Bayesian Approaches to Shrinkage and Sparse Estimation

Other titles in Foundations and Trends® in Econometrics

Performance Analysis: Economic Foundations and Trends
Valentin Zelenyuk
ISBN: 978-1-68083-866-4

Experimetrics: A Survey
Peter G. Moffatt
ISBN: 978-1-68083-792-6

Climate Econometrics: An Overview
Jennifer L. Castle and David F. Hendry
ISBN: 978-1-68083-708-7

Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-based Conjoint Analysis
Moshe Ben-Akiva, Daniel McFadden and Kenneth Train
ISBN: 978-1-68083-526-7

Structural Econometrics of Auctions: A Review
Matthew L. Gentry, Timothy P. Hubbard, Denis Nekipelov and Harry J. Paarsch
ISBN: 978-1-68083-446-8

Bayesian Approaches to Shrinkage and Sparse Estimation

Dimitris Korobilis
University of Glasgow
Dimitris.Korobilis@glasgow.ac.uk

Kenichi Shimizu
University of Glasgow
Kenichi.Shimizu@glasgow.ac.uk

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

D. Korobilis and K. Shimizu. *Bayesian Approaches to Shrinkage and Sparse Estimation*. Foundations and Trends® in Econometrics, vol. 11, no. 4, pp. 230–354, 2022.

ISBN: 978-1-63828-035-4

Foundations and Trends® in Econometrics
Volume 11, Issue 4, 2022

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Econometrics publishes survey and tutorial articles in the following topics:

- Econometric Models
- Simultaneous Equation Models
- Estimation Frameworks
- Biased Estimation
- Computational Problems
- Microeconometrics
- Treatment Modeling
- Discrete Choice Modeling
- Models for Count Data
- Duration Models
- Limited Dependent Variables
- Panel Data
- Time Series Analysis
- Latent Variable Models
- Qualitative Response Models
- Hypothesis Testing
- Econometric Theory
- Financial Econometrics
- Measurement Error in Survey Data
- Productivity Measurement and Analysis
- Semiparametric and Nonparametric Estimation
- Bootstrap Methods
- Nonstationary Time Series
- Robust Estimation

Information for Librarians

Foundations and Trends® in Econometrics, 2022, Volume 11, 4 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Contents

Bayesian Approaches to Shrinkage and Sparse Estimation

Dimitris Korobilis and Kenichi Shimizu

University of Glasgow, UK; Dimitris.Korobilis@glasgow.ac.uk, Kenichi.Shimizu@glasgow.ac.uk

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/nonparametric inference is the norm in several fields of applied econometric work. The purpose of this monograph is to introduce the reader to the world 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

Dimitris Korobilis and Kenichi Shimizu (2022), "Bayesian Approaches to Shrinkage and Sparse Estimation", Foundations and Trends® in Econometrics: Vol. 11, No. 4, pp 230–354. DOI: 10.1561/0800000041.

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[1] allow the reader to replicate many of the algorithms described in this monograph.

---

[1] Online Supplementary Material available from: http://dx.doi.org/10.1561/0800000041_supp.

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 monograph 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 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

[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.

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 monograph 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 non-zeros (where zeros in estimation means absence of some effect or relationship). Shrinkage (or often called "regularization" in machine learning) means estimation where many parameter elements are compressed towards zero, but they are not necessarily zero. While many readers might be familiar with these concepts, interpretation from a Bayesian point of view is slightly different compared to 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 toward zero

[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.

in Bayesian inference is achieved by specifying certain forms of priors; a frequentist statistician usually achieves shrinkage via a penalized likelihood approach.

We explain these differences, and many more concepts, in this detailed monograph. 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.

Throughout the monograph, 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 Section 1.2 and then move to Section 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 $X \in \mathcal{X}$ where $\mathcal{X}$ (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 $\mathcal{X}$ and Θ: a 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 $L(\widehat{\theta}(X), \theta) = \rho(\widehat{\theta}(X), \theta)$ 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)

$$r(\pi, \widehat{\theta}) = \int_{\Theta} E_{\theta}(L(\widehat{\theta}(X), \theta)) d\pi(\theta). \tag{1.1}$$

The quantity $\mathcal{R}(\theta, \hat{\theta}) = E_{\theta}(L(\widehat{\theta}(X), \theta))$ 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 (1.1) is the average of frequentist risk $\mathcal{R}$ with respect to the prior distribution $\pi(\theta)$. Frequentist decision theory aims at making the expected loss $\mathcal{R}(\theta, \hat{\theta})$ small, while Bayesian decision theory aims at finding the minimum of $r(\pi, \widehat{\theta})$. In particular, the quantity

$$r(\pi) = \inf_{\widehat{\theta}} r(\pi, \widehat{\theta}), \tag{1.2}$$

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 $r(\pi, \widetilde{\theta}) < r(\pi, \widehat{\theta})$ then $\widetilde{\theta}$ is admissible. An estimator is $\hat{\theta}_0$ is *minimax* for a given loss function L if

$$\sup_{\theta} \mathcal{R}(\theta, \hat{\theta}_0) = \inf_{\hat{\theta}} \sup_{\theta} \mathcal{R}(\theta, \hat{\theta}), \tag{1.3}$$

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 $\sup_{\theta} \mathcal{R}(\theta, \hat{\theta}_{\pi}) = r(\pi, \hat{\theta}_{\pi})$, 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:

$$MSE(\widehat{\theta}) = E[L(\widehat{\theta}, \theta)] = E[(\widehat{\theta} - \theta)^2] \tag{1.4}$$

$$= E[(\widehat{\theta} - E\{\widehat{\theta}\} + E\{\widehat{\theta}\} - \theta)^2] \tag{1.5}$$

$$= E[(\widehat{\theta} - E\{\widehat{\theta}\})^2] + (E\{\widehat{\theta}\} - \theta)^2. \tag{1.6}$$

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 extent 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 (1.6) 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 $X \sim N_p(\theta, \underline{\sigma}^2 I_p)$ 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

$$\widehat{\theta}^{JS} = \left(1 - \frac{(p-2)\sigma^2}{\sum_{i=1}^{n} X_i}\right) X, \tag{1.7}$$

has lower risk than the MLE, that is, $\mathcal{R}(\widehat{\theta}^{JS}) < \mathcal{R}(\widehat{\theta}^{mle})$. 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 forming 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 spaces, where unbiasedness is usually the holy grail for the econometrician. 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 monograph that 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 monograph, 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

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.8}$$

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 $\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X})$ 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

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y})}, \tag{1.9}$$

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

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2). \tag{1.10}$$

A default prior setting in Bayesian inference is the natural conjugate prior[4] which is defined as

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \tag{1.11}$$

$$= N(\mathbf{0}, \sigma^2\boldsymbol{D}) \times Inv-Gamma\left(\frac{v_0}{2}, \frac{s_0^2}{2}\right) \tag{1.12}$$

$$\propto (\sigma^2)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta}\right\} \tag{1.13}$$

$$\times (\sigma^2)^{-v_0/2-1} \exp\left\{-\frac{s_0^2/2}{\sigma^2}\right\}, \tag{1.14}$$

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 Online Supplementary Technical Document) that the posterior is of the form

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = N(\boldsymbol{V}(\boldsymbol{X}'\boldsymbol{y}), \sigma^2\boldsymbol{V}) \times Inv-Gamma\left(\frac{v}{2}, \frac{s^2}{2}\right), \tag{1.15}$$

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

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

[4] Under a conjugate prior, the prior and the posterior of a parameter are of the same distributional form.

1.2.1 Goodness of Fit Measures: Marginal Likelihood and Information Criteria

While Equation (1.10) is of primary importance for the parameter posterior distributions, the quantity $p(\boldsymbol{y})$ in Equation (1.9) is importance for Bayesian model determination. This is the *prior predictive likelihood*, more commonly known as the *marginal likelihood*:

$$p(\boldsymbol{y}) = \int_{-\infty}^{\infty} \int_{0}^{\infty} p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2, \tag{1.16}$$

which is well-defined for proper priors. It 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. 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 (1.16) can only be calculated when the prior is a proper density, that is, if $p(\boldsymbol{\beta}, \sigma^2)$ integrates to one. The non-informative prior on $\boldsymbol{\beta}$ and σ^2 is a key example where this condition fails and the marginal likelihood does not exist.[5]

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. We can then define the *posterior predictive density*:

$$p(y_{n+1}|\boldsymbol{y}) = \int_{-\infty}^{\infty} \int_{0}^{\infty} p(y_{n+1}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\boldsymbol{\beta} d\sigma^2, \tag{1.17}$$

which is the density of the out-of-sample data point marginalized over the posterior density 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

$$p(\boldsymbol{y}) = \frac{p(y|\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})} \tag{1.18}$$

[5] The non-informative prior is of the form $p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$.

$$= \frac{\Gamma(\frac{v_0}{2})^{-1}(s_0/2)^{\frac{v_0}{2}}}{(2\pi)^{\frac{n}{2}}\Gamma\left(\frac{v}{2}\right)^{-1}(s/2)^{\frac{v}{2}}} \frac{|\boldsymbol{D}|^{-\frac{1}{2}}}{|\boldsymbol{V}|^{-\frac{1}{2}}} \tag{1.19}$$

$$\times \left[\frac{1}{2}(s_0 + \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\mu}^{*}\boldsymbol{V}^{-1}\boldsymbol{\mu}^{*})\right], \tag{1.20}$$

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 (1.15) and $\boldsymbol{\mu}^{*} = \boldsymbol{V}(\boldsymbol{X}'\boldsymbol{y})$.

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

$$y_{n+1}|\boldsymbol{y} \sim t_1(y_{n+1}; \boldsymbol{X}_{n+1}\boldsymbol{V}(\boldsymbol{X}'\boldsymbol{y}), \frac{s}{v}(1 + \boldsymbol{X}_{n+1}\boldsymbol{V}\boldsymbol{X}'_{n+1}), v), \tag{1.21}$$

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

$$t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, d) = \frac{\Gamma(\frac{d+p}{2})}{\Gamma(\frac{d}{2})d^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}}$$

$$\times \left[1 + \frac{1}{d}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]. \tag{1.22}$$

The marginal likelihood is rarely available analytically, and in most cases the integral in Equation (1.16) has to be approximated using Monte Carlo or numerical methods.[6] 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 density in Equation (1.17) using a procedure called *leave one out cross-validation* (LOO-CV). This would involve fitting the model in training data and then using a hold-out sample to evaluate the posterior predictive distribution. Notice that if MCMC samples from the parameter posterior are available, evaluation of Equation (1.17) is straightforward using Monte Carlo integration via

[6] Two early examples are Gelfand and Dey (1994) and Chib (1995); see also Chib and Jeliazkov (2001) for a review. Both Gelfand and Dey (1994) and Chib (1995) esitmators of the marginal likelihood rely on derivation of simple expressions for $p(\boldsymbol{y})$, which explains their popularity in applied research. However, both estimators can be numerically sensitive in certain cases (Geweke, 1999), plus they are not appropriate for multi-model comparisons due to their reliance on computationally intensive Monte Carlo simulation methods.

sampling from[7]

$$p(y_{n+1}|\boldsymbol{\beta}_{(s)}, \sigma^2_{(s)}) \tag{1.23}$$

where $(\boldsymbol{\beta}_{(s)}, \sigma^2_{(s)})$, $s = 1, \ldots, S$, are S MCMC samples from $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$.

When marginal or posterior predictive distributions 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[8] $(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)$ for the logarithm of the term $p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)$ in Equation (1.16), we can write the log-marginal likelihood as

$$\begin{aligned} \log p(\boldsymbol{y}) \quad = \quad & \log p(\boldsymbol{y}|\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) + \log p(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) + \frac{p}{2}\log(2\pi) \\ & -\tfrac{p}{2}\log n - \frac{1}{2}\log|J_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)| + O(n^{-1}), \end{aligned} \tag{1.24}$$

where $J_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)$ is the expected Fisher information matrix of $p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \cdot p(\boldsymbol{\beta}, \sigma^2)$ 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 (1.24) any terms of order $O(1)$ or less, we obtain

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) - \frac{p}{2}\log n + O(1). \tag{1.25}$$

The approximation above provides the basis for defining the Schwarz (1978)'s Bayesian information criterion

$$BIC = -2\log \mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2|\boldsymbol{y}, \boldsymbol{X}) + p\log n, \tag{1.26}$$

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

[7] 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).

[8] 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.

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

$$DIC = -4E_{p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)] + 2\log p(\boldsymbol{y}|\widetilde{\boldsymbol{\beta}},\widetilde{\sigma}^2). \tag{1.27}$$

The first term is the expectation of the data density with respect to the posterior[9] which can be evaluated numerically from the MCMC output by taking the mean of $\log p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)$ 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 (2016), Spiegelhalter *et al.* (2014) and van der Linde (2005).

In hierarchical models, there are latent variables in addition to $(\boldsymbol{\beta},\sigma^2)$. In such case, computing the DIC incorporating the latent variables has a practical advantage that it is easy to obtain from MCMC outputs. However, this approach faces difficulties in some cases because the asymptotic justification of DIC can be provided when the dimensionality of the parameter vector does not grow indefinitely with the number of observations. In many hierarchical models, this dimensionality grows asymptotically. See Quintero and Lesaffre (2018) for more discussion and alternative approaches for computing DIC in hierarchical models.

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

[9]For that reason, the DIC is related to the posterior predictive density, i.e., the integral in Equation (1.17), rather than the marginal likelihood.

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 $j = 1, \ldots, p$. 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

$$BF_{12} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}, \tag{1.28}$$

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

$$PO_{12} \equiv \frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)} \times \frac{p(M_1)}{p(M_2)}, \tag{1.29}$$

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 $p(M_1) = p(M_2) = \frac{1}{2}$ 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 (1.16). 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 (1.8) 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

$$BF_{12} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)} \tag{1.30}$$

$$= \frac{\int_{-\infty}^{\infty}\int_{0}^{\infty} p(\boldsymbol{y}|\beta, \sigma^2, M_1)p(\beta, \sigma^2|M_1)d\beta d\sigma^2}{\int_{0}^{\infty} p(\boldsymbol{y}|\beta^\star, \sigma^2, M_2)p(\beta^\star, \sigma^2|M_2)d\sigma^2} \tag{1.31}$$

$$= \frac{\int_{0}^{\infty} p(\beta^\star, \sigma^2|\boldsymbol{y}, M_2)d\sigma^2}{\int_{0}^{\infty} p(\beta^\star, \sigma^2|M_2)d\sigma^2}, \tag{1.32}$$

that is, SDDR 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 $\int_0^\infty p(\beta^\star, \sigma^2|M_2)d\sigma^2 = p(\beta^\star|M_2)\int_0^\infty p(\sigma^2|M_2)d\sigma^2 = p(\beta^\star|M_2)$, 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

$$E(L(\beta, \widehat{\beta})) = [p(M_1|\boldsymbol{y})E(L(\beta, \widehat{\beta})|M_1) \tag{1.33}$$

$$+ p(M_2|\boldsymbol{y})E(L(\beta, \widehat{\beta})|M_2)], \tag{1.34}$$

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

$$\widehat{\beta}^{BPE} = p(M_1|\boldsymbol{y})E(p(\beta|\boldsymbol{y}, M_1)) + p(M_2|\boldsymbol{y})E(p(\beta|\boldsymbol{y}, M_1)). \tag{1.35}$$

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 monograph 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 $\mathcal{M} = \{M_r\colon r = 1, \ldots, 2^p\}$. 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[10]

$$M_r\colon \boldsymbol{y} = \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}, \tag{1.36}$$

where $\boldsymbol{X}_r$ is $n \times p_r$ and $\boldsymbol{\beta}_r$ is $p_r \times 1$ with $p_r \in \{1, \ldots, p\}$. 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 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

$$p(M_r|\boldsymbol{y}) \propto p(\boldsymbol{y}|M_r)p(M_r). \tag{1.37}$$

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

[10] 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.

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

$$p(\Delta|\boldsymbol{y}) = \sum_{r=1}^{2^p} p(\Delta|\boldsymbol{y}, M_r)p(M_r|\boldsymbol{y}). \tag{1.38}$$

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

$$p(\boldsymbol{\beta}_r, \sigma^2|M_r) = N_{p_r}(\mathbf{0}_{p_r}, \sigma^2\boldsymbol{D}_r) \times Inv-Gamma\left(\frac{v_0}{2}, \frac{s_0^2}{2}\right). \tag{1.39}$$

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 weakly informative 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 (1.20) 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 $\boldsymbol{V}_r = \sigma^2(\boldsymbol{X}_r'\boldsymbol{X}_r + (\tau \boldsymbol{I}_{p_r})^{-1})^{-1}$ 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.[11]

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 (1.7) 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

$$\boldsymbol{\beta}_r | \sigma^2, M_r \sim N_{p_r}\left(\mathbf{0}_{p_r}, \frac{1}{g}\sigma^2(\boldsymbol{X}_r'\boldsymbol{X}_r)^{-1}\right), \tag{1.40}$$

where $\sigma^2(\boldsymbol{X}_r'\boldsymbol{X}_r)^{-1}$ 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 $\boldsymbol{V}_r = \frac{1}{1+g} \times \sigma^2(\boldsymbol{X}_r'\boldsymbol{X}_r)^{-1}$, such that the posterior variance is uniformly affected by selection of g. Consequently, the posterior mean/mode is

$$\boldsymbol{\beta}_r^\star = \frac{1}{1+g}\widehat{\boldsymbol{\beta}}_r. \tag{1.41}$$

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

[11] 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!

way[12] 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 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

$$\boldsymbol{\beta}_r|\sigma^2, M_r \sim N_{p_r}\left(\mathbf{0}_{p_r}, \frac{1}{g}\sigma^2(\boldsymbol{X}_r'\boldsymbol{X}_r)^\lambda\right), \tag{1.42}$$

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,

[12] 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.

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 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$. 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 $\boldsymbol{\gamma} = (0, \ldots, 0)'$ and the model with all predictors is indexed as $\boldsymbol{\gamma} = (1, \ldots, 1)'$. 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. The theoretical underpinning of Bayesian variable selection are well-understood in linear regression with both Gaussian (Hoeting *et al.*, 1999) and non-Gaussian (Klein and Smith, 2021; Kundu and Dunson, 2014) errors, as well as nonparametric regression (Kohn *et al.*, 2001; Smith and Kohn, 1996). At the same time, variable selection with the g-prior provides the ground for some of the most interesting Bayesian work on computation in high-dimensional settings.[13] Ultimately, modern inference with g-prior relies heavily on the benefits of a hierarchical Bayes modeling. 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.

[13] 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).

References

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

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

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

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

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

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

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

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

Assmann, C., J. Boysen-Hogrefe, and M. Pape (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 B. K. Mallick (2004). "Gene selection using a two-level hierarchical Bayesian model". *Bioinformatics*. 20(18): 3423–3430.

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

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

Baumeister, C., D. Korobilis, and T. K. Lee (2020). "Energy markets and global economic conditions". *The Review of Economics and Statistics*: 1–45.

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

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

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

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

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

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

Bernanke, B. S., J. Boivin, and P. Eliasz (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., J. Datta, Y. Li, and N. Polson (2020). "Horseshoe regularisation for machine learning in complex and deep models". *International Statistical Review*. 88(2): 302–320.

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

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

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

Bitto, A. and S. Frühwirth-Schnatter (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., A. Kucukelbir, and J. D. McAuliffe (2017). "Variational inference: A review for statisticians". *Journal of the American Statistical Association*. 112(518): 859–877.

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

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

Brown, P. J., M. Vannucci, and T. Fearn (1998). "Multivariate Bayesian variable selection and prediction". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 60(3): 627–641.

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

Carbonetto, P. and M. Stephens (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 A. Doucet (2008). "Sparse Bayesian nonparametric regression". In: *Proceedings of the 25th International Conference on Machine Learning. ICML '08*. New York, NY, USA: ACM. 88–95.

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

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

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

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

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

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

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

Chen, J. and Z. Chen (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 I. Jeliazkov (2001). "Marginal likelihood from the Metropolis–Hastings output". *Journal of the American Statistical Association*. 96(453): 270–281.

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

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

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

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

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

Datta, J. and J. K. Ghosh (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 F. Spezzaferri (1997). "Alternative Bayes factors for model selection". *Canadian Journal of Statistics*. 25(4): 503–515.

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

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

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

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

Efron, B. and C. Morris (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., C. Papageorgiou, and A. E. Raftery (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., E. Ley, and M. F. Steel (2001a). "Benchmark priors for Bayesian model averaging". *Journal of Econometrics*. 100(2): 381–427.

Fernández, C., E. Ley, and M. F. J. Steel (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 E. I. George (1994). "The risk inflation criterion for multiple regression". *The Annals of Statistics*. 22(4): 1947–1975.

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

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

Frazier, D. T., R. Loaiza-Maya, and G. M. Martin (2022). "Variational Bayes in state space models: Inferential and predictive accuracy". *Tech. rep.* arXiv:2106.12262, ArXiV.

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

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

Gelfand, A. E. and D. K. Dey (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 S. K. Ghosh (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., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC.

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

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

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

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

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

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

Geweke, J. (1999). "Using simulation methods for Bayesian econometric models: Inference, development, and communication". *Econometric Reviews*. 18(1): 1–73.

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

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

Giannone, D., M. Lenza, and G. E. Primiceri (2015). "Prior selection for vector autoregressions". *The Review of Economics and Statistics*. 97(2): 436–451.

Giordano, R., T. Broderick, and M. I. Jordan (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 C. P. Robert (1998). "Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections". *Biometrika*. 85(1): 29–37.

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

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

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

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

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

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

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

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

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

Ji, C. and S. C. Schmidler (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., P. Orenstein, and A. Bhattacharya (2020). "Scalable approximate MCMC algorithms for the Horseshoe prior". *Journal of Machine Learning Research*. 21(73): 1–61.

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

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

Judge, G. G., W. E. Griffith, R. C. Hill, H. Lütkepohl, and T.-C. Lee (1985). *The Theory and Practice of Econometrics*. New York: Wiley.

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

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

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

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

Kass, R. E. and L. Wasserman (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 C. Schumacher (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 J. P. Hobert (2012). "Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression". *Journal of Multivariate Analysis*. 112: 108–116.

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

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

Klein, N. and T. Kneib (2016). "Scale-dependent priors for variance parameters in structured additive distributional regression". *Bayesian Analysis*. 11(4): 1071–1106.

Klein, N. and M. S. Smith (2021). "Bayesian variable selection for non-Gaussian responses: A marginally calibrated copula approach". *Biometrics*. 77(3): 809–823.

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

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

Kohn, R., M. Smith, and D. Chan (2001). "Nonparametric regression using linear combinations of basis functions". *Statistics and Computing*. 11(4): 313–322.

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

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

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

Koop, G. and D. Korobilis (2018). "Bayesian dynamic variable selection in high dimensions". *Tech rep.* arXiv:1809.03031, ArXiV.

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

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

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

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

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

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

Kundu, S. and D. B. Dunson (2014). "Bayes variable selection in semi-parametric linear models". *Journal of the American Statistical Association*. 109(505): 437–447.

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

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

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

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

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

Lewis, S. M. and A. E. Raftery (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 D. Pati (2017). "Variable selection using shrinkage priors". *Computational Statistics & Data Analysis*. 107: 107–119.

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

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

Lim, D., B. Park, D. Nott, X. Wang, and T. Choi (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., V. Ročková, and Y. Wang (2021). "Variable selection with ABC Bayesian forests". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 83(3): 453–481.

Loaiza-Maya, R., M. S. Smith, D. J. Nott, and P. J. Danaher (2021). "Fast and accurate variational inference for models with many latent variables". *Journal of Econometrics*.

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

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

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

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

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

Matusevich, D. S., W. Cabrera, and C. Ordonez (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 J. J. Beauchamp (1988). "Bayesian variable selection in linear regression". *Journal of the American Statistical Association*. 83(404): 1023–1032.

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

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

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

Nakajima, J. and M. West (2015). "Dynamic network signal processing using latent threshold models". *Digital Signal Processing*. 47: 6–15.

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

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

Neville, S. E., J. T. Ormerod, and M. P. Wand (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 R. Kohn (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 M. J. Sillanpää (2009). "A review of Bayesian variable selection methods: What, how and which". *Bayesian Analysis*. 4(1): 85–117.

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

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

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

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

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

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

Peltola, T., P. Marttinen, and A. Vehtari (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.

Quintero, A. and E. Lesaffre (2018). "Comparing hierarchical models via the marginalized deviance information criterion". *Statistics in Medicine*. 37(16): 2440–2454.

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., D. Sparks, K. Khare, and L. Zhang (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*. Vol. Springer Texts in Statistics. Springer-Verlag New York.

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

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

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

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

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

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

Schwarz, G. (1978). "Estimating the dimension of a model". *Annals of Statistics*. 6(2): 461–464.

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

Simpson, D., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye (2017). "Penalising model component complexity: A principled, practical approach to constructing priors". *Statistical Science*. 32(1): 1–28.

Smith, M. and R. Kohn (1996). "Nonparametric regression using Bayesian variable selection". *Journal of Econometrics*. 75(2): 317–343.

Smith, M. and R. Kohn (2000). "Nonparametric seemingly unrelated regression". *Journal of Econometrics*. 98(2): 257–281.

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

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

Srivastava, S., B. E. Engelhardt, and D. B. Dunson (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: *Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. University of California Press. 197–206.

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.

Tran, M.-N., D. J. Nott, and R. Kohn (2017). "Variational Bayes with intractable likelihood". *Journal of Computational and Graphical Statistics*. 26(4): 873–882.

Uribe, P. and H. Lopes (2017). "Dynamic sparsity on dynamic regression models". *Tech. rep.* URL: http://hedibert.org/wp-content/uploads/2018/06/uribe-lopes-Sep2017.pdf.

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

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

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

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

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

Verdinelli, I. and L. Wasserman (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 A. E. Raftery (2000). "Bayesian information criterion for censored survival models". *Biometrics*. 56(1): 256–262.

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

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

Wang, Y. and D. M. Blei (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: *Bayesian Statistics 7*. Ed. by J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West. Oxford University Press. 723–732.

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

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

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

Yuan, M. and Y. Lin (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: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Ed. by P. Goel and A. Zellner. New York: Elsevier Science Publishers, Inc. 233–243.

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

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

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