Consider an $n \times 1$ vector of economic variables $\boldsymbol{X}_t$ observed over $t = 1, ..., T$ (without loss of generality t can measure time series, but it can also be observations on individuals or other cross-sectional units). The dimension n can be inconveniently high[33], such that unrestricted

[33]This description includes both the ultra high-dimensional case where n can be in the order of thousands, or

estimation of models (e.g. linear regressions) using the data $\boldsymbol{X}_t$ is infeasible. Our target is to estimate a lower-dimensional $k \times 1$ vector ($k << n$) of latent variables (factors), that summarizes as much as possible the information contained in $\boldsymbol{X}_t$. For that reason we define the following multivariate model

$$\boldsymbol{X}_t = \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \tag{203}$$

where $\boldsymbol{\Lambda}$ is an $n \times k$ matrix of parameters, $\boldsymbol{F}_t$ are the latent variables and $\boldsymbol{\varepsilon}$ is a disturbance term. This is not a regression model, as both $\boldsymbol{\Lambda}$ and $\boldsymbol{F}_t$ are latent. For simplicity, we follow Lopes and West (2004) and make the assumption that $\boldsymbol{F}_t \sim N_k(\boldsymbol{0}, \boldsymbol{I})$, although we can allow the factors to have a more general covariance matrix such that they are correlated with each other. For the disturbance term we assume $\boldsymbol{\varepsilon}_t \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$.

While the model in Equation 203 looks like a linear regression – in which case application of hierarchical priors would be straightforward – this is not the case due to the fact that all terms on the right-hand side of the equation are latent. In many instances, researchers in economics, finance and other disciplines replace $\boldsymbol{F}$ with the first k principal components (PCs). PCs are nonparametric estimates of the factors, that is, they are only approximate estimators of the true factors implied by the likelihood of the model in Equation 203. Plugging in the place of the latent factors (parameters) $\boldsymbol{F}$ the PC estimates turns the factor model into a regression model and inference is simplified. Conditional on the principal component estimates, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ can be estimated simply via least squares, but also standard Bayesian methods for multivariate regression can be used. The benefit of this two-step approach is that it is simple, both conceptually and computationally, and that principal component estimates always provide a sensible fit. However, in many more complex settings (e.g. macroeconomic dynamic factor models or financial factor models with stochastic volatility), the PC only provide a rough approximation, and it might be preferable to use the likelihood function to estimate $\boldsymbol{F}$. In such cases, it is imperative to make sure the factor estimates are unique. Therefore, we discuss first how to uniquely identify the factors, loadings and other parameters, before discussing Bayesian inference using hierarchical priors.

Identification of the factor model

The factor model implies that the conditional covariance matrix of $\boldsymbol{X}$ can be decomposed as

$$cov\left(\boldsymbol{X}|\boldsymbol{\Lambda}, \boldsymbol{F}, \boldsymbol{\Sigma}\right) \equiv \boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}. \tag{204}$$

This decomposition illustrates the fact that likelihood-based (maximum likelihood or Bayesian) estimation of the factor model suffers from lack of identification of a unique set of parameter estimates. For example, consider the case where $\boldsymbol{\Sigma}$ is a full matrix, then there are infinite ways to construct the decomposition in Equation 204. In order to deal with this issue, it is common

larger, but also the case where n is small but much larger than the number of available observations T.

in factor analysis to set $\boldsymbol{\Sigma}$ to be a diagonal matrix.[34] A consequence of this assumption is that the disturbances $\boldsymbol{\varepsilon}_j$ become idiosyncratic to each variable $\boldsymbol{X}_j$, $j = 1, ..., n$, that is, they capture measurement errors and other idiosyncrasies of each variable that are not attributed to its covariation with the remaining $n-1$ variables. Instead, any comovements/commonalities in the n variables $\boldsymbol{X}$ are solely captured by the *common component* $\boldsymbol{\Lambda F}$.

Having $\boldsymbol{\Sigma}$ diagonal is a big step towards identification in the factor model. As Lopes and West (2004) mention, $\boldsymbol{\Omega}$ has $n(n+1)/2$ unique elements, therefore, the number of elements in the decomposition of Equation 204 should not exceed that threshold. The matrix $\boldsymbol{\Lambda}$ has nk elements, and $\boldsymbol{\Sigma}$ being diagonal has n elements, therefore we obtain the inequality $n(n+1)/2 \geq nk + n$, which provides an upper bound on the number of factors one can extract: with $n = 5$ variables we can extract $k = 2$ factors, and when $n = 20$ the maximum number of factors that can be extracted is $k = 9$. However, there is a further problem impairing identification of the factor model, and this pertains to separating $\boldsymbol{\Lambda}$ from $\boldsymbol{F}$. Without further restrictions, there are infinite ways of finding such matrices that provide exactly the same values for the common component. Put more formally, if $\boldsymbol{P}$ is an $k \times k$ orthogonal matrix such that $\boldsymbol{PP}' = \boldsymbol{I}_k$, then the factor model can be rewritten as

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{\Lambda F}_t + \boldsymbol{\varepsilon}_t \Rightarrow & (205) \\ \boldsymbol{X}_t &= \boldsymbol{\Lambda PP}'\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t \Rightarrow & (206) \\ \boldsymbol{X}_t &= \boldsymbol{\Lambda}^\star \boldsymbol{F}_t^\star + \boldsymbol{\varepsilon}_t, & (207) \end{aligned}$$

where $\boldsymbol{\Lambda}^\star$ and $\boldsymbol{F}_t^\star$ are alternative estimates to $\boldsymbol{\Lambda}$ and $\boldsymbol{F}_t$ that provide exactly the same likelihood value (they are observationally equivalent). Given that the variance of the factors is normalized to be one, unique identification of the loadings and the factors requires an additional $k(k-1)/2$ restrictions on the loadings matrix $\boldsymbol{\Lambda}$. A standard restriction that is imposed in this case (Lopes and West, 2004) is to restrict $\boldsymbol{\Lambda}$ to be lower triangular, that is, the top $k \times k$ block of this matrix has its $k(k-1)/2$ upper triangular elements equal to zero. This restriction provides local identification up to a rotation of the sign, meaning that we could multiply any column of $\boldsymbol{\Lambda}$ with -1 and do the same to the respective column of the factors $\boldsymbol{F}$, and arrive to an observationally equivalent solution. For that reason, Geweke and Zhou (1996) suggest to further assume the k diagonal elements of $\boldsymbol{\Lambda}$ to be restricted to be positive.

When modeling comovements between financial time series, the assumption $\boldsymbol{F}_t \sim N_k(\boldsymbol{0}, \boldsymbol{I})$ is often not empirically relevant, and instead it is assumed that $\boldsymbol{F}_t \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma}^F)$ with $\boldsymbol{\Sigma}^F$ a diagonal matrix, with possibly heteroskedastic elements that capture changing (over time) volatility of financial variables (see for example Chib et al., 2006). In this case further restrictions on $\boldsymbol{\Lambda}$ are needed, and Chib et al. (2006) choose to fix the diagonal elements of the

[34] This is known as the *exact factor model* assumption, as opposed to the class of *approximate factor models* that allow for "some" weak correlation among the variables in $\boldsymbol{X}$ and a $\boldsymbol{\Sigma}$ covariance that has some non-zero off-diagonal elements. Approximate factor models are typically not estimated with likelihood-based methods, so we don't consider this class of models here.

loadings matrix to be one. Bernanke et al. (2005) extract factors from a large macroeconomic dataset and in their methodology it is imperative for $\mathbf{\Sigma}^F$ to be a full covariance matrix (it is the covariance matrix of a VAR from which they want to identify shocks and estimate impulse response functions). Therefore, with $\mathbf{\Sigma}^F$ a full matrix, these authors further restrict the upper $k \times k$ block of the loadings matrix to be the identity matrix.

All these identification restrictions in various applications of the factor model do not come at no cost. Imposing zeros in the loadings matrix means that certain variables are excluded from determining the factors. In the case of Bernanke et al. (2005), in particular, the identity restriction means that the first variable exclusively loads on the first factor, the second variable exclusively on the second factor, and so on. Therefore, the ordering of the variables in $\boldsymbol{X}$ ends up affecting the estimates of $\boldsymbol{F}$, and in their case this restriction turns out to be empirically detrimental.[35]

Bayesian shrinkage and variable selection in the factor model

Despite the fact that statistical identification of the factor model using zero and sign restrictions on the loadings might contradict evidence in the data, once the factor model is fully identified Bayesian inference becomes straightforward. To see this, we re-write for convenience the factor model including now all relevant identification restrictions that are imposed prior to estimation

$$\begin{aligned}
\boldsymbol{X}_t &= \mathbf{\Lambda}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, ..., T, && (208)\\
\boldsymbol{F}_t &\sim N_k(\mathbf{0}, \boldsymbol{I}), && (209)\\
\boldsymbol{\varepsilon}_t &\sim N_n(\mathbf{0}, \mathbf{\Sigma}), && (210)\\
\Sigma_{ij} &= 0, \quad i \neq j, \quad i, j = 1, ..., n && (211)\\
\Lambda_{ij} &= 0, \quad i < j, \quad i = 1, ..., n, \quad j = 1, ..., k, && (212)\\
\Lambda_{ij} &> 0, \quad i = j, \quad i = 1, ..., n, \quad j = 1, ..., k. && (213)
\end{aligned}$$

Lopes and West (2004) show that this model is conveniently estimated sequentially via the Gibbs sampler, by sampling from conditional posteriors. The priors are of the form

$$\begin{aligned}
\Lambda_{ij} &\sim \begin{cases} N(0, \tau^2), & , i > j \\ N(0, \tau^2) I(\Lambda_{ij} > 0), & , i = j \\ \delta_0(\Lambda_{ij}), & , i < j \end{cases} && (214)\\
\Sigma_{ii} &\sim Inv - Gamma(a_0, b_0), && (215)
\end{aligned}$$

[35] While their factors are statistically identified, they do not carry any economic content (i.e. they are not "structurally identified"). Bernanke et al. (2005) is one of the few papers that estimates a factor model both with principal components and least squares (the "plug-in" approach explained previously) and with Bayesian inference. Comparing the impulse response functions for some key variables using the two estimation methods, there are marked differences. Whenever PCA has been used to estimate the unknown factors, impulse response functions have the signs and shapes expected by economic theory. When likelihood-based factors have been used, the impulse responses of variables such as inflation degenerate to zero for all horizons; (compare Bernanke et al., 2005, Figures II and IV).

where $\delta_0(\Lambda_{ij})$ is the Dirac delta function, that is, a point mass function for Λ_{ij} that is concentrated at zero. The conditional posteriors are

$$\begin{aligned}
\boldsymbol{F}_i|\bullet &\sim N\left(\boldsymbol{V}_F\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right), \boldsymbol{V}_F\right), && (216)\\
\boldsymbol{\Lambda}_i|\bullet &\sim N\left(\boldsymbol{V}_{L,i}\left(\frac{1}{\Sigma_{ii}}\boldsymbol{F}\boldsymbol{X}_i\right), \boldsymbol{V}_{L,i}\right), \quad i=1,...,n, && (217)\\
\Sigma_{ii}|\bullet &\sim Inv-Gamma\left(a_0+n/2, b_0+SSE_i\right), \quad i=1,...,n, && (218)
\end{aligned}$$

where $\boldsymbol{V}_F=(\boldsymbol{I}+\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}$, $\boldsymbol{V}_{L,i}=(\boldsymbol{D}^{-1}+\frac{1}{\Sigma_{ii}}\boldsymbol{F}'\boldsymbol{F})^{-1}$ and $SSE_i=(\boldsymbol{X}_i-\boldsymbol{F}\boldsymbol{\Lambda}_i)'(\boldsymbol{X}_i-\boldsymbol{F}\boldsymbol{\Lambda}_i)$. The $|\bullet$ notation above is used to denote conditioning on other parameters and data.

By updating $\boldsymbol{\Lambda}$ conditional on $\boldsymbol{F}$ and vice-versa, the Gibbs sampler works around the issue that the product of these two parameters shows up in the likelihood function. Of course, this sequential updating of the common component by updating each of $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ conditional on the other, will inevitably generate unwanted correlation in the Gibbs chain. In order to deal with the sampling inefficiency associated with correlated MCMC draws, Chib et al. (2006) proposed an alternative Metropolis-Hasting step for updating $\boldsymbol{\Lambda}$ conditional on the factors, while Ghosh and Dunson (2009) proposed a parameter-expanded Gibbs sampler (see also Ročková and George, 2016, for a similar parameter expanded factor model estimated with an EM algorithm). Notice that the fact that $\boldsymbol{\Lambda}$ has the required zero and sign restrictions imposed prior to estimation, means that every time we sample $\boldsymbol{F}$ conditional on $\boldsymbol{\Lambda}$ the factors will be sampled from a unique, identified posterior distribution. Had we not imposed these restrictions, the Gibbs sampler would still work numerically, but lack of identification means that each sample could correspond to different pairs of solutions for $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$. That is, in the case where $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ are not separately identified, their product (the common component $\boldsymbol{\Lambda}\boldsymbol{F}$) is always identified. There are certain inference exercises, such as prediction, where it might be the case that identification and interpretation of the factors is not required; see for example the arguments in favor of this approach in Bhattacharya and Dunson (2011) and in Korobilis (2020).

An early attempt to full-Bayes inference in factor models is West (2003) who developed a variable selection prior in the loadings of the static factor model. Under the Lopes and West (2004) identification scheme, the extension proposed by West (2003) simply involves replacing the prior in Equation 214 with a spike and slab prior. As long as the identification restrictions are maintained, the presence of the variable selection prior can be used to find further data-based restrictions in the loadings matrix. Carvalho et al. (2008) extend the variable ideas in West (2003) to create a very sparse static factor model for genome data; see our discussion of their prior in subsection 2.10. Knowles and Ghahramani (2011) further extend these ideas to a spike and slab prior that is semiparametric, utilizing the ability of an Indian Buffet Process prior to allow for infinitely many factors. These authors use a Metropolis-within-Gibbs algorithm for inference. Ročková and George (2016) propose a similar spike and slab formulation based on the Indian Buffet Process, but unlike Knowles and Ghahramani (2011)

they propose maximum a posteriori (MAP) inference by means of approximating the posterior mode using an EM algorithm. The Bayesian asymptotic theory and posterior contraction rates for the sparse static factor model with continuous spike and slab priors is explored in detail in Pati et al. (2014).

Ghosh and Dunson (2009) proposed a heavy-tailed prior on $\mathbf{\Lambda}$ (using a normal/inverse gamma mixture prior) which they argue performs better than the default normal/truncated-normal prior in Equation 214. Bhattacharya and Dunson (2011) proposed a novel multiplicative gamma process prior on the factor loadings that shrinks more aggressively columns of $\mathbf{\Lambda}$ that correspond to a higher number of factors. They call their approach a sparse infinite factor model, as it allows to specify a maximum number of factors and the prior is able to determine zero and non-zero loadings, as well as the number of factors. The gamma process prior for the loadings matrix is of the following "global-local shrinkage" form

$$\begin{aligned}
\Lambda_{ij}|\phi_{ij},\tau_j &\sim N(0,\phi_{ij}^{-1}\tau_j^{-1}), && (219)\\
\phi_{ij} &\sim Gamma(v/2,v/2), && (220)\\
\tau_j &= \prod_{l=1}^{j}\delta_l, \quad j=1,...,k, && (221)\\
\delta_1 &= Gamma(a_1,1), && (222)\\
\delta_l &= Gamma(a_2,1), \quad l\geq 2. && (223)
\end{aligned}$$

While the local shrinkage parameter is the same for each element of $\mathbf{\Lambda}$, the global shrinkage parameter τ_j is shrinking more aggressively as the index j increases, where $j=1,...,k$ indexes the number of factors. This is because τ_j is a j-dimensional product of gamma-distributed random variables. Legramanti et al. (2020) propose a cumulative shrinkage process prior and Srivastava et al. (2017) propose a multi-scale generalized double Pareto prior; both these priors are similar in spirit to the Bhattacharya and Dunson (2011) prior in terms of shrinking the loadings towards zero and selecting the appropriate number of factors at the same time.

We close this section by mentioning ongoing research on alternative solutions to the identification problem (rotational interdeterminacy) in factor models, that do not rely on preimposing zero restrictions on the loadings matrix. The expanded parametric forms proposed in papers such as Bhattacharya and Dunson (2011) and Legramanti et al. (2020) discussed above, deal with this issue efficiently. Other approaches include the ex-post processing approaches of Assmann et al. (2016) and Kaufmann and Schumacher (2019). Früwirth-Schnatter and Lopes (2018) introduce a generalized lower triangular representation of the factor model and propose a sparsity-inducing prior that overshrinks. While papers like West (2003) apply sparsity after imposing zero identifying restrictions, the parameterization of Früwirth-Schnatter and Lopes (2018) allows the prior to impose zeros that are sufficient for identification and inference, thus, not suffering from the rotation problem. Finally, Chan et al. (2018) propose an invariant parameterization of the static factor model that is based on the singular value

decomposition.

4.3 Dynamic sparsity and shrinkage

When working in a time series setting the concepts of shrinkage, variable selection, and model averaging need not be static. This is true for economics where there has always been evidence that predictors can be unstable. There is significant theoretical and empirical evidence that when forecasting oil prices, stock prices, consumer prices, exchange rates, and numerous other economic/financial variables there is hardly a single exogenous predictor that can be claimed to be important over a substantial time sample. In practice, we observe "pockets of predictability", that is, short periods where a specific variable might have predictive information for another variable of interest. This concept of unstable predictors has been popularized since the global financial crisis of 2007-2009, a period when it was obvious that all constant parameter relationships between economic variables completely broke down. Combined with the availability of new Bayesian tools for high-dimensional inference, a large literature has emerged since then that uses terms such as "time-varying sparsity" or "dynamic model averaging" or "time-varying dimension models". Koop and Korobilis (2018) provide a detailed discussion of this literature.[36]

The starting point for imposing dynamic sparsity and dynamic shrinkage is a regression with time-varying parameters (TVPs) and stochastic volatility (henceforth, abbreviated as *TVP regression*) of the form

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta}_t + h_t z_t, \tag{224}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t, \quad \boldsymbol{\beta}_0 \sim N_p(\boldsymbol{0}, \boldsymbol{P}), \tag{225}$$

$$\log h_t = \log h_{t-1} + v_t, \quad \log h_0 \sim N(0, \delta), \tag{226}$$

where y_t is a scalar time series observation for $t = 1, ..., T$, $\boldsymbol{X}_t$ is a p-dimensional vector of covariates (that can include an intercept, own lags of y and exogenous predictors), $\boldsymbol{\beta}_t$ is a vector of time-varying (or drifting) regression coefficients, and h_t is the scalar time-varying (or stochastic) variance/volatility parameter. Additionally, we assume $z_t \sim N(0, 1)$, $\boldsymbol{u}_t \sim N_p(\boldsymbol{0}, \boldsymbol{Q})$ with $\boldsymbol{Q}$ a $p \times p$ covariance matrix, and $v_t \sim N(0, \delta^2)$ with δ^2 a scalar variance parameter.

As is the case with the constant parameter regression, shrinkage is mainly desirable in the TVPs $\boldsymbol{\beta}_t$, but this can take now a dual form: shrinkage towards time-invariance ($\boldsymbol{\beta}_t$ becomes a constant parameter) and "traditional" shrinkage towards zero.[37] Notice that the TVP regression as is specified in equations (224) - (226) is a hiearchical model, and Equation 225 in particular can be though of as a hiearchical prior for $\boldsymbol{\beta}_t$ of the form $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{Q} \sim N(\boldsymbol{\beta}_{t-1}, \boldsymbol{Q})$.

[36] At the same time, in the field of signal processing there is a related literature on "dynamic compressive sensing" for streaming signals (e.g. video); see Ziniel and Schniter (2013).

[37] Shrinkage of h_t towards a constant variance specification is feasible, but it is not desirable for economic and financial time series data, since we know that economic shocks are very volatile and constant parameter specifications are always inferior (both using in-sample and out-of-sample measures of fit).

Seen like this, it is straightforward to assume that $\boldsymbol{Q}$ is diagonal and allow its elements to follow of the hyperpriors we examined previously (e.g. Student-t, Laplace etc). However, doing so would only regularize the evolution of $\boldsymbol{\beta}_t$ around $\boldsymbol{\beta}_{t-1}$, where in the limit of $\boldsymbol{Q} = \mathbf{0}$ then $\boldsymbol{\beta}_t$ becomes a constant parameter. Shrinking $\boldsymbol{\beta}_t$ towards zero requires different treatment, and there are numerous ways one can deal with this problem.

Dynamic variable selection or dynamic model averaging can be implemented in this setting by simply placing appropriate hierarchical priors that will allow to test the hypothesis $H_0 : \beta_{jt} = 0$ vs $H_1 : \beta_{jt} \neq 0$ for all $j = 1, ..., p$ and for all $t = 1, ..., T$. Recall that in "static" Bayesian model averaging the challenge is to average over 2^p regressions. Therefore, the dynamic version of model averaging implies that one has to average over 2^p regression models for each $t = 1, ..., T$. It is not surprising then that many proposed approaches in the literature for dealing with this problem are not based on computationally intensive MCMC algorithms. For example, Koop and Korobilis (2012) and Dangl and Halling (2012) use variance discounting methods (see for example West and Harrison, 1997) in order to provide plug-in estimators of h_t and Q and estimate a single time-varying parameter regression very quickly. Subsequently, dynamic variable selection and model averaging can be implemented by enumerating and estimating all 2^p possible models – as long as p is fairly small (e.g. 20 predictors).

In terms of directly introducing shrinkage and sparsity via hierarchical priors, there are numerous ways of doing so in a TVP regression model. Belmonte et al. (2014) and Bitto and Frühwirth-Schnatter (2019) place hierarchical priors in an equivalent "non-centered" parameterization of the TVP regression that takes the form

$$y_t = \boldsymbol{X}_t\boldsymbol{\theta} + \boldsymbol{X}_t\boldsymbol{W}\boldsymbol{\theta}_t + h_t z_t, \tag{227}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{d}_t, \quad \boldsymbol{\theta}_0 = \mathbf{0}, \tag{228}$$

$$\log h_t = \log h_{t-1} + v_t, \quad \log h_0 \sim N(0, \delta) \tag{229}$$

where $\boldsymbol{d}_t \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$ and $\boldsymbol{W}$ is a diagonal matrix with elements w_j, $j = 1, ..., p$. This formulation is observationally equivalent to the TVP regression of equations (224) - (226). As long as the initial condition is $\boldsymbol{\theta}_0$ it holds that $\boldsymbol{\theta} + \boldsymbol{\theta}_t = \boldsymbol{\beta}_t$. This allows to split the time-varying parameter into a constant parameter level $\boldsymbol{\theta}$ (determined by data $\boldsymbol{X}_t$) and the additive time-variation around the constant level. Additionally, notice that the state equation is now standardized ($\boldsymbol{d}_t$ has unit variance) which can be done by setting $\boldsymbol{W}'^{-1}\boldsymbol{W}^{-1} = \boldsymbol{Q}$. Belmonte et al. (2014) place a Bayesian lasso (Laplace) prior on $\boldsymbol{\theta}$ and on the diagonal elements of $\boldsymbol{W}$. By doing so, they can shrink the total coefficient $\beta_{j,t}$ into a constant parameter θ_j (if $w_j \to 0$), or shrink it to zero (when both $\beta_{j,t}$ and w_j are shrunk towards zero). Alternatively, the model can become an unrestricted TVP regression when both $\beta_{j,t}$ and w_j are not shrunk towards zero by the Laplace prior.

Nakajima and West (2013a) convert the TVP regression into a latent threshold dynamic

regression of the form

$$
\begin{aligned}
y_t &= \boldsymbol{X}_t \boldsymbol{b}_t + h_t z_t, && (230)\\
\boldsymbol{b}_t &= \boldsymbol{\beta}_t \boldsymbol{S}_t, && (231)\\
\boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t, \quad \boldsymbol{\beta}_0 \sim N(\mathbf{0}, \boldsymbol{P}), && (232)\\
\log h_t &= \log h_{t-1} + v_t, \quad \log h_0 \sim N(0, \delta), && (233)
\end{aligned}
$$

where $\boldsymbol{S}_t$ is a $p \times p$ diagonal matrix with element $s_{j,t} = I(\beta_{j,t} \geq d_j)$. That is, the $s_{j,t}$ are 0/1 indicators that can shrink $b_{j,t}$ either towards zero or towards the unrestricted TVP $\beta_{j,t}$. The threshold value d_j can be estimated endogenously such that the data decide which coefficients are zero (or not) at each point in time. Of course, similar to interpretation of spike and slab priors, the condition $I(\beta_{j,t} \geq d_j)$ is a soft, rather than a hard, thresholding rule, due to the fact that $s_{j,t}$ (in a Bayesian setting) is a random variable. This means that once considering the full uncertainty in the posterior the approach of Nakajima and West (2013a) provides a class of smooth thresholding models; see also Nakajima and West (2013b, 2015, 2017).

Ročková and McAlinn (2017) specify a dynamic spike and slab prior of the form

$$
\begin{aligned}
\boldsymbol{\beta}_t | \boldsymbol{\mu}_t, \boldsymbol{\gamma}, \lambda_0, \lambda_1 &\sim (\boldsymbol{I} - \boldsymbol{\Gamma}) N(\mathbf{0}, \lambda_0 \boldsymbol{I}_p) + \boldsymbol{\Gamma} N(\boldsymbol{\mu}_t, \lambda_1 \boldsymbol{I}_p), && (234)\\
\boldsymbol{\mu}_t &\sim N(\boldsymbol{\mu}_{t-1}, \boldsymbol{Q}), && (235)
\end{aligned}
$$

where $\boldsymbol{\Gamma} = diag(\boldsymbol{\gamma}) = diag(\gamma_1, ..., \gamma_p)$ and λ_0 and λ_1 can also have further exponentional prior distributions, converting this prior into a dynamic version of the spike and slab lasso of Ročková and George (2018). This prior is a spike and slab prior for $\boldsymbol{\beta}_t$ but it is only the slab component that incorporates the random walk evolution via the prior mean for $\boldsymbol{\mu}_t$. In contrast, Koop and Korobilis (2018) propose a similar but non-hierarchical prior of the form

$$
\begin{aligned}
\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}, \boldsymbol{Q} &\sim N_p(\boldsymbol{\beta}_{t-1}, \boldsymbol{Q}), && (236)\\
\boldsymbol{\beta}_t | \boldsymbol{\gamma}, \boldsymbol{\tau}^2 &\sim (\boldsymbol{I} - \boldsymbol{\Gamma}) N_p(\mathbf{0}, c\boldsymbol{D}_\tau) + \boldsymbol{\Gamma} N_p(\mathbf{0}, \boldsymbol{D}_\tau), && (237)
\end{aligned}
$$

where $\boldsymbol{D}_\tau = diag(\boldsymbol{\tau}^2) = diag(\tau_1^2, ..., \tau_p^2)$ and c is a small constant (set to $c = 0.0001$ in Koop and Korobilis, 2018). Again it is trivial to allow $\boldsymbol{\tau}^2$ to have its own hyperprior, such that we can combine shrinkage with sparsity in one setting. Finally, Kalli and Griffin (2014) modify Equation 225 and introduce a normal-gamma mixture evolution for $\boldsymbol{\beta}_t$, which can be written in the following hyperprior form

$$
\begin{aligned}
\beta_{j,t} | \beta_{j,t-1}, \psi_{j,t}, \psi_{j,t-1} &\sim N\left(\frac{\psi_{j,t}}{\psi_{j,t-1}} \phi_j \beta_{j,t-1}, \left(1 - \phi_j^2\right) \psi_{j,t} \right), && (238)\\
\psi_{j,t} | \kappa_{j,t-1} &\sim Gamma\left(\lambda_j + \kappa_{j,t-1}, \frac{\lambda_j}{\mu_j (1 - \rho_j)} \right), && (239)\\
\kappa_{j,t} &\sim Poisson\left(\frac{\rho_j \lambda_j \psi_{j,t-1}}{\mu_j (1 - \rho_j)} \right), && (240)
\end{aligned}
$$

which makes this a normal-gamma-Poisson mixture distribution. While this mixture having a very flexible distributional form, implying interesting shapes for $\boldsymbol{\beta}_t$, there might be sensitivity to the choice of the key hyperparameters $(\lambda_j, \rho_j, \mu_j)$.

All the examples above use the state-space form of the TVP regression and rely on recursive estimation methods, either in the form of the simple Kalman filter or (within the context of simulation methods) forward filtering backward sampling (FFBS) algorithms. However, as noted by Korobilis (2021) one can simply discard the prior $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{Q} \sim N(\boldsymbol{\beta}_{t-1}, \boldsymbol{Q})$ and treat the TVP regression as a constant parameter regression. This can be seen if we stack all observations in Equation 224 and write it as

$$\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_{T-1} \\ y_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \ddots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & ... & \boldsymbol{X}_{T-1} & \mathbf{0} \\ \mathbf{0} & ... & \mathbf{0} & \boldsymbol{X}_T \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ ... \\ \boldsymbol{\beta}_{T-1} \\ \boldsymbol{\beta}_T \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_{T-1} \\ \varepsilon_T \end{bmatrix}, \tag{241}$$

$$\boldsymbol{y} \qquad\qquad \mathfrak{X} \qquad\qquad \boldsymbol{B} \qquad\qquad \boldsymbol{\varepsilon}, \tag{242}$$

where $\varepsilon_t \sim N(0, h_t^2)$. In this form, the TVP regression is a model with T observations and Tp covariates and it can be estimated as a high-dimensional "static" regression with data matrices $\boldsymbol{y}$ and $\mathfrak{X}$ as defined in Equation 242. Korobilis (2021) shows that a large class of hierarchical shrinkage priors can be placed on the $Tp \times 1$ parameter vector $\boldsymbol{B}$, and inference can proceed using the regression form in Equation 242 without the need to rely on state-space methods. Since $\boldsymbol{B}$ has T time copies of parameters on the p predictors in $\boldsymbol{X}_t$, more structured shrinkage can be placed by using a group lasso or other similar grouping prior.

Applying a shrinkage or variable selection prior directly to the vector of parameters $\boldsymbol{\beta}_t$ means that a certain $\beta_{j,t}$ might be unrestricted in period s, then restricted to zero in period $s+1$, then switch back to being unrestricted in $s+2$, and so on, for $s \in \{1, ..., T\}$. This is a noisy approach to dynamic shrinkage/sparsity, and more persistent estimates over time might be desirable such that we prevent an important coefficient from becoming sparse just for a period or two, and vice-versa for a sparse coefficient. In many economic data, there is evidence of prolonged regimes where coefficients are either important or not important (e.g. macroeconomic recessions vs expansions, or bull vs bear stock markets). In this case, it might be desirable to incorporate the information in $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{Q} \sim N(\boldsymbol{\beta}_{t-1}, \boldsymbol{Q})$ alongside a hierarchical shrinkage prior. A simple way to do this is to write the TVP regression as a static regression

for the parameters $\Delta\boldsymbol{\beta}_t = \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}$. This takes the form

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_{T-1} \\ y_T \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & ... & \mathbf{0} \\ \boldsymbol{X}_2 & \boldsymbol{X}_2 & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{X}_{T-1} & ... & \boldsymbol{X}_{T-1} & \mathbf{0} \\ \boldsymbol{X}_T & ... & \boldsymbol{X}_T & \boldsymbol{X}_T \end{bmatrix}}_{\mathfrak{X}} \underbrace{\begin{bmatrix} \boldsymbol{\beta}_1 \\ \Delta\boldsymbol{\beta}_2 \\ ... \\ \Delta\boldsymbol{\beta}_{T-1} \\ \Delta\boldsymbol{\beta}_T \end{bmatrix}}_{\boldsymbol{B}^{\Delta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_{T-1} \\ \varepsilon_T \end{bmatrix}}_{\boldsymbol{\varepsilon}}, \tag{243}$$

(244)

where we implicitly assume that $\boldsymbol{\beta}_0 = 0$ such that $\Delta\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1$. The t-th equation of the system above can be written as:

$$\begin{aligned} y_t &= \boldsymbol{X}_t\Delta\boldsymbol{\beta}_t + \boldsymbol{X}_t\Delta\boldsymbol{\beta}_{t-1} + ... + \boldsymbol{X}_t\Delta\boldsymbol{\beta}_2 + \boldsymbol{X}_t\boldsymbol{\beta}_1 + \varepsilon_t, && (245) \\ &= \boldsymbol{X}_t(\Delta\boldsymbol{\beta}_t + \Delta\boldsymbol{\beta}_{t-1} + ... + \boldsymbol{\beta}_1) + \varepsilon_t, && (246) \\ &= \boldsymbol{X}_t\boldsymbol{\beta}_t + \varepsilon_t, && (247) \end{aligned}$$

that is, equations (224) and (243) are observationally equivalent. Under this specification the prior implied by Equation 225 becomes (in matrix form)

$$\boldsymbol{B}^{\Delta} \sim N_{[Tp]}(\mathbf{0}, \boldsymbol{Q}), \tag{248}$$

and this prior can now be converted into a hierarchical prior by placing appropriate hyper-prior distributions on $\boldsymbol{Q}$.

Dynamic shrinkage and sparsity is a very active area of research, and there are several other important contributions that we don't explore here due to space constraints. For further readings we direct the reader to Chan et al. (2012), Irie (2019), Kowal et al. (2019) and Uribe and Lopes (2017), among others.

4.4 High-dimensional causal inference

Let y_i denote an outcome variable and T_i be some treatment variable. Suppose that the p-dimensional vector of cofounders $\boldsymbol{x}_i$ is high-dimensional. The parameter of interest is the treatment effect α in the model below:

$$y_i = \beta_0 + \alpha T_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \tag{249}$$

A *naive post selection* approach would be to apply the lasso to the equation above, excluding α from the ℓ_1-penalty and then regress y_i on T_i as well as on the selected covariates to estimate and conduct inference about the treatment effect. Any control variable that is highly correlated with T_i but weakly with y_i tends to drop out of the selection in the first stage, and could lead to omitted variable bias in estimating α in the second stage. Belloni et al. (2014) propose

post-double selection to overcome such bias. Hahn et al. (2018) and Antonelli et al. (2019) offer Bayesian counterparts in linear models, using shrinkage priors.

Using the model (249) as a benchmark, Hahn et al. (2018) consider the following system of equations:

$$\begin{aligned} T_i &= \boldsymbol{x}_i'\boldsymbol{\gamma} + \epsilon_i, & \epsilon_i &\sim N(0, \sigma_\epsilon^2) \quad \text{(Selection eq.)} \\ y_i &= \alpha T_i + \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, & v_i &\sim N(0, \sigma_v^2) \quad \text{(Response eq.)} \end{aligned}$$

The likelihood can be factorized:

$$f\left(Y, T|\boldsymbol{X}, \boldsymbol{\gamma}, \alpha, \boldsymbol{\beta}, \sigma_\epsilon, \sigma_v\right) = f\left(Y|\boldsymbol{X}, T, \alpha, \boldsymbol{\beta}, \sigma_\epsilon\right) f\left(T|\boldsymbol{X}, \boldsymbol{\gamma}, \sigma_v\right)$$

With re-parameterization $(\alpha, \boldsymbol{\beta} + \alpha\boldsymbol{\gamma}, \boldsymbol{\gamma})' \to (\alpha, \boldsymbol{\beta}_d, \boldsymbol{\beta}_c)'$, the system can be written as

$$\begin{aligned} T_i &= \boldsymbol{x}_i'\boldsymbol{\beta}_c + \epsilon_i, & \epsilon_i &\sim N(0, \sigma_\epsilon^2) \quad \text{(Selection eq.)} \\ y_i &= \alpha\left(T_i - \boldsymbol{x}_i'\boldsymbol{\beta}_c\right) + \boldsymbol{x}_i'\boldsymbol{\beta}_d + v_i, & v_i &\sim N(0, \sigma_v^2) \quad \text{(Response eq.)} \end{aligned}$$

The authors place independent shrinkage priors over $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_d$:

$$\begin{aligned} \pi(\beta_{ij}) &\propto \frac{1}{v}\log\left(1 + \frac{4}{(\beta_j/v)^2}\right) \quad j = 1, ..., p, \quad i = c, d \\ v &\sim C^+(0, 1) \end{aligned}$$

where $C^+(0,1)$ denotes a folded standard Cauchy distribution. This prior is a proxy of the horseshoe prior. Non-informative priors are used for other parameters $\alpha \propto 1$, $\sigma_\epsilon \propto \frac{1}{\sigma_\epsilon}$, and $\sigma_v \propto \frac{1}{\sigma_v}$. They use a slice sampler for posterior sampling. Hahn et al. (2020) extends the approach to nonparametric case using regression trees.

Antonelli et al. (2019) propose a spike-and-slab lasso prior approach. Their proposed framework is

$$\begin{aligned} y_i|T_i, \boldsymbol{x}_i, \alpha, \boldsymbol{\beta}, \sigma^2 &\sim N\left(\beta_0 + \alpha T_i + \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2\right) \\ p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2) &= \prod_{j=1}^{p} \gamma_j\psi_1(\beta_j; \lambda_1, \sigma^2) + (1 - \gamma_j)\psi_0(\beta_j; \lambda_0, \sigma^2) \\ p(\boldsymbol{\gamma}|\theta) &= \prod_{j=1}^{p} \theta^{\omega_j\gamma_j}\left(1 - \theta^{\omega_j}\right)^{1-\gamma_j} \\ p(\theta|a, b) &\sim Beta(1, 0.1p) \\ p(\sigma^2|c, d) &\sim Inv - Gamma(c, d) \\ \beta_0, \alpha &\sim N(0, K) \end{aligned}$$

where $\psi_0(\beta_j; \lambda_0, \sigma^2) = \frac{\lambda_0}{2\sigma}e^{-\lambda_0|\beta_j|/\sigma}$ and $\psi_1(\beta_j; \lambda_1, \sigma^2) = \frac{\lambda_1}{2\sigma}e^{-\lambda_1|\beta_j|/\sigma}$. λ_1 is fixed to be a

small value, say 0.1, so that the prior variance in the slab component is high enough to be uninformative.

The hyperparameter λ_0 is chosen via empirical Bayes. A new feature that they introduce is the weights ω_j which are tuning parameters that they use to prioritize variables to be included (i.e. $\gamma_j = 1$) if they are associated with the treatment. Specifically, they first fit the standard lasso on the model for predicting T given X. If the jth covariate x_j has non-zero coefficient from the lasso, they set $\omega_j = \delta$ for some $\delta \in (0, 1)$. For other variables, $\omega_j = 1$. On the one hand, a smaller value of δ leads to higher inclusion probability and hence more protection against the omitted variable bias. On the other hand, one needs to ensure a small enough inclusion probability for an unimportant variable in the outcome model. To balance the trade off, the authors choose $\delta \in (0, 1)$ as the smallest value of ω_j such that the inclusion probability of $\beta_j = 0$ is less than 0.1. See also Antonelli et al. (2020), who introduce how posterior distributions of treatment and outcome models can be used together with doubly robust estimators.

4.5 Bayesian quantile regression

A regression specification can be represented more generally using the formulation

$$y_i = f(y_i|\boldsymbol{X}_i) + \varepsilon, \tag{250}$$

where $f(y_i|\boldsymbol{X}_i)$ is a conditional mean function for y (conditional on covariates $\boldsymbol{X}$). Using this notation, the linear regression model can be recovered if we set $f(y_i|\boldsymbol{X}_i) = E(y_i|\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, that is, the linear regression only models the (conditional) mean of y. The distribution of y is fully determined by the assumptions we make about the disturbance term ε. In many cases it is desirable to use the information in covariates in such a way that the full distribution of y is determined by $\boldsymbol{X}$. While such feature can also be incorporated implicitly in a traditional linear regression setting[38], a structured (and popular) way is to model the conditional quantiles of y, $\mathbb{Q}_r(y_i|\boldsymbol{X}_i)$, where $r \in (0, 1)$ denotes the quantile of y. While the conditional quantile can be modeled using either linear or nonlinear functional forms, the linear form is by far the most widely used.

In this case, we replace in Equation 250 $f(y_i|\boldsymbol{X}_i) = \mathbb{Q}_r(y_i|\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}_r$ and obtain the following *quantile regression specification*

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta}_r + \varepsilon_i. \tag{251}$$

The model above is a linear regression for each quantile level r. Koenker and Bassett (1978)

[38] For example, we can assume $\varepsilon_i \sim N(0, \sigma_i^2)$, where σ_i^2 can be some function of $\boldsymbol{X}_i$.

show that an estimator of this quantile regression model can be obtained as

$$\widehat{\boldsymbol{\beta}}_r = \min_{\boldsymbol{\beta}_r} \mathbb{E}\left(\sum_{i=1}^{n} \rho_r(\varepsilon_i)\right), \tag{252}$$

where $\rho_r(u) = (r - \mathbb{I}(u < r))$ is a loss function. Yu and Moyeed (2001) show that the same estimator $\widehat{\boldsymbol{\beta}}_r$ can be obtaining by obtaining the maximum likelihood estimator under the assumption that ε_i is distributed as asymmetric Laplace, i.e. if it has the density

$$p(\varepsilon_i; r, \sigma^2) \propto \frac{\tau(1-\tau)}{\sigma_r^2}\left[e^{(1-r)\frac{\varepsilon_i}{\sigma_r^2}}\mathbb{I}(\varepsilon_i \leq 0) + e^{-r\frac{\varepsilon_i}{\sigma_r^2}}\mathbb{I}(\varepsilon_i > 0)\right], \tag{253}$$

where σ_r^2 is a scale parameter. Therefore, the contribution of Yu and Moyeed (2001) provides a parametric framework for implementing Bayesian inference.[39] In particular, Kozumi and Kobayashi (2011) take advantage of the fact that the asymmetric Laplace likelihood can be written as a normal-exponential mixture of the form

$$\varepsilon_i | u_i, z_{i,r} \sim \theta_r z_{i,r} + \sqrt{\sigma_r^2 \kappa_r^2 z_{i,r}} u_t, \tag{254}$$

where $z_{i,r} \sim Exp(\sigma_r^2)$ and $u_t \sim N(0,1)$, while θ_r, κ_r^2 are parameters defined as $\theta_r = \frac{1-2r}{r(1-r)}$, $\kappa_r^2 = \frac{2}{r(1-r)}$. If we marginalise Equation (254) over the exponentially distributed $z_{i,r}$ we obtain Equation (253); see a derivation in Khare and Hobert (2012).

Therefore, the Bayesian quantile regression model has the following representation

$$y_i = \boldsymbol{X}_i \boldsymbol{\beta}_r + \theta_r z_{i,r} + \sqrt{\sigma_r^2 \kappa_r^2 z_{i,r}} u_t, \tag{255}$$

where $z_{i,r} \sim Exp(\sigma_r^2)$ and $u_t \sim N(0,1)$ can either be thought of two disturbance terms (similar to frontier models in econometrics), or equivalently u_t can be interpreted as the disturbance term and $z_{i,r}$ can be thought of as an unobserved factor (similar to the factors we examined in subsection 4.2). In any case, conditional posteriors are trivial to derive as conditional on all other parameters, the posterior of each parameter of interest has standard form. This is easier

[39] Of course here we have similar conceptual issues as with the Bayesian representation of the lasso estimator: while Tibshirani (1996) showed that the ℓ_1 optimization problem for the lasso is equivalent to the posterior mode of Bayesian regression estimator under a Laplace prior, Castillo et al. (2015) show that the posterior distribution does not contract at the same rate as the posterior mode. Similarly here, there is an equivalence between quantile regression and maximizing the likelihood under an asymmetric Laplace likelihood as both problems provide unique point solutions. Bayesian inference, in contrast, assumes that coefficients are random variables and (unless one focuses on MAP or MMSE estimation) cannot be obtained as the solution to an optimization problem. In practice, however, it turns out that Bayesian quantile regression estimation using the asymmetric Laplace likelihood is a very flexible model, even if it is not identical to the model introduced by Koenker and Bassett (1978).

to see for instance for the coefficients $\boldsymbol{\beta}_r$ by rewriting the model as

$$(y_i - \theta_r z_{i,r}) = \boldsymbol{X}_i \boldsymbol{\beta}_r + \sqrt{\sigma_r^2 \kappa_r^2 z_{i,r}} u_t \Rightarrow \tag{256}$$

$$y_i^\star = \boldsymbol{X}_i \boldsymbol{\beta}_r + \sigma_r^\star u_t, \tag{257}$$

where $y_i^\star = (y_i - \theta_r z_{i,r})$ and $\sigma_r^\star = \sqrt{\sigma_r^2 \kappa_r^2 z_{i,r}}$. As long as we condition on $z_{i,r}, \sigma_r^2, \kappa_r^2, \theta_r$, we can obtain a sample of $\boldsymbol{\beta}_r$ (assuming a normal prior), from the standard formulas for a linear regression of $y_i^\star$ on $\boldsymbol{X}_i$ with known variance $\sigma_r^\star$.

In more detail, we assume the following priors

$$\boldsymbol{\beta}_r \sim N(0, \boldsymbol{D}_{\tau,r}), \tag{258}$$

$$\sigma_r^2 \sim Inv - Gamma(n_{0,r}, s_{0,r}), \tag{259}$$

$$z_{i,r} \sim Exponential(\sigma_r^2), \tag{260}$$

where for simplicity assume that $\boldsymbol{D}_{\tau,r} = \boldsymbol{D}_\tau = \tau \times I_p$ where τ is fixed and known for all r. The conditional posteriors (Khare and Hobert, 2012) are of the form

$$\boldsymbol{\beta}_r | \bullet \sim N_p\left(\boldsymbol{V} \times \left(\boldsymbol{X}'\boldsymbol{U}^{-1}\widetilde{\boldsymbol{y}}\right), \boldsymbol{V}\right), \tag{261}$$

$$\sigma_r^2 | \bullet \sim Inv - Gamma\left(a_r, s_{0,r} + \sum_{i=1}^{n} \frac{(y_i^\star)^2}{2 z_{i,r} \kappa_r^2} + \sum_{i=1}^{n} z_{i,r}\right), \tag{262}$$

$$z_{i,r}^{-1} | \bullet \sim IG\left(\frac{\sqrt{\theta_r^2 + 2\kappa_r^2}}{|y_i - \boldsymbol{X}_i \boldsymbol{\beta}_r|}, \frac{\theta_r^2 + 2\kappa_r^2}{\sigma_r^2 \kappa_r^2}\right), \tag{263}$$

where the notation $|\bullet$ means "conditioning on other parameters and data", $\boldsymbol{V} = \left(\boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{X} + D_{\tau,r}^{-1}\right)^{-1}$, $\boldsymbol{U} = \left(\sigma_r^2 \kappa_r^2\right) \times diag\left(z_{1,r}, ..., z_{n,r}\right)$, $\widetilde{\boldsymbol{y}} = (\boldsymbol{y} - \theta_r \boldsymbol{z}_r)$, $y_i^\star = (y_i - \boldsymbol{X}_i \boldsymbol{\beta}_r - \theta_r z_{i,r})$, and $a_r = n_{0,r} + \frac{3n}{2}$.

We need to note here a few important points regarding the implementation of this model

1. The Gibbs sampler in equations (261) - (263) is only one of the many implementations of the Bayesian quantile regression using an asymmetric Laplace likelihood. Kozumi and Kobayashi (2011) first developed a Gibbs sampler were $z_{i,r}$ is sampled from a three-parameter generalized inverse Gaussian (GIG) distribution. Khare and Hobert (2012) proved that this Gibbs sampler is ergodic, but proposed to sample $z_{i,r}^{-1}$ from the two-parameter inverse Gaussian (IG) posterior of (263). While the two sampling steps are identical (the transformation utilizes the ability of GIG distribution to be written as an equivalent IG distribution), the consequences in programming might be more important. For example, MATLAB only has a built-in random number generator for the IG distribution but not for the GIG (although contributed packages on the internet do exist), while in R there are libraries that provide reliable generators for both the IG and GIG distributions.

2. The Gibbs sampler needs to be run for each quantile level r. Therefore, we need to choose quantile levels that are reasonable. For most empirical cases the grid $r = 0.05, 0.10, 0.25, 0.5, 0.75, 0.90, 0.95$ covers the most important areas of a distribution of interest, but of course one can consider much finer grids at the cost of increased computation.

3. Estimation of the quantile $r = 0.5$ is the most accurate, as 50% of the data lie on the left/right of the median. As r approaches 0 or 1, estimation accuracy might decrease as for some problems the number of observations in the tails could be too low (e.g. short, quarterly macroeconomic time series). If ultra high-frequency financial data are available (e.g. 1-min data) then typically the researcher is able to consider values of r closer to 0 or 1 without problems.

4. The conditional posteriors are applied for each quantile level independently. If we consider small grid of quantiles, for example $r = 0.05, 0.10, 0.25, 0.5, 0.75, 0.90, 0.95$ then one can use vectorized operations (for matrix programming languages, such as MATLAB) to obtain $\boldsymbol{\beta}_r$ for all r – despite the fact that these samples of $\boldsymbol{\beta}_r$ will be uncorrelated across r. If we consider a very fine grid for values of r, then further benefits can be achieved if we parallelize and split the sampling equations for each r in different cores.

5. The fact that the regression for each quantile is estimated independently, means that one can obtain estimates of the conditional quantile function $\mathbb{Q}_r(y_i|\boldsymbol{X}_i)$ that are not ordered. That is, solutions of the form $\widehat{\mathbb{Q}}_{r_1}(y_i|\boldsymbol{X}_i) > \widehat{\mathbb{Q}}_{r_2}(y_i|\boldsymbol{X}_i)$ for $r_1 < r_2$ are not compatible with the concept of quantile. Putting this in context, we can't allow the model to predict that the conditional median of inflation is 2% while its first quartile is 2.5%! This problem, which is known as quantile crossing, pertains to all quantile models for which every conditional quantile level is estimated independently from the others (regardless of whether estimation is Bayesian or not). However, Rodrigues and Fan (2017), among numerous others, provide a fully Bayesian algorithm for post-processing MCMC draws of the conditional quantiles in order to ensure monotonicity. The idea is to use a nonparametric smoothing function in order to ensure that this monotinicity exists. The approach of Rodrigues and Fan (2017) is attractive from a Bayesian parametric perspective because it uses the properties of the asymmetric Laplace distribution in order to derive the implied information that a quantile level r' conveys for some other quantile level r, thus, using an expanded information set when smoothing conditional quantile estimates.

After taking the modeling points above into consideration, estimation of the Bayesian quantile regression model is not much more challenging than the normal (Gaussian) linear regression model. All the hierarchical priors described previously can be readily applied to the quantile model, with very minor modifications to the conditional posteriors presented in equations (261)

- (263). The interesting feature is that because a separate regression needs to be run for each r, we can also use shrinkage or sparsity priors to allow different covariates in $\boldsymbol{X}$ to affect different quantiles of the distribution of y. Exactly because application of hierarchical priors is so trivial, we won't provide detailed examples and derivations here. Kozumi and Kobayashi (2011), while proposing a Gibbs sampler for the Bayesian quantile model, they also show in a subsection how easy it is to adopt the double-exponential (lasso) prior. Other applications include Alhamzawi and Yu (2012), Yu et al. (2013), Korobilis (2017) and Korobilis et al. (2021), and the reader can consult these papers for modeling details. Lim et al. (2020) is the case of a paper that estimates a Bayesian quantile regression using variational Bayes methods.

5 Concluding remarks

We have reviewed a wide range of concepts and algorithms for Bayesian sparse and shrinkage estimation. Our focus was on recent contributions in the field, covering the mainly academic publications during the decade 2010-2020, that are increasingly focusing on efficient computation and inference in high-dimensional models. A major contribution of our work is to collect in a single document all these recent contributions, as other reviews and surveys we are aware of provide only very focused reviews of certain hierarchical priors and algorithms. While we believe contributions to the field of Bayesian sparse and shrinkage estimation will keep expanding at a polynomial rate, we do hope that this review will become a useful manual for PhD students and researchers who want an accessible introduction to the field.

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):111–142.

Alhamzawi, R. and Ali, H. T. M. (2018). The Bayesian adaptive lasso regression. *Mathematical Biosciences*, 303:75 – 82.

Alhamzawi, R. and Yu, K. (2012). Variable selection in quantile regression via gibbs sampling. *Journal of Applied Statistics*, 39(4):799–813.

Antonelli, J., Papadogeorgou, G., and Dominici, F. (2020). Causal inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. *Biometrics*.

Antonelli, J., Parmigiani, G., and Dominici, F. (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Analysis*, 14(3):805 – 828.

Armagan, A., Clyde, M., and Dunson, D. (2011). Generalized beta mixtures of Gaussians. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica*, 23:119–143.

Armagan, A. and Zaretzki, R. L. (2010). Model selection via adaptive shrinkage with t priors. *Computational Statistics*, 25(3):441–461.

Assmann, C., Boysen-Hogrefe, J., and Pape, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics*, 192(1):190–206.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–3430.

Bai, R., Rockova, V., and George, E. I. (2021). Spike-and-slab meets lasso: A review of the spike-and-slab lasso. *arXiv preprint arXiv:2010.06451*.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Baumeister, C., Korobilis, D., and Lee, T. K. (2020). Energy Markets and Global Economic Conditions. *The Review of Economics and Statistics*, pages 1–45.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belmonte, M. A., Koop, G., and Korobilis, D. (2014). Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33(1):80–94.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, volume Springer Series in Statistics. Springer-Verlag New York.

Berger, J. O. and Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, 94(446):542–554.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Berger, J. O. and Pericchi, L. R. (1998). Accurate and stable Bayesian model selection: The median intrinsic Bayes factor. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 60(1):1–18.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In Lahiri, P., editor, *Model selection*, volume Volume 38 of *Lecture Notes–Monograph Series*, pages 135–207. Institute of Mathematical Statistics, Beachwood, OH.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *The Quarterly Journal of Economics*, 120(1):387–422.

Bhadra, A., Datta, J., Li, Y., and Polson, N. (2020). Horseshoe regularisation for machine learning in complex and deep models. *International Statistical Review*, 88(2):302–320.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, pages 291–306.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490. PMID: 27019543.

Bitto, A. and Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1):75 – 97. Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective".

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 39(3):1551–1579.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.*, 5(3):583–618.

Cao, X., Khare, K., and Ghosh, M. (2020). High-dimensional posterior consistency for hierarchical non-local priors in regression. *Bayesian Analysis*, 15(1):241 – 262.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73 – 108.

Caron, F. and Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 88–95, New York, NY, USA. ACM.

Carriero, A., Clark, T. E., and Marcellino, M. (2019). Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212(1):137 – 154. Big Data in Dynamic Predictive Econometric Modeling.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069 – 2101.

Chan, J., Leon-Gonzalez, R., and Strachan, R. W. (2018). Invariant inference and efficient computation in the static factor model. *Journal of the American Statistical Association*, 113(522):819–828.

Chan, J. C. and Grant, A. L. (2016). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics & Data Analysis*, 100:847 – 859.

Chan, J. C., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2012). Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3):358–367.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

Chib, S., Nardari, F., and Shephard, N. (2006). Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, 134(2):341–371.

Chipman, H., George, E. I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In Lahiri, P., editor, *Model selection*, volume Volume 38 of *Lecture Notes–Monograph Series*, pages 65–116. Institute of Mathematical Statistics, Beachwood, OH.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In Bernardo, J., Dawid, A., Berger, J., and Smith, A., editors, *Bayesian Statistics 6*. Oxford University Press.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157 – 181.

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Anal.*, 8(1):111–132.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika*, 73(2):323–332.

De Santis, F. and Spezzaferri, F. (1997). Alternative Bayes factors for model selection. *Canadian Journal of Statistics*, 25(4):503–515.

Dehaene, G. and Barthelmé, S. (2018). Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):199–217.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using mcmc. *Statistics and Computing*, 12(1):27–36.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915.

Dunson, D. B., Herring, A. H., and Engel, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, 103(482):534–546.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Eicher, T. S., Papageorgiou, C., and Raftery, A. E. (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, 26(1):30–55.

Fernández, C., Ley, E., and Steel, M. F. (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Fernández, C., Ley, E., and Steel, M. F. J. (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

Figueiredo, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1150–1159.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Fourdrinier, D., Strawderman, W. E., and Wells, M. T. (2018). *Shrinkage Estimation*, volume Springer Texts in Statistics. Springer International Publishing.

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1):1–28.

Früwirth-Schnatter, S. and Lopes, H. (2018). Sparse Bayesian factor analysis when the number of factors is unknown. Technical Report arXiv:1804.04231v1, ArXiV.

Früwirth-Schnatter, S. and Wagner, H. (2010). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9*. Oxford University Press.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515 – 534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, 3rd ed. edition.

Gelman, A. and Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society Series A*, 180(4):967–1033.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

Geweke, J. and Zhou, G. (1996). Measuring the price of the arbitrage pricing theory. *The Review of Financial Studies*, 9(2):557–587.

Ghosh, J. and Clyde, M. A. (2011). Rao–blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association*, 106(495):1041–1052.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19(51):1–49.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532.

Goutis, C. and Robert, C. P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika*, 85(1):29–37.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Griffin, J. E. and Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442.

Griffin, J. E. and Brown, P. J. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12(1):135–159.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965 – 1056.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(1):251–278.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, 89(425):309–319.

Irie, K. (2019). Bayesian dynamic fused lasso. *arXiv preprint arXiv:1905.12275*.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*, 98(462):438–455.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Ji, C. and Schmidler, S. C. (2013). Adaptive markov chain monte carlo for Bayesian variable selection. *Journal of Computational and Graphical Statistics*, 22(3):708–728.

Jiang, W. (2006). On the consistency of Bayesian variable selection for high dimensional binary regression and classification. *Neural Computation*, 18(11):2762–2776.

Johndrow, J., Orenstein, P., and Bhattacharya, A. (2020). Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hvoothesis tests hypothesis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 72(2):143–170.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649.

Judge, G. G., Griffith, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.-C. (1985). *The theory and practice of econometrics*. Wiley, New York.

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290.

Kahn, M. J. and Raftery, A. E. (1992). Fast exact Bayesian inference for the hierarchical normal model: Solving the improper posterior. Technical report, University of Washington.

Kalli, M. and Griffin, J. E. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779 – 793.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.

Kaufmann, S. and Schumacher, C. (2019). Bayesian estimation of sparse dynamic factor models with order-independent and ex-post mode identification. *Journal of Econometrics*, 210(1):116 – 134. Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective".

Khare, K. and Hobert, J. P. (2012). Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *Journal of Multivariate Analysis*, 112:108 – 116.

Khare, K. and Hobert, J. P. (2013). Geometric ergodicity of the Bayesian lasso. *Electron. J. Statist.*, 7:2150–2163.

Kim, A. S. I. and Wand, M. P. (2016). The explicit form of expectation propagation for a simple statistical model. *Electronic Journal of Statistics*, 10(1):550–581.

Knowles, D. and Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koop, G. and Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, 3(4):267–358.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Koop, G. and Korobilis, D. (2016). Model uncertainty in panel vector autoregressive models. *European Economic Review*, 81:115–131.

Koop, G. and Korobilis, D. (2018). Bayesian dynamic variable selection in high dimensions. Technical Report arXiv:1809.03031, ArXiV.

Koop, G., Korobilis, D., and Pettenuzzo, D. (2019). Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210(1):135–154.

Korobilis, D. (2013a). Bayesian forecasting with highly correlated predictors. *Economics Letters*, 118(1):148 – 150.

Korobilis, D. (2013b). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230.

Korobilis, D. (2016). Prior selection for panel vector autoregressions. *Computational Statistics & Data Analysis*, 101:110 – 120.

Korobilis, D. (2017). Quantile regression forecasts of inflation under model uncertainty. *International Journal of Forecasting*, 33(1):11–20.

Korobilis, D. (2020). Sign restrictions in high-dimensional vector autoregressions. Working Paper series 20-09, Rimini Centre for Economic Analysis.

Korobilis, D. (2021). High-dimensional macroeconomic forecasting using message passing algorithms. *Journal of Business & Economic Statistics*, 39(2):493–504.

Korobilis, D., Landau, B., Musso, A., and Phella, A. (2021). The time-varying evolution of inflation risks. Working Paper Series 2600, European Central Bank.

Korobilis, D. and Pettenuzzo, D. (2019). Adaptive hierarchical priors for high-dimensional vector autoregressions. *Journal of Econometrics*, 212(1):241 – 271.

Korobilis, D. and Pettenuzzo, D. (2020). Machine learning econometrics: Bayesian algorithms and methods. *Oxford Research Encyclopedia of Economics and Finance*.

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

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565–1578.

Krishna, A., Bondell, H. D., and Ghosh, S. K. (2009). Bayesian variable selection using an adaptive powered correlation prior. *Journal of Statistical Planning and Inference*, 139(8):2665 – 2674.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 60(1):65–81.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):247–262.

Legramanti, S., Durante, D., and Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752.

Leng, C., Tran, M.-N., and Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244.

Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655.

Li, H. and Pati, D. (2017). Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107–119.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Lim, D., Park, B., Nott, D., Wang, X., and Choi, T. (2020). Sparse signal shrinkage and outlier detection in high-dimensional quantile regression with variational Bayes. *Statistics and Its Interface*, 13(2):237–249.

Lindley, D. V. (1983). Parametric empirical Bayes inference: Theory and applications: Comment. *Journal of the American Statistical Association*, 78(381):61–62.

Liu, Y., Ročková, V., and Wang, Y. (2019). Variable selection with abc Bayesian forests. Technical Report arXiv:1806.02304v2, ArXiV.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–67.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique*, 63(2):215–232.

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Mallick, H. and Yi, N. (2014). A new Bayesian lasso. *Statistics and its Interface*, 7(4):571–582.

Martini, A. S. and Spezzaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):296–303.

Matusevich, D. S., Cabrera, W., and Ordonez, C. (2016). Accelerating a Gibbs sampler for variable selection on genomics data with summarization and variable pre-selection combining an array DBMS and R. *Machine Learning*, 102(3):483–504.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Moran, G. E., Ročková, V., and George, E. I. (2019). Variance prior forms for high-dimensional Bayesian variable selection. *Bayesian Analysis*, 14(4):1091–1119.

Nakajima, J. and West, M. (2013a). Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31(2):151–164.

Nakajima, J. and West, M. (2013b). Bayesian dynamic factor models: Latent threshold approach. *Journal of Financial Econometrics*, 11:116–153.

Nakajima, J. and West, M. (2015). Dynamic network signal processing using latent threshold models. *Digital Signal Processing*, 47:6–15. https://doi.org/10.1016/j.dsp.2015.04.008.

Nakajima, J. and West, M. (2017). Dynamics and sparsity in latent threshold factor models: A study in multivariate EEG signal processing. *Brazilian Journal of Probability and Statistics*, 31:701–731.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Narisetty, N. N., Shen, J., and He, X. (2018). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association*, 0(0):1–13.

Neville, S. E., Ormerod, J. T., and Wand, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electronic Journal of Statistics*, 8(1):1113–1151.

Nott, D. J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika*, 92(4):747–763.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–117.

Ormerod, J. T., You, C., and Müller, S. (2017). A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549–3594.

Pal, S. and Khare, K. (2014). Geometric ergodicity for Bayesian shrinkage models. *Electronic Journal of Statistics*, 8(1):604 – 645.

Pal, S., Khare, K., and Hobert, J. P. (2017). Trace class Markov chains for Bayesian inference with generalized double Pareto shrinkage priors. *Scandinavian Journal of Statistics*, 44(2):307–323.

Papaspiliopoulos, O. and Rossell, D. (2017). Bayesian block-diagonal variable selection and model averaging. *Biometrika*, 104(2):343–359.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102–1130.

Peltola, T., Marttinen, P., and Vehtari, A. (2012). Finite adaptation and multistep moves in the metropolis-hastings algorithm for variable selection in genome-wide association analysis. *PLOS ONE*, 7(11):1–11.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 9*. Oxford University Press.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266.

Rajaratnam, B., Sparks, D., Khare, K., and Zhang, L. (2019). Uncertainty quantification for modern high-dimensional regression via scalable Bayesian methods. *Journal of Computational and Graphical Statistics*, 28(1):174–184.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, volume Springer Texts in Statistics. Springer-Verlag New York.

Rodrigues, T. and Fan, Y. (2017). Regression adjustment for noncrossing Bayesian quantile regression. *Journal of Computational and Graphical Statistics*, 26(2):275–284.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Ročková, V. and McAlinn, K. (2017). Dynamic variable selection with spike-and-slab process priors. Technical Report arXiv:1708.00085v2, ArXiV.

Rue, H. (2001). Fast sampling of Gaussian markov random fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2):325–338.

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053–1078.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97(460):1141–1153.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493.

Srivastava, S., Engelhardt, B. E., and Dunson, D. B. (2017). Expandable factor analysis. *Biometrika*, 104(3):649–663.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Neyman, J., editor, *Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206. University of California Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244.

Uribe, P. and Lopes, H. (2017). Dynamic sparsity on dynamic regression models. Technical report, Available at http://hedibert.org/wp-content/uploads/2018/06/uribe-lopes-Sep2017.pdf.

van den Boom, W., Dunson, D., and Reeves, G. (2015a). Quantifying uncertainty in variable selection with arbitrary matrices. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 385–388.

van den Boom, W., Dunson, D., and Reeves, G. (2015b). Scalable approximations of marginal posteriors in variable selection. Technical Report arXiv:1506.06629v1, ArXiV.

van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica*, 59(1):45–56.

van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, H. and Pillai, N. S. (2013). On a class of shrinkage priors for covariance matrix estimation. *Journal of Computational and Graphical Statistics*, 22(3):689–707.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7*, pages 723–732. Oxford University Press.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, volume Springer Series in Statistics. Springer-Verlag New York.

Yu, K., Chen, C., Reed, C., and Dunson, D. (2013). Bayesian variable selection in quantile regression. *Statistics and its Interface*, 6(2):261–274. cited By 22.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437 – 447.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243, New York. Elsevier Science Publishers, Inc.

Zhang, Y. and Bondell, H. D. (2018). Variable selection via penalized credible regions with Dirichlet–Laplace global-local shrinkage priors. *Bayesian Analysis*, 13(3):823 – 844.

Ziniel, J. and Schniter, P. (2013). Dynamic compressive sensing of time-varying signals via approximate message passing. *IEEE Transactions on Signal Processing*, 61(21):5270–5284.

Zou, X., Li, F., Fang, J., and Li, H. (2016). Computationally efficient sparse Bayesian learning via generalized approximate message passing. In *2016 IEEE International Conference on Ubiquitous Wireless Broadband (ICUWB)*, pages 1–4.