2011/61

■

Bayesian methods

Luc Bauwens and Dimitris Korobilis

DISCUSSION PAPER

Center for Operations Research
and Econometrics

Voie du Roman Pays, 34
B-1348 Louvain-la-Neuve
Belgium
http://www.uclouvain.be/core

CORE DISCUSSION PAPER
2011/61

Bayesian methods

Luc BAUWENS [1] and Dimitris KOROBILIS[2]

December 2011

Abstract

Chapter written for the Handbook of Research Methods and Applications on Empirical Macroeconomics, edited by Nigar Hashimzade and Michael Thornton, forth- coming in 2012 (Edward Elgar Publishing). This chapter presents an introductory review of Bayesian methods for research in empirical macroeconomics

Keywords: Bayesian inference, dynamic regression model, prior distributions, MCMC methods.

JEL Classification: C11

[1] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: luc.bauwens@uclouvain.be. This author is also member of ECORE, the association between CORE and ECARES.
[2] University of Glasgow.

Research supported by the contract "Projet d'Actions de Recherche Concertées" 07/12-002 of the "Communauté française de Belgique", granted by the "Académie universitaire Louvain".

This paper presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

1 Introduction

The scope of this Chapter is to introduce applied macroeconomists to the world of Bayesian estimation methods. Why would an empirical macroeconomist invest in learning Bayesian estimation after having invested hours learning estimation methods like maximum likelihood and generalized method of moments (see previous Chapters)?

Nowadays, with the advancement of computing power and the establishment of new simulation techniques, it is probably much easier to answer this question compared to, say, thirty years ago. First, Bayesian methods offer a wide range of estimation tools for macroeconomic models, ranging from simple time series models to structural macroeconometric models. When optimizing the likelihood function becomes a daunting task (due to its high dimensionality or multimodality, or due to under-identification of specific model parameters), Bayesian methods can prove more robust since they do not rely on using complex optimization techniques that might get stuck in local optima. Second, Bayesian methods allow the researcher to incorporate prior beliefs in her model. We argue in this chapter that such beliefs are not so difficult to formalize as one may fear a priori, and that they may help to get reasonable estimates and forecasts, e.g. through soft shrinkage constraints.

The purpose of this Chapter is to provide the reader with a soft introduction to the Bayesian world. Exhaustively presenting the advances of Bayesian methods in macroeconomics is beyond the scope of this Chapter. The interest reader should consult Bauwens et al. (1999), Koop (2003), Lancaster (2004) and Geweke (2005), in order to delve deeper into Bayesian methods in econometrics. We start by introducing the basic principles of Bayesian inference and the numerical tools that are necessary for its implementation (Section 2). We apply all this in Section 3 to the dynamic linear regression model. The last section contains a short guide to the Bayesian literature for more sophisticated models.

2 Basics of Bayesian inference

2.1 Prior, posterior and likelihood

At the core of the Bayesian paradigm lies 'Bayes Theorem'. This is our starting point for statistical inference. Assume two random events A and B, and a probability measure P such that $P(B) \neq 0$, then Bayes Theorem states that

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

It holds, by definition of conditional probability, that $P(A|B) = P(A \cap B)/P(B)$ and similarly $P(B|A) = P(A \cap B)/P(A)$, which gives $P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A)$, and by rearranging terms we end up with Bayes theorem. However this exposition from probability theory may not be interesting to the applied econometrician who wants to estimate a parametric model which could potentially be connected to economic theory.

Subsequently assume that the random event B is a stochastic process y from which our observed data (for instance US inflation) occur, and the random event A is our parameters θ which take values in a space Θ. An econometric model might have other assumptions apart from observed data and model parameters, but we can ignore them at this level of generality. The parameters θ might come from a regression model, a vector autoregressive (VAR) model, a dynamic stochastic general equilibrium (DSGE) model, or another type of model. We can write Bayes Theorem in terms of y and θ in the following form:

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)}. \tag{1}$$

where p denotes generically a probability density function. The density $p(\theta)$ is representing our prior beliefs about the values of our parameters. The conditional density of the data given the parameters, $p(y|\theta)$, is the familiar likelihood function $L(\theta; y)$ (once y is 'realized' into observations). The density $p(y)$ is the marginal 'likelihood' or prior predictive density of the data. It is the density $p(y|\theta)$ marginalized with respect to the parameters θ, that is

$$p(y) = \int_{\Theta} p(y|\theta)\, p(\theta)\, d\theta, \text{when} \theta \text{is continuous} \tag{2}$$

$$p(y) = \sum_{\Theta} p(y|\theta)\, p(\theta), \text{when} \theta \text{is discrete} \tag{3}$$

The resulting conditional density $p(\theta|y)$ in Eq. (1) is called the posterior of the parameters: it is the density of the parameters after observing the data (hence conditional on y). Note that $p(y)$ does not depend on θ since it is integrated out from this density. Subsequently the formula of the posterior can be written as

$$p(\theta|y) \propto p(y|\theta)\, p(\theta) \tag{4}$$

or

$$p(\theta|y) \propto L(\theta; y)\, p(\theta), \tag{5}$$

where the symbol $\propto$ means 'proportional to' and is used extensively to avoid writing uninteresting constants in many formulas. This formulation makes it clear that the posterior density of the parameters, $p(\theta|y)$, updates our prior beliefs (before seeing the data) with the data information embedded in the likelihood function that measures the probability that our data come from the specified model.

Consequently, the Bayesian econometrician tries to average the likelihood with the prior $p(\theta)$, whereas the frequentist tries to maximize the likelihood $L(\theta; y) \propto p(y|\theta)$. The equation above has several implications. One of those is that we assumed that we can assign a probability density not only to our data, but also to our parameters θ. That is, the parameters are considered as random variables, with a well defined probability density. Subsequently, for the Bayesian, a parameter θ is not only identified by the likelihood, but also by the prior distribution, something that has very important implications in possibly underidentified macroeconometric models, especially DSGE models.

Thus, what is common to Bayesian and frequentist econometricians is the need to define a model that is assumed to have generated the data, yielding a likelihood function. What differentiates them is that the Bayesian finds it convenient to treat the parameters as random variables, while the frequentist views the parameters as unknown fixed constants. Notice that these viewpoints are not necessarily contradictory, since the Bayesian is just adding that unknown true constants may be the object of (subjective or objective) probability judgements, precisely because they are unknown. The frequentist then maximizes the likelihood (or another) function and reports point estimates and typically asymptotic standard errors. The Bayesian computes the posterior by integral calculus and reports posterior means as point estimates and posterior standard deviations as measures of uncertainty about these estimates.

The presence of a prior distribution is thus one of the main elements that differentiate substantially frequentist from Bayesian analysis. At the same time, the prior is a useful tool for the modern macroeconomist in that it allows to incorporate beliefs from economic theory, personal experience and opinions about the structure or the future of the economy coming from analysts in business, academia, or simply consumers' beliefs. Given that macroeconomic series at monthly or quarterly frequencies are short, thus making the likelihood not very informative (especially when trying to model the economy using many variables), the prior often plays a favorable role in producing reasonable results. Subsequently the prior should be seen as an ally for the applied macroeconomist, and not as one more trouble to solve during the process of statistical inference. In the next section, we discuss in detail the elaboration of priors for the parameters of regression models, and we give examples of informative priors that have been adopted by applied econometricians.

This explains why applied macroeconomics is one of the few fields in economics where the philosophical dispute about being subjective (as opposed to being objective) plays less of a significant role. Nevertheless, Bayesian analysis allows us to produce results that are comparable to maximizing the likelihood (objective analysis). For instance, it is customary in Bayesian econometrics to use a noninformative prior on a parameter θ when we have no prior information about this parameter. Suppose that this parameter is the mean of a normal density with possible values (support) on the space $(-\infty, +\infty)$ and that since we have no information to restrict this support a priori,[1] we use the uniform prior $p(\theta) \propto c$ (an arbitrary constant), so that all values of θ are a priori equally likely. This leads to write the posterior as $p(\theta|y) \propto L(\theta; y) \times c \propto L(\theta; y)$, so that the posterior is proportional to the likelihood function (up to a normalizing constant that does not involve θ). Subsequently the mode of the posterior density is equal to the maximum likelihood estimate (MLE), and the posterior mean is close to the MLE if the likelihood is symmetric around its mode.

Finally we should note that another benefit of using Bayesian analysis is that the computation of the posterior (see next subsection) usually does not require optimization techniques that can fail in complex models, when the likelihood function is multimodal or

[1] As an alternative, if θ is the mean of GDP growth, a prior belief would be to restrict this parameter on the interval, say, -20 to 25 per cent, implying that we do not realistically expect to observe growth beyond these bounds. Letting the bounds become arbitrarily large may be viewed as a convenient simplification.

highly dimensional. Additionally, basing predictions (that are often essential for applied macroeconomists) on the predictive density of future observations means that uncertainty about the values of the parameters is accounted for. Let y_f denote future observations to be predicted, i.e. y_f occurs after our sample y that serves to compute $p(\theta|y)$. Then the predictive density of y_f is obtained as

$$p(y_f|y) = \int p(y_f|y,\theta)p(\theta|y)d\theta. \tag{6}$$

The integrand in the previous formula shows that the density of future observations conditional on the past data and the parameters (i.e. the likelihood function of the potential future data) is weighted by the evidence obtained about the parameters from the past data. The integration means that this evidence is taken into account for all possible values of θ, and not only at the MLE as in a frequentist analysis. Indeed, a natural frequentist point predictor of y_f is $E(y_f|y,\hat{\theta})$, where $\hat{\theta}$ is the MLE, whereas a Bayesian point predictor is naturally taken to be $E(y_f|y)$ (or any other central measure like the median). Notice that $E(y_f|y) = E[E(y_f|y,\theta)]$, implying that $E(y_f|y,\hat{\theta})$ is just one of the averaged values in the Bayesian formula.

2.2 Methods to compute the posterior

In essence all Bayesian computation techniques try to estimate the formula in (5). In the Bayesian context, 'estimate' means finding the normalizing constant of the posterior (which is nothing else but $p(y)$), and, more importantly, whatever features of it are of interest. Such features typically include the expected value, the covariance matrix, univariate marginal densities of θ and their quantiles, and often also of functions of θ, denoted by $g(\theta)$. In general the quantities we wish to compute can be expressed as

$$E[g(\theta)|y] = \int g(\theta)p(\theta|y)d\theta, \tag{7}$$

where it is understood that $p(\theta|y)$ is a properly normalized density, i.e. $\int p(\theta|y)d\theta = 1$. To be concrete, if $g(\theta) = \theta$, we get the expected value of θ, if $g(\theta) = \theta\theta'$, we get the matrix of uncentered second moments, and from these two, we get easily the covariance matrix of θ. If we are interested in the posterior probability that the i-th element of θ is in the interval (a,b), we define $g(\theta)$ as the indicator function that is equal to one if $\theta_i \in (a,b)$ and to zero otherwise. Moreover, if $g(\theta)$ is taken to be $p(y_f|y,\theta)$, we get $p(y_f|y)$, and if it is $E(y_f|y,\theta)$ (which may be available analytically), we get $E(y_f|y)$.

For a few (simple) econometric models (in particular the normal linear regression model), some choices of prior densities, and some simple functions g, there is no need to compute the integrals above by numerical methods, because they are known analytically. For example, if the posterior is a Normal density, we know all its moments, but if we are interested in a special function of θ, we may not know the result.[2] In this case,

[2]For example, if $\theta = (\theta_1\,\theta_2)$, and we are interested in θ_1/θ_2, we do not know the density of this ratio.

we need numerical methods to compute integrals as in (7). We also need these numerical integration techniques in most econometric models. This is due to the fact that their likelihood function is so complex that whatever the prior we choose and the functions g we are interested in, the integrals are not known analytically. In this kind of situation, the typical tool to compute integrals as in (7) is Monte Carlo simulation (see also Chapter 17 of this Handbook). The principle of this technique is to simulate on a computer a large sample of values of θ that are distributed according to the posterior density $p(\theta|y)$. Let us denote by $\{\theta^{(r)}\}_{r=1}^{R}$ a sample of size R of such values, called (random) draws or replicates (of the posterior). Then we can estimate consistently[3] $E[g(\theta)|y]$ (if it is finite) by the sample mean of the draws, i.e.

$$\frac{1}{N}\sum_{r=1}^{R} g(\theta^{(r)} \xrightarrow{p} E[g(\theta)|y] \tag{8}$$

as R tends to infinity. We present next the most useful ways of generating draws of the posterior density of the parameters of econometric models.

2.2.1 Direct Sampling

As mentioned above, in linear regression models under the normality assumption, it is possible to obtain analytically the posterior. Write the regression model for observation t as $y_t = \beta' x_t + \varepsilon_t$, where x_t and β are vectors of k elements, $\varepsilon_t \sim N(0, \sigma^2)$ and $t = 1, 2, \ldots, T$. In matrix form, this is written as $y = X\beta + \varepsilon$ where X is the matrix of T observations on the k regressors, and we assume $T > k$. The likelihood function is then

$$L(\beta, \sigma^2; y, X) \propto (\sigma^2)^{-T/2} \exp\left[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\right]. \tag{9}$$

We use the noninformative prior $p(\beta, \sigma^2|) \propto 1/\sigma^2$. This means that we consider each regression coefficient and the logarithm of σ^2 to be uniformly distributed on the real line.[4] For sure, such a prior is not a density, since it is not integrable ('improper'). However the posterior is integrable (proper) as we explain below. The reader who feels uneasy about using an improper prior may be reassured by the following argument: instead of saying that a regression coefficient is uniformly distributed on $(-\infty, +\infty)$, we could decide that it should be uniformly distributed on the bounded interval $(-B, +B)$. By choosing B to be very large but finite, the prior is proper and the posterior will be the same as if we used the improper uniform prior. A way to choose B that ensures this equivalence is to choose B as the smallest value such that when the likelihood function is evaluated at $\beta = \pm B$, the computer returns zero (an underflow). It would be a waste of time to search for this value, so that using an improper uniform prior is a convenient shortcut.

By multiplication of the likelihood and the prior, we get the posterior

$$p(\beta, \sigma^2|y, X) \propto (\sigma^2)^{-(T+2)/2} \exp\left[-\frac{(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) + s^2}{2\sigma^2}\right] \tag{10}$$

[3] Invoking the ergodic theorem, we do not even need that the sample be independent.

[4] If $p(\log \sigma^2) \propto 1$, then $p(\sigma^2) \propto 1/\sigma^2$ due to the Jacobian $\partial \log \sigma^2 / \partial \sigma^2$.

where $\hat{\beta} = (X'X)^{-1}X'y$ is the OLS estimator and $s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})$ is the sum of squared OLS residuals. The equality of $(y - X\beta)'(y - X\beta)$ and $(\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) + s^2$ can be checked by using the definitions of $\hat{\beta}$ and s^2 in the latter. To obtain the posterior density of β alone, we must integrate the above expression with respect to σ^2. This yields a proper density if $T > k$, given by

$$p(\beta|y, X) \propto \left[(\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) + s^2\right]^{-\frac{\nu+k}{2}}, \tag{11}$$

where $\nu = T - k$ is the degrees of freedom parameter. The posterior density of β is a multivariate t, with parameters $\hat{\beta}$, $X'X$, s^2 and ν. To write it as a normalized density, we should multiply the expression in the above formula by some constants (see Bauwens et al. (1999), Appendix A, for a complete definition and properties). Using the properties of the t density, we can state that the posterior mean of β is $\hat{\beta}$, and its posterior covariance matrix is $s^2(X'X)^{-1}/(T - k - 2)$. Thus the posterior mean is the OLS estimator, and the posterior covariance differs only slightly from the covariance of the OLS estimator, which is equal to $s^2(X'X)^{-1}/(T - k)$. For the reader who is not familiar with the multivariate t density, but well with the Normal, the difference disappears as ν tends to infinity. Thus if the sample size is large relative to the number of regressors, the posterior (11) is very well approximated by the $N(\hat{\beta}, s^2(X'X)^{-1}/(T - k))$ density.

This is an example of a model and prior which give numerical results that are quasi-identical from the frequentist and Bayesian perspectives. However the interpretation of the results is different. The Bayesian says that given the observed unique sample that is available, the unknown parameter β has a posterior density centered on $\hat{\beta}$, while the frequentist says that the (sampling) distribution of $\hat{\beta}$ is centered on the (unknown) true value of β. Thus the frequentist refers to a hypothetically infinitely many times repeated experiment of sampling the data from the population model, whereas the Bayesian just refers to the single sample that has been observed.

Subsequently all we need to do to obtain features of the posterior density that are not known analytically is to generate independent draws of the t density with the parameters specified above. This can be done easily in any programming language using a random number generator from the t density (see Bauwens et al. (1999), Appendix B, for a generator). For example, if we are interested to generate the marginal posterior density of $(\beta_1 + \beta_2)/(1 - \beta_3)$, we proceed as follows:

1. *Generate R draws $\{\beta^{(r)}\}_{r=1}^{R}$ of β from the t density with parameters $\hat{\beta}$, $X'X$, s^2 and ν.*
2. *Compute $(\beta_1^{(r)} + \beta_2^{(r)})/(1 - \beta_3^{(r)})$ for $r = 1, 2, \ldots, R$.*
3. *Use a kernel method to obtain an estimate of the posterior density.*

We draw the reader's attention to the fact that the posterior mean of $(\beta_1 + \beta_2)/(1 - \beta_3)$ does not exist, so that a point estimator of that quantity could be the median of its posterior. The median is obtained by ordering the R draws of $(\beta_1^{(r)} + \beta_2^{(r)})/(1 - \beta_3^{(r)})$ by increasing value and selecting the value ranked in position $R/2$ (if R is an even number).

2.2.2 Methods to simulate posteriors that are not tractable

There are cases where multiplying the prior with the likelihood gives a mathematical expression of the posterior density that that does not belong to any known family of densities or is not easy to simulate by direct sampling. If this happens, we are unable to simulate directly samples $\{\theta^{(r)}\}_{r=1}^{R}$ from the posterior of the parameters using known random number generators. In these cases we need to rely on other simulation methods. One useful class of methods to do this is called 'Markov chain Monte Carlo' (MCMC). Two MCMC sampling methods, very much used, are the Metropolis-Hastings (MH) algorithm and the Gibbs algorithm. The 'Markov chain' part of the name reveals that the samples generated by these methods have a 'Markov property', implying that the draws are not independent, contrary to the direct sampling method. To explain MCMC, we first define the Markov property.

A Markov Chain is a stochastic process (i.e. a sequence of random variables) which has the Markov property, i.e. the probability of the next state of the stochastic process depends only the current state, and not on any other state in the distant past. Formally, consider a process $\{s_t\}$, $t = 1, 2, \ldots, T$. If the probability of moving from one state s_t to the next s_{t+1} satisfies $P\left(s_{t+1}|s_t, s_{t-1}, \ldots, s_1\right) = P\left(s_{t+1}|s_t\right)$, the process satisfies the Markov property. If the variable s_t is continuous, the above property holds using densities (i.e. P is replaced by p). All you need to remember in order to understand MCMC is that we have a variable s_t, with initial state (or initial condition) s_0, and *transition density* $p\left(s_{t+1}|s_t\right)$ which satisfies the Markov property.

Gibbs sampling In the rest of this section, we denote the posterior by $p(\theta)$ instead of $p\left(\theta|y\right)$. The Gibbs sampler requires us to partition the parameter vector θ (of k elements) into b sub-vectors ('blocks'), with $b \leq k$, denoted by $\theta_{[i]}$, i.e. $\theta = (\theta'_{[1]}\theta'_{[2]}\ldots\theta'_{[b]})'$, such that for each block, the 'full' conditional density $p(\theta_{[i]}|\theta_{-[i]})$, where $\theta_{-[i]}$ denotes θ without $\theta_{[i]}$, can be directly simulated. To generate a sample of size R of draws from $p(\theta)$ (after warming-up the algorithm with R_0 draws), the algorithm proceeds as follows:

1. *Choose an initial value* $\theta^{(0)}_{-[1]}$ *that belongs to the parameter space.*
2. *Set* $r = 1$.
3. *Draw successively*

$$
\begin{array}{lll}
\theta^{(r)}_{[1]} & \textit{from} & p(\theta_{[1]}|\theta^{(r-1)}_{-[1]}) \\
\theta^{(r)}_{[2]} & \textit{from} & p(\theta_{[2]}|\theta^{(r)}_{[1]},\, \theta^{(r-1)}_{[3]}, \ldots, \theta^{(r-1)}_{[b]}) \\
\vdots & & \\
\theta^{(r)}_{[i]} & \textit{from} & p(\theta_{[i]}|\theta^{(r)}_{[1]}, \ldots, \theta^{(r)}_{[i-1]}, \theta^{(r-1)}_{[i+1]}, \ldots, \theta^{(r-1)}_{[b]}) \\
\vdots & & \\
\theta^{(r)}_{[b]} & \textit{from} & p(\theta_{[b]}|\theta^{(r)}_{-[b]}).
\end{array}
$$

4. *Set $r = r + 1$ and go to step 3 unless $r > R_0 + R$.*

5. *Discard the first R_0 values of $\theta^{(r)} = (\theta_{[1]}^{(r)'}\theta_{[2]}^{(r)'} \dots \theta_{[b]}^{(r)'})'$. Compute what you are interested in (estimates of posterior means, variances...) from the last R generated values.*

In step 3, we sample successively from the full conditional posterior densities of each block. Each full conditional density is updated by the values of the previously generated blocks in the current iteration (r), while the blocks that have not yet been generated are set at the values of the previous iteration $(r-1)$. This creates the dependence in the sample (through the Markov property). Remark that if there is only one block, the method boils down to direct sampling, which should be used whenever possible. The number of blocks should be chosen as small as possible but this choice is constrained by our ability to find full conditional densities that can be directly simulated. It may happen that for some blocks, we are not able to perform the direct simulation of the full conditional. We then resort to an indirect method to sample from the full conditional (a 'Metropolis step', see Section 3.2.2 for an example) within the Gibbs sampler.

Notice that we must warm-up the algorithm with R_0 draws that are discarded. The purpose of this is to get rid of the impact of the initial value $\theta_{[1]}^{(0)}$ and to let the algorithm converge to the target distribution. Convergence means that the sampled values of θ are a valid sample from the target. The issue of convergence is important and too often overlooked by applied researchers. Though it is often easy to state theoretical conditions for convergence,[5] it is not possible to prove convergence practically (for a given run). There exists convergence diagnostics that should always be used and reported (see e.g. Bauwens et al. (1999), Ch. 3, 90-92 for details and references).

Metropolis-Hastings algorithm This is a useful algorithm when Gibbs sampling is not applicable, because there is no way to partition the parameter θ into blocks whose full conditional densities are easy to simulate. The MH algorithms requires to elaborate a density that approximates the target (posterior) and is easy to simulate (e.g. a Normal, a finite mixture of Normals...). Parameter values are drawn form the approximating density (called candidate density) and subject to an acceptance-rejection test to decide if the drawn value (called candidate) is a draw of the target, in which case it is kept as a valid draw $\theta^{(r)}$. If the candidate is rejected, the previously accepted draw is accepted once again (i.e. $\theta^{(r)} = \theta^{(r-1)}$). Thus there will be sequences of identical draws in the posterior sample, which directly indicates that the draws are dependent. If the approximating density is identical to the target, all draws are accepted and the method boils down to direct simulation of independent draws of the posterior. Thus, the approximating density should be designed to be as close as possible to the target, which is more easily said than done in large dimension.

The candidate density may depend on the last accepted draw and is therefore denoted by $q(\theta|\theta^{(r-1)})$. The steps of the MH algorithm are:

[5] A sufficient condition is that the full conditional densities are always strictly positive in the parameter space.

1. *Set* $r = 1$. *Choose an initial value* $\theta^{(0)}$ *that belongs to the parameter space.*

2. *Draw* $\theta^{(cand)} \sim q(\theta|\theta^{(r-1)})$. *Compute* $p = \min\left\{\frac{p(\theta^{(cand)})}{p(\theta^{(r-1)})}\frac{q(\theta^{(r-1)}|\theta^{(cand)})}{q(\theta^{(cand)}|\theta^{(r-1)})}, 1\right\}$.

3. *Set* $\theta^{(r)} = \theta^{(cand)}$ *with probability* p, *and set* $\theta^{(r)} = \theta^{(r-1)}$ *with probability* $1 - p$.

4. *Set* $r = r + 1$ *and go to step 2 unless* $r > R_0 + R$.

5. *Identical to step 5 of the Gibbs algorithm.*

Step 3 is implemented by drawing a random number U from the Uniform$(0, 1)$ density, and if $U < p$, setting $\theta^{(r)} = \theta^{(cand)}$, otherwise to $\theta^{(r-1)}$). The ratio in the min of Step 2 is called the MH ratio. It may be larger than one, in which case the candidate is accepted surely. Indeed it is the ratio of the posterior to candidate densities evaluated at the candidate draw, to the same ratio evaluated at the previous draw. If that ratio is larger than one, the new candidate must be accepted. If it is smaller than one, it is accepted only with probability $p < 1$.

Some choices of proposal density are of interest. If q does not depend on $\theta^{(r-1)}$), the algorithm is known as the independent MH algorithm. If q is symmetric in the sense that $q(\theta^{(r-1)}|\theta^{(cand)}) = q(\theta^{(cand)}|\theta^{(r-1)})$ the MH ratio is simplified. One way to let q depend on $\theta^{(r-1)}$ is the random walk MH algorithm. This generates $\theta^{(cand)}$ as $\theta^{(r-1)} + v$ where v is a draw from a distribution (e.g. Normal) centered on 0 and with a variance matrix to be selected not too small so as to allow the candidate draw to walk in the parameter space without staying too close to the previous draw.

3 Linear regression model

Since choosing a prior is an important step in an application using Bayesian inference, and this step may look like a daunting task (which is it not), we provide in this section several useful approaches to do this in the context of the dynamic linear regression model. We also describe the corresponding algorithms to compute the posterior.

Consider a univariate time-series of interest y_t (GDP, price inflation...) observed over the period $t = 1, ..., T$. The empirical macroeconomist usually assumes that y_t depends on an intercept, some own lags, and current or past values of some predictor variables z_t. Then a popular model for y_t is the dynamic regression model of the form

$$y_t = \kappa + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{j=0}^{q} \lambda_j' z_{t-q} + \varepsilon_t \tag{12}$$

with the usual assumption that $\varepsilon_t \sim N(0, \sigma^2)$. This model can be cast in the standard regression form already introduced in Section 2.2.1, $y_t = \beta' x_t + \varepsilon_t$, where $x_t = (1, y_{t-1}', ..., y_{t-p}', z_{t-1}', ..., z_{t-q}')'$ collects all the regressors, and $\beta' = \left(\kappa, \varphi_1, ..., \varphi_p, \lambda_1', ..., \lambda_q'\right)$ the coefficients. The first thing the researcher needs to do is chose a 'sensible' prior for

the parameters $\theta = (\beta', \sigma^2)$. Here we will guide the reader step-by-step on what exactly Bayesians mean by choosing a sensible prior. We have already presented an easy-to-use and sensible prior in Section 2.2.1. It is actually a particular case of what is called a conjugate prior.

3.1 Conjugate priors

The uninformative prior presented in section 2.2.1 does not allow the researcher to add information: it just allows the likelihood to determine the final result about β. Presumably the researcher is free to use any other prior density of the form

$$p(\theta) = f(\theta|\underline{a}_1, \underline{a}_2, ..., \underline{a}_n) \tag{13}$$

where $f(\cdot)$ is a generic density function (Beta, Gamma, Dirichlet, Normal and so on) and $(\underline{a}_1, \underline{a}_2, ..., \underline{a}_n)$ are parameters of this distribution. In practice, there are three major considerations to keep in mind when deciding on the specific form of the prior:

1. The prior distribution must be congruent with the support of the parameters. As an example, an inverse-Gamma density[6] denoted by $iG(\nu$,q) has support on $[0, \infty)$, so it is not an appropriate choice for a regression coefficient β that is expected to take negative values, but it is appropriate for the variance parameter σ^2 of the regression model.

2. The prior must be of a form that allows to easily choose sensible values of the prior parameters $(\underline{a}_1, \underline{a}_2, ..., \underline{a}_n)$. Subsequently Bayesians usually focus on standard distributions, such as the Normal, Bernoulli, Beta, Gamma, Exponential etc. which have one or two prior parameters to choose.

3. The prior distribution must be such that the resulting posterior is either known analytically or easy to sample from using simulation methods such as reviewed in Section 2.2.

In that respect, for many models that are the focus of macroeconomists, default choices of prior densities exist. In practice, Bayesian macroeconomists usually focus on *conjugate priors* that have all three of the above properties. A prior distribution as (13) is said to be conjugate to the likelihood function $L(\theta; y)$ if the resulting posterior $p(\theta|y)$ also belongs to the family $f(\theta|\overline{a}_1, \overline{a}_2, ..., \overline{a}_n)$, but obviously the posterior parameters (overlined) update the prior ones (underlined) with some functions of the data.

[6]The iG(ν, q) density with $\nu > 0$ degrees of freedom and scale parameter $q > 0$, for the random variable U, is given by $[1/\Gamma(\nu/2)](q/2)^{\nu/2}u^{-(\nu+2)/2}\exp -[q/(2u)]$. Its mean is $q/(\nu-2)$ (if $\nu > 2$) and its variance is $2q^2/[(\nu-2)(\nu-4)]$ (if $\nu > 4$).

Conjugate prior for the regression model The conjugate prior on β is the Normal density $N\left(\underline{\beta}, \sigma^2\underline{V}_\beta\right)$ where the parameters $\underline{\beta}, \sigma^2\underline{V}_\beta$ are the prior mean and prior covariance matrix. Notice that for conjugacy the inclusion of σ^2 is needed as a proportionality factor in the prior covariance. This is not very convenient since it forces us to interpret the variances and covariances in units of σ, an unknown parameter, though an idea of its value is given by the usual OLS estimator. We explain how to avoid this problem in Section 3.2. Written explicitly, that normal prior is

$$p\left(\beta|\sigma^2\right) \propto (\sigma^2)^{-\frac{k}{2}}|\underline{V}_\beta|^{-\frac{1}{2}} \exp\left[-\frac{(\beta-\underline{\beta})'\underline{V}_\beta^{-1}(\beta-\underline{\beta})}{2\sigma^2}\right]. \tag{14}$$

Since the likelihood has the same functional form, i.e. Normal – see (9) and (10) –, the resulting posterior of β given σ^2 has the same form: $\beta|\sigma^2, y \sim N\left(\overline{\beta}, \sigma^2\overline{V}_\beta\right)$,[7] where

$$\overline{V}_\beta \;=\; \left(\underline{V}_\beta^{-1} + X'X\right)^{-1}, \quad \overline{\beta} = \overline{V}_\beta\left(\underline{V}_\beta^{-1}\underline{\beta} + X'X\hat{\beta}\right). \tag{15}$$

We notice an interesting feature: the posterior mean $\overline{\beta}$ does not depend on σ^2 and it is therefore the unconditional mean of β, $E(\beta|y)$, as well as its conditional mean, $E(\beta|y, \sigma^2)$. That posterior mean is a matrix weighted average of the prior mean and of the OLS estimator. On the contrary, the posterior covariance matrix $Var(\beta|y, \sigma^2)$ is proportional to σ^2 and is thus not the unconditional covariance matrix that we need to make inferences. For example, to compute a highest posterior density (HPD) interval for a particular element of β, we need to know its marginal variance.[8]

Since the posterior we have obtained for β is conditioned on σ^2, we must still marginalize it to use it for inferences. For this we need the posterior of σ^2, since marginalization means computing $\int p(\beta|\sigma^2, y)p(\sigma^2|y)d\sigma^2$ to get $p(\beta|y)$. The conjugate prior for σ^2 is the $\text{iG}\left(\underline{\nu}, \underline{q}\right)$ density. The resulting posterior density of σ^2 is $\text{iG}(\overline{\nu}, \overline{q})$, where

$$\overline{\nu} \;=\; \underline{\nu} + T, \quad \overline{q} = \underline{q} + y'y + \underline{\beta}'\underline{V}_\beta^{-1}\underline{\beta} - \overline{\beta}'\overline{V}_\beta^{-1}\overline{\beta}.$$

Finally, the marginal density of $\beta|y$ is multivariate t with parameters $\overline{\beta}$ (the mean), $\overline{V}_\beta^{-1}$, $\overline{\nu}$ (degrees of freedom) and $\overline{q}$, such that the posterior covariance matrix is $\overline{q}\overline{V}_\beta/(\overline{\nu}-2)$. The posterior mean of σ^2 is $\overline{q}/(\overline{\nu}-2)$, so that we have obtained that the posterior marginal (or unconditional) covariance of β is equal to the posterior mean of σ^2 multiplied by the matrix $\overline{V}_\beta$.

The above results are fully analytical and useful if we are just interested in β and σ^2. However, if we are interested in functions of these parameters, we may need to simulate

[7] For a detailed proof, see Bauwens et al. (1999), Ch. 2, 58-59.

[8] A HPD interval of (probability) level α for a scalar parameter θ is the shortest interval of values (θ_i, θ_s) such that $P[\theta \in (\theta_i, \theta_s)] = \alpha$. If the density of θ is $\text{N}(m, s^2)$, it is given by $(m - z_{\alpha/2}, m + z_{\alpha/2}s)$, where $z_{\alpha/2}$ is the quantile of level $\alpha/2$ of the right tail of the standard Normal. For a Bayesian, a HPD interval is an interval estimator. A HPD interval resembles a frequentist confidence interval, but it has a quite different interpretation: for a Bayesian, θ is random, for a frequentist, the interval is random.

the posterior. Though we can do this by direct simulation (as explained at the end of section 2.2.1) we can also use the Gibbs sampler to sample from the posterior. The algorithm iterates between the conditional posterior of β given σ^2, which is none other than $\mathrm{N}\big(\overline{\beta}, \sigma^2\overline{V}_\beta\big)$, and that of σ^2 given β, which can be easily shown to be $\mathrm{iG}(\overline{\nu}^*, \overline{q}^*)$, where

$$\overline{\nu}^* \;=\; \underline{\nu} + T + k, \quad \overline{q}^* = \underline{q} + (y - X\beta)'(y - X\beta) + (\beta - \underline{\beta})'\underline{V}_\beta^{-1}(\beta - \underline{\beta}).$$

Here is the Gibbs sampling algorithm to generate a sample of size R of draws from the posterior distribution of β and σ^2 (after warming-up the algorithm with R_0 draws):

1. *Choose an initial value* $(\sigma^2)^{(0)}$ *(e.g. the OLS sum of squared residuals divided by* $T - k$*).*
2. *Set* $r = 1$.
3. *Draw successively* $\beta^{(r)}$ *from* $\mathrm{N}\big(\overline{\beta}, (\sigma^2)^{(r-1)}\overline{V}_\beta\big)$ and $(\sigma^2)^{(r)}$ *from* $\mathrm{iG}(\overline{\nu}^*, (\overline{q}^*)^{(r)})$*, where* $(\overline{q}^*)^{(r)}$ *is* $\overline{q}^*$ *evaluated at* $\beta = \beta^{(r)}$.
4. *Set* $r = r + 1$ *and go to step 3 unless* $r > R_0 + R$.
5. *Discard the first* R_0 *values of* $\beta^{(r)}$ *and* $(\sigma^2)^{(r)}$*. Compute what you are interested in (estimates of posterior means, variances...) from the last* R generated *values.*

Noninformative conjugate prior It is worth mentioning at this point that even though conjugate priors are not by default noninformative, they can (almost) always become noninformative by taking their parameters to some limit. For instance, the bell-shaped Normal density becomes almost flat when its variance is large. Therefore, the conjugate prior $\mathrm{N}(0, \sigma^2 10^6 I_k)$ implies that for each element of β, values in the range (-1000,1000) are practically speaking 'equally likely' a priori. For the regression variance parameter, the inverse Gamma density becomes noninformative (variance close to infinity) when both $\underline{\nu}$ and $\underline{q}$ tend to zero. Then it is customary in practice to use the $\mathrm{iG}(0.001, 0.001)$ in the absence of prior information. The fully noninformative prior $p(\beta, \sigma^2|) \propto 1/\sigma^2$ of Section 2.2.1 is the conjugate prior for β, σ^2 given by $N\left(\underline{\beta}, \sigma^2\underline{V}_\beta\right) \times iG\left(\underline{\nu}, \underline{q}\right)$ when $\underline{\beta} = 0$ (a vector), $\underline{V}_\beta^{-1} = 0$ (a matrix) and $\underline{\nu} = \underline{q} = 0$ (scalars).

Practical recommendations for fixing $\underline{\beta}$ and $\underline{V}_\beta$ It is recommended to choose a value for the inverse of $\underline{V}_\beta$ since it is the inverse that appears in the formulas (15) defining the posterior parameters. The researcher who wants to be very little informative on an element of β should choose a very small positive value for the corresponding diagonal element of $\underline{V}_\beta$, and zero values for the off-diagonal elements on the corresponding row and column of that matrix. For the prior mean, the element of $\underline{\beta}$ should be fixed to zero. For being informative on an element of β, the researcher should assign his belief of what could be the most likely value of this coefficient to the corresponding element of $\underline{\beta}$. Such a belief may be inspired by theory or by empirical results on similar (but different) data. For

example, if a theory or past results suggest that the parameter should be between two values, the average of these is a sensible prior mean. The prior variance can then be fixed in such a way that with high prior probability the parameter lies in the interval in question. The corresponding diagonal element of $\underline{V}_{\beta}^{-1}$ is then just the inverse of this variance if one assumes, as is almost always the case, that $\underline{V}_{\beta}^{-1}$ is a diagonal matrix. More examples and ways to choose the prior parameters are discussed in Bauwens et al. (1999), Ch. 4.

A researcher might have a subjective opinion about a parameter of choice. For instance if in the dynamic regression model written in (12) the GDP growth rate is the dependent variable, one might want to incorporate the belief that the intercept should be in the bounds, say, -15 to 15 percent since the researcher might not expect with certainty to observe a growth rate beyond these bounds in her economy of interest. This could be translated to the subjective conjugate prior for the intercept $\kappa/\sigma \sim N(0, 9)$. This prior gives almost all the prior weight in the support $(-15, 15)$. Additionally, due to bell shape of the Normal distribution, more prior probability goes to values of the growth rate around zero and less probability is given to tail (extreme) values.

As another example consider the AR(1) coefficient in a dynamic regression model for the case of price inflation. The AR(1) coefficient is not expected to be more than one (the process can be assumed to be stationary or near-stationary, but definitely not explosive), hence the prior $\varphi_1/\sigma \sim \mathrm{N}(0, 1)$ seems more appropriate than the noninformative option $\mathrm{N}(0, 10^6)$, since it will attract the likelihood towards a more realistic posterior mean value. In models with several lags, the researcher might choose for more distant lags to use the prior $\varphi_i/\sigma \sim \mathrm{N}(0, 1/\mathrm{i})$, so that more distant lags are increasingly penalized.[9]

Other than these specific examples to elicit the prior parameters, macroeconomists (for instance working in Central Banks) have well defined economic theories to guide them empirically[10] as well as strong opinions about the state of the economy. In that respect, researchers have used estimates from national econometric models to form priors for regional models for which data are sparse or simply at yearly frequency, see Adkins et al. (2003). Other researchers have used priors informed from estimated DSGE models, see Ingram and Whiteman (1994).

The g-prior Zellner (1986) proposed to use a conjugate prior of the form

$$\beta|\sigma^2, X \sim \mathrm{N}\left(0, g\sigma^2 \left(X'X\right)^{-1}\right)$$

that is, scale the prior variance by using the inverse of the information matrix. The resulting Bayes estimate also provides shrinkage over the least squares estimate, since the posterior

[9] This is similar in spirit, but not identical to, the so-called Minnesota prior for VAR models; see Doan et al. (1984).

[10] At least, one can argue that macroeconomic theory can guide the empirical researcher on what outcome to expect, as well as what empirical result makes sense. In that respect, empirical results like the price puzzle in VAR models (the fact that inflation responds with an increase after a contractionary monetary policy shock), have been solved by using prior restrictions about the expected signs of the responses of each variable; see Uhlig (2001)

mean in (15) is simplified into

$$\overline{\beta} = \frac{g}{1+g}\hat{\beta}.$$

For $g \to \infty$ we get the least squares estimate, while for $g \to 0$ we have shrinkage towards zero (the prior mean). Despite the shrinkage properties, this prior has mainly been used because it allows analytical calculation of the marginal likelihood. The latter is a prevalent model choice criterion among Bayesian econometricians. Between two models for the same data y, the model that has the highest marginal likelihood is preferred.

This prior has been used extensively in situations where (macroeconomic) theory fails to give directions for constructing the empirical model. This is the case of the famous "growth regressions" where researchers try to identify factors affecting growth from a large pool of potential factors; see Fernandez et al. (2001). A major criticism of the g-prior in dynamic models is that X includes data from y (the vector of data of the dependent variable) through lags, so that prior depends on y and Bayes theorem is inapplicable. The ridge prior presented in the next subsection is a shrinkage prior that avoids this criticism.

3.2 Non-conjugate priors

A reason for using a non-conjugate prior is convenience in choosing the parameters of the prior. We have seen that the normal (conjugate) prior for β depends on σ^2 through its covariance matrix that is proportional to σ^2. Thus if we choose for example the prior $\beta \sim \mathrm{N}(0, \sigma^2 \underline{\mathrm{V}}_\beta)$ (assume β is scalar here) and $\sigma^2 \sim IG(\underline{\nu}, \underline{q})$ with ν smaller than two, the unconditional prior variance of β is 'infinite' (i.e. it does not exist). This implies that however small we fix the value of $\underline{V}_\beta$, a choice we would like to make if we have precise information on β, we will be actually noninformative on β. This happens because a value of $\underline{a}$ smaller than two implies that $E(\sigma^2)$ does not exist, and though $E(\beta|\sigma^2) = \sigma^2 \underline{V}_\beta$ exists for any finite value of σ^2, $E(\beta) = E(\sigma^2)\underline{V}_\beta$ does not exist if $E(\sigma^2)$ does not exist. In practice, it is very practical to be noninformative on σ^2 since this is a parameter about which we usually have no prior ideas. A convenient noninformative prior on σ^2 is proportional to $1/\sigma^2$, even if an iG$(0.001, 0.001)$ is practically noninformative as well. To avoid the problem outlined above when we want to be informative about at least one element of β, we recommend therefore to use a prior on β that does not depend on σ^2.

3.2.1 Normal priors

There are many possible choices on non-conjugate priors for β, and we consider first the case where the prior is Normal, say $\beta \sim N(\underline{\beta}, \underline{V}_\beta)$, multiplied by the noninformative prior $1/\sigma^2$ for the variance parameter of the regression model. The price to pay for avoiding conjugacy is that the posterior results are not available analytically and must be computed numerically. However, a Gibbs sampling algorithm is easily constructed to simulate the posterior. It is the same as the algorithm described in Section 3.1 except that the distributions to simulate in steps 3 are given below. Indeed, calculations similar to those of

Section 2.2.1 provide the following conditional posteriors:[11]

$$\begin{aligned}\beta|y,X,\sigma^2 &\sim \mathrm{N}(\overline{\beta}^*,\overline{\mathrm{V}}^*_\beta)\\ \sigma^2|y,X,\beta &\sim \mathrm{iG}(\mathrm{T},(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta))\end{aligned}$$

where

$$\overline{V}^*_\beta = \left(\underline{V}^{-1}_\beta + \sigma^{-2}X'X\right)^{-1}, \quad \overline{\beta^*} = \overline{V}^*_\beta\left(\underline{V}^{-1}_\beta\underline{\beta} + \sigma^{-2}X'X\hat{\beta}\right). \tag{16}$$

Apart from subjective choices of $\underline{\beta}$ and $\underline{V}_\beta$, we review briefly other choices that have been proposed and are less demanding in terms of prior elicitation.

Ridge regression priors A Normal prior of the form

$$p(\beta) \sim \mathrm{N}(0,\tau I_k)$$

where τ is a prior parameter, is called a 'ridge regression' prior. It leads to a posterior mean similar to the estimate obtained from classical ridge regression. In this case, $\beta|\sigma^2$ has a Normal posterior with posterior mean

$$\overline{\beta}^* = \left(\sigma^{-2}X'X + \frac{1}{\tau}I_k\right)^{-1}\sigma^{-2}X'y.$$

As for the g-prior, the case where the prior variance is infinite ($\tau \to \infty$) leads to the OLS estimate as unconditional posterior mean. For $\tau \to 0$ the unconditional posterior mean of β also tends to zero, thus this prior can provide shrinkage over the OLS estimate. Apart from these two limit cases, the unconditional posterior mean must be computed by Gibbs sampling. Ridge regression priors impose prior independence between the coefficients β, since the prior covariance matrix is diagonal, and cannot incorporate prior beliefs of correlations between elements of β.

Empirical Bayes priors The Empirical Bayes technique relies on the information in the observations to estimate the parameters of the prior distribution. In that respect, they are subject to the major criticism that Bayes theorem is not applicable if the prior depends on the data y. Depending on the problem at hand, there are many options for defining an Empirical Bayes prior. For instance, Judge and Bock (1978) suggested the Empirical Bayes prior

$$\beta \sim \mathrm{N}\left(0,\tau\left(X'X\right)^{-1}\right)$$

[11]The normality of $p(\beta|y,X,\sigma^2)$ comes from the fact that its functional form is the product of two functions that are like Normal densities: $\exp[-(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})/(2\sigma^2)]$ (from the likelihood) and $\exp[-(\beta-\underline{\beta})'\underline{V}^{-1}_\beta(\beta-\underline{\beta})/2]$ (from the prior). Actually, the Normal prior used here is conjugate for the likelihood when σ^2 is fixed, though the joint prior of β and σ^2 is not conjugate to the likelihood for both parameters. Thus we have 'partial conjugacy'.

where $\tau = \frac{\widehat{\sigma}^2}{\widehat{\xi}^2}$, $\widehat{\sigma}^2 = \left(y - X\hat{\beta}\right)' \left(y - X\hat{\beta}\right)/T$ and $\widehat{\xi}^2 = \frac{\hat{\beta}'\hat{\beta}}{\mathrm{tr}(\mathrm{X}'\mathrm{X})^{-1}} - \widehat{\sigma}^2$. This empirical Bayes prior is Stein-like, also shrinking β towards zero, since the posterior mean of β given σ^2 writes

$$\overline{\beta}^* = \left(\frac{\tau}{\tau + \sigma^2}\right)\hat{\beta}.$$

Full-Bayes (hierarchical) priors Ridge and g-priors are based on the subjective choice of a 'tuning' prior parameter that provides shrinkage, and hence are difficult to justify among objective researchers. On the other hand, Empirical Bayes priors are less favoured by Bayesians because they do not respect Bayes Theorem, though from an empirical viewpoint, they often prove to be helpful. Since anyway the coefficients β are assumed to be random variables, why not treat also the prior parameters as random variables which admit a prior on their own and can be updated from information in the likelihood (using again Bayes Theorem)? Although this choice may seem abstract conceptually, by definining a *hyperprior* distribution on unknown prior parameters, we can accommodate a variety of shrinkage and model selection estimators.

To give an example, we consider hierarchical shrinkage priors. In the ridge regression prior $\beta \sim \mathrm{N}(0, \tau I_k)$, all k coefficients in β share the same shrinkage factor τ. If we want the different coefficients in β to be shrunk to a different degree, we can use the prior

$$\beta_i|\tau_i \sim \mathrm{N}(0, \tau_i) \tag{17}$$

for $i = 1, 2, ..., k$. Choosing all the different τ_i is very demanding. If instead we assume a common prior on all τ_is, we allow the data to determine their values. To see this assume the conjugate prior for this variance parameter of the form

$$\tau_i \sim \mathrm{iG}(q_1, q_2). \tag{18}$$

Then we can easily derive the conditional posterior densities for this model and use the Gibbs sampler to simulate all parameters:

1. Draw τ_i conditional on β_i from

$$\mathrm{iG}\left(q_1 + 1, q_2 + \beta_i^2\right), \text{for } i = 1, 2, \ldots, k. \tag{19}$$

2. Draw σ^2 conditional on β and the data from $\mathrm{iG}(T, (y - X\beta)'(y - X\beta))$.

3. Draw β conditional on all τ_is, σ^2 and the data from

$$\mathrm{N}\left(\left(\sigma^{-2}X'X + (\underline{V})^{-1}\right)^{-1} X'y, \left(\sigma^{-2}X'X + (\underline{V})^{-1}\right)^{-1}\right) \tag{20}$$

where $\underline{V} = \mathrm{diag}(\tau_1, ..., \tau_k)$ is the prior covariance matrix constructed from the τ_i.

Note that in step 1 the data do not revise τ_i directly but only through β_i, in step 2 the τ_is influence σ^2 only through β, and in step 3 they influence β directly. The data appear directly only in steps 2 and 3.

Numerous such examples exist in the literature. For instance, once can specify a Uniform (noninformative) prior on τ_i, while an Exponential prior on τ_i gives a posterior mean with shrinkage properties identical to the LASSO (least absolute shrinkage and selection operator) algorithm. Other authors have used hierarchical priors for model selection and model averaging. For instance, one can replace the prior in (17) by

$$p\left(\beta_i\right) \sim \mathrm{N}\left(0, \gamma_i \tau_i\right) \tag{21}$$

where τ_i may, or may not, have a prior but the crucial assumption is that γ_i is a 0/1 variable. This prior is a mixture of Normal priors: when $\gamma_i = 0$, we have a $N\left(0, 0\right)$ prior, i.e. a point mass at zero, which by definition will restrict the posterior of β_i to have point mass at zero; when $\gamma_i = 1$, we have a $N\left(0, \tau_i\right)$ prior, i.e. an unrestricted prior (for non-zero values of τ_i) and hence β_i is updated by the likelihood. The Bayesian can allow the information in the likelihood to determine which γ_i will be zero and which will be one, by placing a prior on γ_i. The conjugate prior is of the form

$$\gamma_i \sim \text{Bernoulli}\left(\pi_i\right). \tag{22}$$

It leads to a Gibbs sampler algorithm which gives: i) a posterior estimate of β_i, which is shrunk towards zero if and only if $\gamma_i = 0$, and ii) a posterior estimate of γ_i indicating which coefficients (and hence which predictor variables) should be included in the model.

One can take this hierarchical analysis to a further step and combine algorithms and ideas. For instance, if in (21) we assume $\tau_i \to \infty$ we just let all coefficients with $\gamma_i = 1$ to have a very flat and uninformative prior. However, we can use the prior (18) instead. In that case, if a coefficient is not restricted to be exactly 0 (i.e. if $\gamma_i = 1$), it can still be shrunk towards zero by allowing τ_i to vary according to information in the likelihood. Similarly, if we are unsure about choosing a precise value for the hyperparameter π_i in (22), we can easily introduce one more hierarchical layer and place a prior on this hyperparameter! In this case, the conjugate prior on π_i is in the family of Beta densities, so that the posterior of π_i is also a Beta density, and hence is easy to sample from.

3.2.2 Non-Normal priors

If a researcher wishes to use a non-Normal prior for β, denoted by $p(\beta)$, the conditional posterior of $\beta|y, X, \sigma^2$ is not Normal.[12] We can only say that

$$p(\beta|y, X, \sigma^2) \propto p(\beta) \exp[-(\beta - \hat{\beta})' X' X (\beta - \hat{\beta})/(2\sigma^2)], \tag{23}$$

[12]For example, we way wish to use an asymmetric prior if our prior belief is that a parameter is of the order of 0.95 (prior mean or mode), definitely smaller than one, and with high probability in the interval $(0.5, 1)$. A normal prior centered on 0.9 with a small standard deviation (such as 0.015) implies that the parameter has a negligible probability to be larger than one, but it also implies that the parameter is in the interval $(0.9, 1)$ with probability (very close to) one rather than in the desired interval $(0.5, 1)$. A Beta density for that parameter can be easily chosen to satisfy all the prior beliefs.

and that it will not be possible to simulate it directly. In such a case, the simulation of $\beta|y, X, (\sigma^2)^{(r-1)}$ in step 3 of the Gibbs sampling algorithm of Section 3.1 has to be done using a Metropolis step. To do this, at iteration r, we must approximate (23) (where σ^2 is set at the value $(\sigma^2)^{(r-1)}$ generated at step 3 of the previous iteration) by a density that we can simulate directly, denoted by $q(\beta|(\sigma^2)^{(r-1)})$. The sampling of β at step 3 of the algorithm of Section 3.1 is done like this:

Draw $\beta^{(cand)}$ *from* $q(\beta|(\sigma^2)^{(r-1)})$. *Set* $\beta^{(r)} = \beta^{(cand)}$ *with probability* α *and* $= \beta^{(r-1)}$ *with probability* $1-\alpha$, *where* $\alpha = \min\left\{\frac{p(\beta^{(cand)}|y,X,(\sigma^2)^{(r-1)})}{p(\beta^{(r-1)}|y,X,(\sigma^2)^{(r-1)})}\frac{q(\beta^{(r-1)}|\beta^{(cand)},(\sigma^2)^{(r-1)})}{q(\beta^{(cand)}|\beta^{(r-1)},(\sigma^2)^{(r-1)})}, 1\right\}$ *and* $p(.|.)$ *defined in (23).*

If the prior $p(\beta)$ is not very informative, or if it is not highly non-normal, (23) can be easily approximated by replacing the non-normal prior by a normal prior approximating it. Then the candidate $q(\beta|\sigma^2)$ will be the normal posterior defined in Section 3.2.1 above formula (16) and the Metropolis step is easy to implement. If the prior is highly non-normal and very sharp, one will have to think harder to design a good proposal, i.e. one that does not lead to reject often the candidate draws of β. Too many rejections would slow down (or prevent) the convergence of the algorithm to the targeted posterior distribution.

4 Other models: A short guide to the literature

Several books cover in detail Bayesian inference in econometrics. The most comprehensive ones for applied macroeconomists are probably Bauwens et al. (1999) and Koop (2003). Geweke et al. (2011) has chapters on time series state space models, macroeconometrics and MCMC methods. Geweke (2005) and Lancaster (2004) include each one chapter on time series models.

Bayesian inference on DSGE models is covered in Chapter 22 of this Handbook.

Bayesian inference for VAR models is reviewed in Ch. 9 of Bauwens et al. (1999). A recent treatment of VAR models with shrinkage, time-varying parameters and stochastic volatility, as well as factor augmented VARs can be found in Koop and Korobilis (2010). The authors provide also MATLAB code to estimate the models using analytical or MCMC methods of the form introduced in Section 2 of this Chapter.

Markov Switching and state-space models are covered extensively in the two excellent books by Frühwirth-Schnatter (2006) and Kim and Nelson (1999).

The list of resources related to Bayesian analysis in macroeconomics is not by all means restricted to the books and monographs just presented. However, this referenced material is a good starting point for the inexperienced student or researcher who would want to start producing research using Bayesian methods.

References

Adkins, L. C., D. S. Rickman, and A. Hameed (2003). Bayesian estimation of regional production for CGE modeling. *Journal of Regional Science 43*, 641–661.

Bauwens, L., M. Lubrano, and J. Richard (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press.

Doan, T., L. R.B., and C. Sims (1984). Forecasting and conditional projection under realistic prior distributions. *Econometric Reviews 3*, 1–100.

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

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Wiley.

Geweke, J., G. Koop, and H. van Dijk (2011). *The Oxford Handbook of Bayesian Econometrics*. Oxford University Press.

Ingram, B. F. and C. H. Whiteman (1994). Supplanting the Minnesota prior: Forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics 34*, 497–510.

Judge, G. G. and M. Bock (1978). *Statistical Implications of Pre-Test and Stein Rule Estimators in Econometrics*. North-Holland.

Kim, C.-J. and C. R. Nelson (1999). *State-Space Models with Regime Switching*. MIT Press.

Koop, G. (2003). *Bayesian Econometrics*. Wiley.

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

Lancaster, T. (2004). *Introduction to Modern Bayesian Econometrics*. Blackwell.

Uhlig, H. (2001). Did the Fed surprise the markets in 2001? A case study for VARs with sign restrictions. Technical report, CESifo Working Paper Series.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In P. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques*. North-Holland.

Recent titles

CORE Discussion Papers

2011/17. Juan A. MAÑEZ, Rafael MONER-COLONQUES, José J. SEMPERE-MONERRIS and Amparo URBANO Price differentials among brands in retail distribution: product quality and service quality.
2011/18. Pierre M. PICARD and Bruno VAN POTTELSBERGHE DE LA POTTERIE. Patent office governance and patent system quality.
2011/19. Emmanuelle AURIOL and Pierre M. PICARD. A theory of BOT concession contracts.
2011/20. Fred SCHROYEN. Attitudes towards income risk in the presence of quantity constraints.
2011/21. Dimitris KOROBILIS. Hierarchical shrinkage priors for dynamic regressions with many predictors.
2011/22. Dimitris KOROBILIS. VAR forecasting using Bayesian variable selection.
2011/23. Marc FLEURBAEY and Stéphane ZUBER. Inequality aversion and separability in social risk evaluation.
2011/24. Helmuth CREMER and Pierre PESTIEAU. Social long term care insurance and redistribution.
2011/25. Natali HRITONENKO and Yuri YATSENKO. Sustainable growth and modernization under environmental hazard and adaptation.
2011/26. Marc FLEURBAEY and Erik SCHOKKAERT. Equity in health and health care.
2011/27. David DE LA CROIX and Axel GOSSERIES. The natalist bias of pollution control.
2011/28. Olivier DURAND-LASSERVE, Axel PIERRU and Yves SMEERS. Effects of the uncertainty about global economic recovery on energy transition and CO_2 price.
2011/29. Ana MAULEON, Elena MOLIS, Vincent J. VANNETELBOSCH and Wouter VERGOTE. Absolutely stable roommate problems.
2011/30. Nicolas GILLIS and François GLINEUR. Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization.
2011/31. Nguyen Thang DAO and Julio DAVILA. Implementing steady state efficiency in overlapping generations economies with environmental externalities.
2011/32. Paul BELLEFLAMME, Thomas LAMBERT and Armin SCHWIENBACHER. Crowdfunding: tapping the right crowd.
2011/33. Pierre PESTIEAU and Gregory PONTHIERE. Optimal fertility along the lifecycle.
2011/34. Joachim GAHUNGU and Yves SMEERS. Optimal time to invest when the price processes are geometric Brownian motions. A tentative based on smooth fit.
2011/35. Joachim GAHUNGU and Yves SMEERS. Sufficient and necessary conditions for perpetual multi-assets exchange options.
2011/36. Miguel A.G. BELMONTE, Gary KOOP and Dimitris KOROBILIS. Hierarchical shrinkage in time-varying parameter models.
2011/37. Quentin BOTTON, Bernard FORTZ, Luis GOUVEIA and Michael POSS. Benders decomposition for the hop-constrained survivable network design problem.
2011/38. J. Peter NEARY and Joe THARAKAN. International trade with endogenous mode of competition in general equilibrium.
2011/39. Jean-François CAULIER, Ana MAULEON, Jose J. SEMPERE-MONERRIS and Vincent VANNETELBOSCH. Stable and efficient coalitional networks.
2011/40. Pierre M. PICARD and Tim WORRALL. Sustainable migration policies.
2011/41. Sébastien VAN BELLEGEM. Locally stationary volatility modelling.
2011/42. Dimitri PAOLINI, Pasquale PISTONE, Giuseppe PULINA and Martin ZAGLER. Tax treaties and the allocation of taxing rights with developing countries.
2011/43. Marc FLEURBAEY and Erik SCHOKKAERT. Behavioral fair social choice.
2011/44. Joachim GAHUNGU and Yves SMEERS. A real options model for electricity capacity expansion.
2011/45. Marie-Louise LEROUX and Pierre PESTIEAU. Social security and family support.
2011/46. Chiara CANTA. Efficiency, access and the mixed delivery of health care services.

Recent titles

CORE Discussion Papers - continued

2011/47. Jean J. GABSZEWICZ, Salome GVETADZE and Skerdilajda ZANAJ. Migrations, public goods and taxes.
2011/48. Jean J. GABSZEWICZ and Joana RESENDE. Credence goods and product differentiation.
2011/49. Jean J. GABSZEWICZ, Tanguy VAN YPERSELE and Skerdilajda ZANAJ. Does the seller of a house facing a large number of buyers always decrease its price when its first offer is rejected?
2011/50. Mathieu VAN VYVE. Linear prices for non-convex electricity markets: models and algorithms.
2011/51. Parkash CHANDER and Henry TULKENS. The Kyoto *Protocol*, the Copenhagen *Accord*, the Cancun *Agreements*, and beyond: An economic and game theoretical exploration and interpretation.
2011/52. Fabian Y.R.P. BOCART and Christian HAFNER. Econometric analysis of volatile art markets.
2011/53. Philippe DE DONDER and Pierre PESTIEAU. Private, social and self insurance for long-term care: a political economy analysis.
2011/54. Filippo L. CALCIANO. Oligopolistic competition with general complementarities.
2011/55. Luc BAUWENS, Arnaud DUFAYS and Bruno DE BACKER. Estimating and forecasting structural breaks in financial time series.
2011/56. Pau OLIVELLA and Fred SCHROYEN. Multidimensional screening in a monopolistic insurance market.
2011/57. Knud J. MUNK. Optimal taxation in the presence of a congested public good and an application to transport policy.
2011/58. Luc BAUWENS, Christian HAFNER and Sébastien LAURENT. Volatility models.
2011/59. Pierre PESTIEAU and Grégory PONTHIERE. Childbearing age, family allowances and social security.
2011/60. Julio DÁVILA. Optimal population and education.
2011/61. Luc BAUWENS and Dimitris KOROBILIS. Bayesian methods.

Books

J. HINDRIKS (ed.) (2008), *Au-delà de Copernic: de la confusion au consensus* ? Brussels, Academic and Scientific Publishers.
J-M. HURIOT and J-F. THISSE (eds) (2009), *Economics of cities.* Cambridge, Cambridge University Press.
P. BELLEFLAMME and M. PEITZ (eds) (2010), *Industrial organization: markets and strategies*. Cambridge University Press.
M. JUNGER, Th. LIEBLING, D. NADDEF, G. NEMHAUSER, W. PULLEYBLANK, G. REINELT, G. RINALDI and L. WOLSEY (eds) (2010), *50 years of integer programming, 1958-2008: from the early years to the state-of-the-art*. Berlin Springer.
G. DURANTON, Ph. MARTIN, Th. MAYER and F. MAYNERIS (eds) (2010), *The economics of clusters – Lessons from the French experience*. Oxford University Press.
J. HINDRIKS and I. VAN DE CLOOT (eds) (2011), *Notre pension en heritage*. Itinera Institute.
M. FLEURBAEY and F. MANIQUET (eds) (2011), *A theory of fairness and social welfare*. Cambridge University Press.
V. GINSBURGH and S. WEBER (eds) (2011), *How many languages make sense? The economics of linguistic diversity*. Princeton University Press.

CORE Lecture Series

D. BIENSTOCK (2001), Potential function methods for approximately solving linear programming problems: theory and practice.
R. AMIR (2002), Supermodularity and complementarity in economics.
R. WEISMANTEL (2006), Lectures on mixed nonlinear programming.