Simulation Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances

Lennart F. Hoogerheide[*], Herman K. van Dijk[†] & Rutger D. van Oest[‡]

January 2007

Econometric Institute report EI 2007-03

Abstract

In this paper we discuss several aspects of simulation based Bayesian econometric inference. We start at an elementary level on basic concepts of Bayesian analysis; evaluating integrals by simulation methods is a crucial ingredient in Bayesian inference. Next, the most popular and well-known simulation techniques are discussed, the Metropolis-Hastings algorithm and Gibbs sampling (being the most popular Markov chain Monte Carlo methods) and importance sampling. After that, we discuss two recently developed sampling methods: adaptive radial based direction sampling [ARDS], which makes use of a transformation to radial coordinates, and neural network sampling, which makes use of a neural network approximation to the posterior distribution of interest. Both methods are especially useful in cases where the posterior distribution is not well-behaved, in the sense of having highly non-elliptical shapes. The simulation techniques are illustrated in several example models, such as a model for the real US GNP and models for binary data of a US recession indicator.

[*]Center for Operations Research and Econometrics, Université catholique de Louvain, Belgium

[†]Econometric and Tinbergen Institutes, Erasmus Univerity Rotterdam, The Netherlands

[‡]CentER, Faculty of Economics and Business Administration, Tilburg University, The Netherlands

1 Introduction

In this paper we discuss several aspects of simulation based Bayesian econometric inference [SBBEI]. In recent decades there has been a huge increase in the use of simulation methods for the Bayesian analysis of econometric models. This 'Simulation Revolution' in Bayesian econometric inference is to a large extent due to the advent of computers with ever-increasing computational power; see e.g. the discussion in Geweke (1999), Van Dijk (1999) and Hamilton (2006). This computational power allows researchers to apply elaborate Bayesian simulation techniques for estimation in which extensive use is made of pseudo-random numbers generated on computers.

The basic principle in this line of research is that in most cases of empirical econometric models one can not directly simulate from the distribution of interest. Thus one applies in such cases an *indirect* sampling method. Two classes of indirect simulation methods are Importance Sampling and Markov chain Monte Carlo. The theory of Markov chain Monte Carlo [MCMC] methods starts with Metropolis et al. (1953) and Hastings (1970). The Gibbs sampling method, the most well-known MCMC method, is due to Geman and Geman (1984). Importance sampling, due to Hammersley and Handscomb (1964), was introduced in econometrics and statistics by Kloek and Van Dijk (1978), and further developed by Van Dijk and Kloek (1980, 1984) and Geweke (1989).

The Gibbs sampler has, in particular, become a popular tool in econometrics for analyzing a wide variety of problems; see Chib and Greenberg (1995) and Geweke (1999). Judging from numerous articles in recent literature, Gibbs sampling is still gaining more and more momentum. Recent textbooks such as Bauwens, Lubrano and Richard (1999), Koop (2003), Lancaster (2004), and Geweke (2005) discuss how Gibbs sampling is used in a wide range of econometric models, in particular in models with latent variables.

Evaluating integrals is a crucial ingredient in the Bayesian analysis of any model. The reason is that the basic principle, Bayes' rule, provides (a kernel of) the joint posterior density of all parameters occurring in the model. One is typically interested in the posterior means and standard deviations of some of the parameters; the posterior probability that a parameter lies in a certain interval; and/or the marginal likelihood of the model. For these purposes - and, of course, for prediction and decision analysis - one has to integrate the joint posterior density kernel with respect to all parameters. Therefore, the development of advanced sampling methods, that perform this integration operation efficiently, makes Bayesian inference possible in a wider class of complex models. This allows for more realistic descriptions of processes in many situations, for example in finance and macro-economics, leading to more accurate forecasts and a better quantification of uncertainty.

In order to make this paper self contained we start with a discussion of basic principles of Bayesian inference such as prior & posterior density, Bayes' rule, Highest Posterior Density [HPD] region, Bayes factor, and posterior odds. Good knowl-

edge of these principles is necessary for understanding the application of simulation methods in Bayesian econometric inference. After the introduction to Bayesian inference we proceed and discuss basic ideas of simulation methods. These methods are applicable to posterior densities that are reasonably well-behaved. Recent work in SBBEI deals with cases where the posterior is not well-behaved. We also discuss some methods that can be used in such a situation. Highly non-elliptical shapes in posterior distributions typically arise when some parameters have a substantial amount of posterior probability near or at the boundary of the parameter region. This feature may occur and is relevant in several econometric models. A practical example is a dynamic economic process that is possibly non-stationary. Other examples are the presence of very weak instruments in an instrumental variable regression model, and models with multiple regimes in which one regime may have neglectable probability.

The contents of this paper is structured as follows. In Section 2 we briefly review the basic principles of Bayesian inference. In Section 3 we first discuss several well-known simulation techniques such as Importance Sampling, the Metropolis-Hastings algorithm and the Gibbs sampler. Next, we discuss two recently developed simulation methods: adaptive radial based direction sampling [ARDS], which makes use of a transformation to radial coordinates, and neural network sampling, which makes use of a neural network approximation to the posterior distribution of interest. The final section provides some concluding remarks.

2 A Primer on Bayesian Inference

2.1 Motivation for Bayesian Inference

The dissatisfaction that many applied economic researchers feel when they consider the 'significance' of regression coefficients, using the frequentist/classical approach, is one major motivation to start with Bayesian inference. Consider the following example.

Example: growth of real GNP in the US

Throughout this paper we use the (annualized) quarterly growth rate of the real Gross National Product (GNP) in the United States several times for illustrative purposes. The data are shown in Figure 1. Consider the ordinary least squares (OLS) regression for $T = 126$ observations y_t from 1975 to the second quarter of 2006 (with t-values in parentheses):

$$\begin{array}{rccccccc} y_t = & 1.99 & + & 0.22\ y_{t-1} + & 0.13\ y_{t-2} + \hat{u}_t & & (t = 1, \ldots, T) \\ & (4.80) & & (2.57) & (1.50) & & \end{array}$$

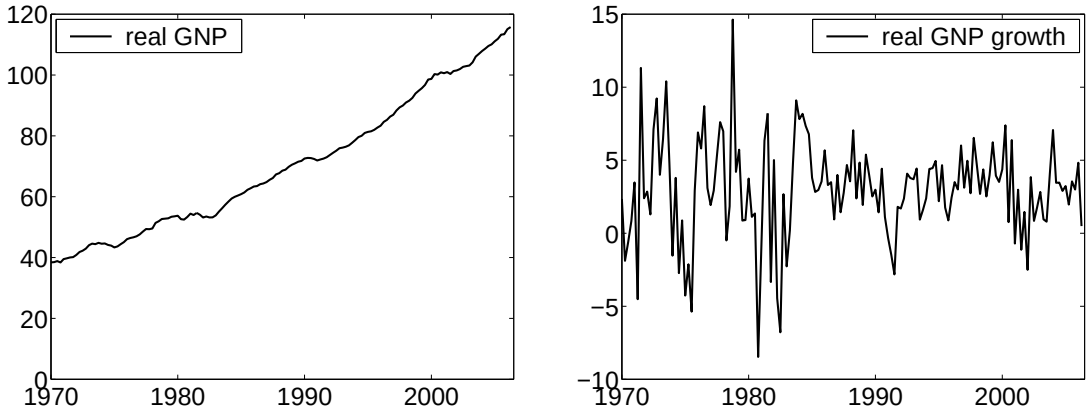

Figure 1: U.S. real Gross National Product - quantity index, 2000=100 (left), and corresponding (annualized) growth rates in percents (right). The data are seasonally adjusted. Source: U.S. Department of Commerce, Bureau of Economic Analysis.

where $\hat{u}_t$ are OLS residuals. Now suppose one fixes the coefficient of y_{t-2} at zero; then one obtains:

$$\begin{array}{rccc} y_t = & 2.26 & + & 0.27\ y_{t-1} + \hat{v}_t \qquad (t = 1, \ldots, T) \\ & (6.03) & & (3.13) \end{array}$$

where $\hat{v}_t$ are the OLS residuals. A naive researcher might conclude that in the second model the influence of y_{t-1} on y_t is "much more significant". However, according to a proper interpretation of the frequentist/classical approach, this is not a meaningful statement. The reason for this is that in classical inference only the falsification of the null hypothesis is possible. Otherwise stated, it is only relevant whether or not the null hypothesis is rejected.

Another point is that the concept of 'unbiasedness' of an estimator is not meaningful in non-experimental sciences: an unbiased estimator takes on average the correct value when the process is repeated (infinitely) many times. However, in non-experimental sciences this idea of repeating the process is not realistic. In non-experimental sciences, a researcher cannot repeat the process he/she studies, and he/she has to deal with only one given data set.

A proper way to consider the sensitivity of estimates and to use probability statements that indicate a 'degree of confidence' is given by the framework of Bayesian inference. So, apart from dissatisfaction with existing practice of the frequentist/classical approach, there also exists a constructive motive to apply Bayesian inference. That is, a second major motivation to start with Bayesian inference is

that the Bayesian framework provides a natural learning rule, that allows for optimal learning and (hence) optimal decision making under uncertainty.

In this section the basic principle of Bayesian inference, Bayes' theorem, will first be discussed. After that, some concepts that play an important role within the Bayesian framework will be described, and a comparison will be made between Bayesian inference and the frequentist/classical approach.

2.2 Bayes' theorem as a learning device

Econometric models may be described by the joint probability distribution of $y = \{y_1, \ldots, y_N\}$, the set of N available observations on the endogenous variable y_i, where y_i may be a vector itself, that is known upto a parameter vector θ. Bayesian inference proceeds from the likelihood function $L(\theta) = p(y|\theta)$, which is either the density of the data given the parameters in case of a continuous distribution or the probability function in case of a discrete distribution, and a prior density $p(\theta)$ reflecting prior beliefs on the parameters before the data set has been observed. So, in the Bayesian approach the parameters θ are considered as random variables of which the prior density $p(\theta)$ is updated by the information contained in the data, incorporated in the likelihood function $L(\theta) = p(y|\theta)$, to obtain the posterior density of the parameters $p(\theta|y)$. This process is formalized by Bayes' theorem:

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

Note that this is merely a result of rewriting the identity $p(y)p(\theta|y) = p(\theta)p(y|\theta)$, the two ways of decomposing the joint density $p(y, \theta)$ into a marginal and a conditional density, see Figure 2 for a graphical interpretation of Bayes' theorem. The marginal likelihood $p(y) = \int p(\theta, y)d\theta = \int p(y|\theta)p(\theta)d\theta$ is the marginal density of the data y, after the parameters θ of the model have been integrated out with respect to their prior distribution. The marginal likelihood can be used for model selection, see subsection 2.3.

Formula (1) can be rewritten as:

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{2}$$

where the symbol $\propto$ means "is proportional to", i.e. the left-hand side is equal to the right-hand side times a scaling constant $(1/p(y) = 1/\int p(\theta)p(y|\theta)d\theta)$ that does not depend on the parameters θ; just like the integrating constant $\sqrt{2\pi}$ in the standard normal density.

The basic idea behind the Bayesian approach is that the prior density $p(\theta)$ and the posterior density $p(y|\theta)$ are *subjective* evaluations of possible states of nature and/or outcomes of some process (or action). A famous quote of De Finetti (1974) is: "probabilities do not exist", that is, probabilities are not physical quantities that one can measure in practice, but they are states of the mind.

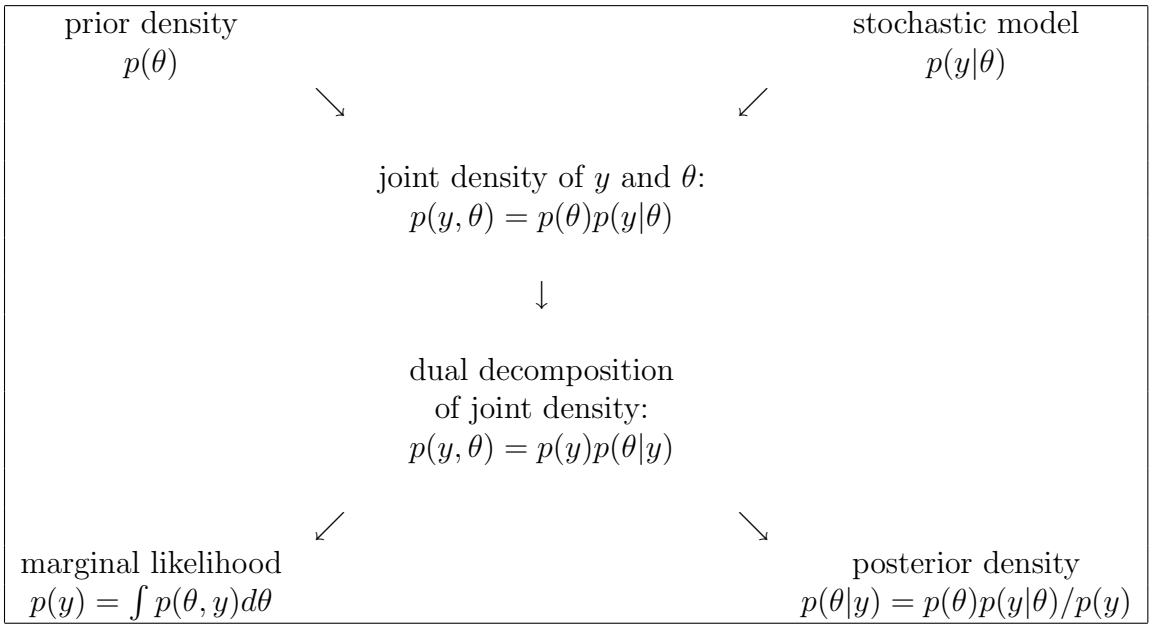

Figure 2: Bayes' theorem as a learning device

Bayes' rule can be interpreted as follows. One starts with the prior density $p(\theta)$; this contains intuitive, theoretical or other ideas on θ, that may stem from earlier or parallel studies. Then one learns from data through the likelihood function $p(y|\theta)$. This yields the posterior $p(\theta|y)$. Briefly stated, Bayes' paradigm is a *learning principle*, which can be depicted as follows:

$$\begin{array}{ccccccc} \text{posterior density} & & \propto & \text{prior density} & \times & \text{likelihood} \\ \\ \begin{array}{c}\text{beliefs after}\\ \text{having observed data}\end{array} & & \Leftarrow & \begin{array}{c}\text{beliefs before}\\ \text{observing data}\end{array} & \& & \begin{array}{c}\text{influence}\\ \text{of the data}\end{array} \end{array}$$

Note that we can apply Bayes' rule sequentially: when new data will become available, we can treat the posterior density that is based on the current data set as the prior density.

The key problems in Bayesian inference are the determination of the probability laws in the posterior kernel, i.e. what families of posterior densities are defined, and the computation of the marginal likelihood.

Example: growth of real GNP in the US (continued)

In order to illustrate Bayes' theorem, we consider the quarterly data on U.S. real GNP. Figure 1 displays real GNP and the corresponding growth rate in percents. Here we consider the naive model

$$y_t = \theta + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, 25) \quad \text{i.i.d.}, \qquad t = 1, \ldots, T, \tag{3}$$

where y_t is the (annualized) growth rate in period t and θ is the average growth rate. So, growth rates are assumed to obey a normal distribution with known standard deviation 5. Clearly, the likelihood function is given by

$$p(y|\theta) \propto \exp\left(-\frac{\sum_{t=1}^{T}(y_t - \theta)^2}{2 \cdot 25} \right), \tag{4}$$

where we have omitted the scaling constant of the normal density, as it is irrelevant in the analysis. Next, a prior density has to be specified for θ. Suppose that it is a priori expected that average real GNP growth is approximately 4 (percent), and that one believes that there is a 95% probability that average real GNP growth lies between 0 and 8 (percent). Such prior beliefs can be captured by a normal distribution with mean 4 and standard deviation 2 (percent), so that the prior density is given by

$$p(\theta) \propto \exp\left(-\frac{(\theta - 4)^2}{2 \cdot 4} \right). \tag{5}$$

Applying Bayes' theorem (2) to formulas (4) and (5) results in a posterior

$$\begin{aligned}
p(\theta|y) &\propto \exp\left(-\frac{\theta^2 - 8\theta + 16}{2 \cdot 4} \right) \exp\left(-\frac{\sum_{t=1}^{T}(\theta - y_t)^2}{2 \cdot 25} \right) \\
&\propto \exp\left(-\frac{\theta^2 - 8\theta}{2 \cdot 4} \right) \exp\left(-\frac{T\theta^2 - 2\theta \sum_{t=1}^{T} y_t}{2 \cdot 25} \right) \\
&= \exp\left(-\frac{1}{2} \left\{ \left[\frac{T}{25} + \frac{1}{4} \right] \theta^2 - 2 \left[\frac{\sum_{t=1}^{T} y_t}{25} + 1 \right] \theta \right\} \right) \\
&\propto \exp\left(-\frac{1}{2} \left[\frac{T}{25} + \frac{1}{4} \right] \left\{ \theta - \frac{\sum_{t=1}^{T} y_t/25 + 1}{T/25 + 1/4} \right\}^2 \right) \\
&= \exp\left(-\frac{1}{2} \left[\frac{T}{25} + \frac{1}{4} \right] \left\{ \theta - \frac{\sum_{t=1}^{T} y_t + 25}{T + 25/4} \right\}^2 \right)
\end{aligned} \tag{6}$$

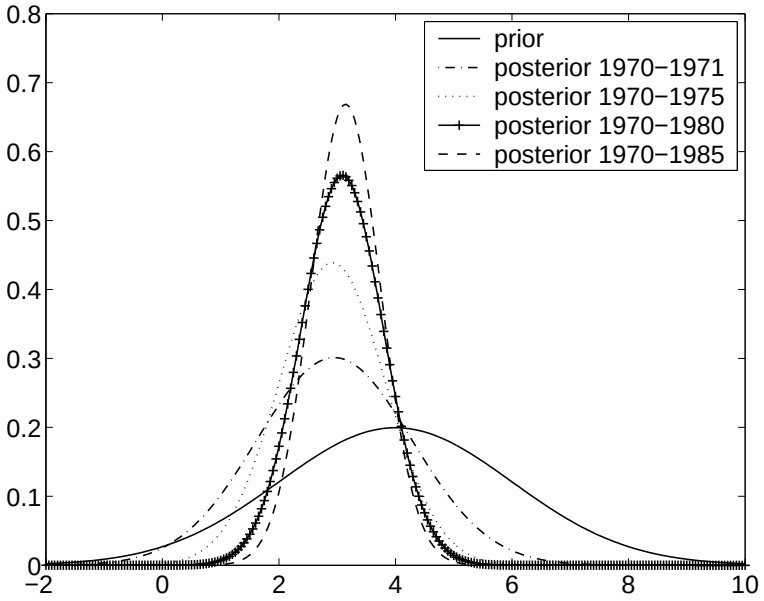

Figure 3: Illustration of Bayesian learning: average U.S. real GNP growth.

which is a kernel (= proportionality function) of a normal density with mean $\frac{\sum_{t=1}^{T} y_t + 25}{T + 25/4}$ and variance $\left(\frac{T}{25} + \frac{1}{4}\right)^{-1}$. So,

$$\theta | y \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T} y_t + 25}{T + 25/4}, \left(\frac{T}{25} + \frac{1}{4} \right)^{-1} \right). \tag{7}$$

Note that for $T \to \infty$ the posterior mean of θ approaches the sample mean $\sum_{t=1}^{T} y_t / T$ and the posterior variance goes to 0, whereas filling in $T = 0$ (and $\sum_{t=1}^{T} y_t = 0$) yields the prior distribution.

Figure 3 provides a graphical illustration of Bayesian learning. It shows how the distribution of the real GNP growth parameter θ changes when more observations become available. In the graph, the posterior distributions are obtained from (7), where the considered observations run from 1970 to 1971, 1975, 1980 and 1985, respectively. For instance, the first posterior density includes the years 1970 and 1971, that is, 8 quarterly observations. All the posterior distributions are located to the left of the prior, suggesting that the prior belief of 4 percent growth overestimates the actual growth rate. It is further seen that parameter uncertainty is reduced when more observations are used.

Conjugate priors

The example above demonstrates that a normal prior applied to a normal data generating process results in a normal posterior. This phenomenon that the posterior density has the same form as the prior density is called "conjugacy". Conjugate priors are useful, as they greatly simplify Bayesian analysis. There exist several forms of conjugacy. Without the intention to be exhaustive, we mention that a Beta prior results in a Beta posterior for a binomial data process, and that a gamma prior results in a gamma posterior for a Poisson data process. Although using conjugacy facilitates Bayesian analysis, a possible critical remark is that conjugate priors are often more driven by convenience than by realism.

General case of the normal model with known variance

The example above can be generalized. Suppose that the data $y = (y_1, \ldots, y_T)$ are generated from a normal distribution $\mathcal{N}(\theta, \sigma^2)$ where the variance σ^2 is known, and the prior distribution for the parameter θ is $\mathcal{N}(\theta_0, \sigma_0^2)$. So, we consider the same model as before, but now we do not fill in specific values for the process variance σ^2 and the prior parameters θ_0 and σ_0^2. In a similar fashion as before, it can be shown that for this more general case

$$\theta|y \sim \mathcal{N}\left(\frac{\theta_0\sigma^2 + \sigma_0^2\sum_{t=1}^{T} y_t}{\sigma^2 + T\sigma_0^2}\,,\left(\frac{1}{\sigma_0^2} + \frac{T}{\sigma^2}\right)^{-1}\right). \tag{8}$$

Interestingly, both the posterior expectation and the (inverse of the) posterior variance in (8) can be decomposed into a prior component and a sample component. By defining the sample mean $\hat{\theta} = \frac{1}{T}\sum_{t=1}^{T} y_t$ and its variance $\sigma_{\hat{\theta}}^2 = \frac{\sigma^2}{T}$, (8) can be written as

$$\theta|y \sim \mathcal{N}\left(\frac{\sigma_0^{-2}}{\sigma_0^{-2} + \sigma_{\hat{\theta}}^{-2}}\,\theta_0 + \frac{\sigma_{\hat{\theta}}^{-2}}{\sigma_0^{-2} + \sigma_{\hat{\theta}}^{-2}}\,\hat{\theta},\left(\sigma_0^{-2} + \sigma_{\hat{\theta}}^{-2}\right)^{-1}\right). \tag{9}$$

In order to interpret (9), we note that the inverted variances σ_0^{-2} and $\sigma_{\hat{\theta}}^{-2}$ essentially measure the informativeness of prior beliefs and available data, respectively. For instance, if σ_0^{-2} is much smaller than $\sigma_{\hat{\theta}}^{-2}$, then the prior density is flat relative to the likelihood function, so that the shape of the posterior density is mainly determined by the data. It is seen from (9) that the posterior expectation of θ is a weighted average of the prior expectation θ_0 and the sample mean $\hat{\theta}$; the weights reflect the amount of prior information relative to the available sample information.

A practical problem, which we have ignored in the analysis so far, is that prior beliefs are often difficult to specify and extremely subjective. So, it might happen that

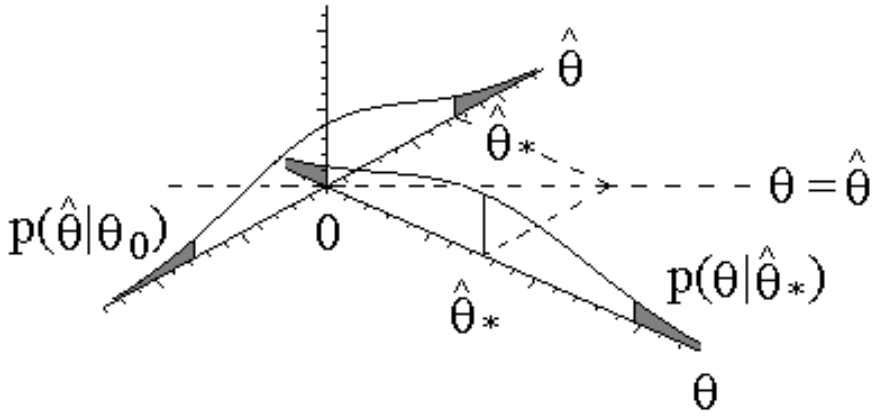

Figure 4: Illustration of symmetry of Bayesian inference and frequentist approach in linear regression model and difference between these approaches

researchers strongly disagree on which prior density is appropriate for the inference problem. As prior beliefs directly affect the posterior results, different researchers may arrive at different conclusions. In order to reach some form of consensus, *non-informative* priors are therefore frequently considered. Such priors are constructed in such a way that they contain little information relative to the information coming from the data. In the (generalized) example above, a "non-informative" prior can be obtained by making the prior distribution $\mathcal{N}(\theta_0, \sigma_0^2)$ diffuse, that is, by letting the prior variance σ_0^2 go to infinity. This essentially amounts to choosing a uniform prior $p(\theta) \propto 1$, reflecting no a priori preference for specific θ values. This implies that the posterior becomes proportional to the likelihood function. It immediately follows from (9) that an infinitely large prior variance results in

$$\theta|y \sim \mathcal{N}(\hat{\theta}, \sigma_{\hat{\theta}}^2), \tag{10}$$

which shows a nice symmetry with classical maximum likelihood [ML], as the ML estimator $\hat{\theta}$ is $\mathcal{N}(\theta, \sigma_{\hat{\theta}}^2)$ distributed. Note that to do classical inference some "true" value has to be assumed for the unknown parameter θ, as otherwise the distribution $\mathcal{N}(\theta, \sigma_{\hat{\theta}}^2)$ would contain unknown elements. *Classical analysis is conditioned on postulated "true" parameter values, whereas Bayesian analysis is conditioned on the data.* This is an important difference between the two approaches. Figure 4 illustrates the difference between Bayesian inference and the frequentist/classical approach in the linear regression model. For example, a Bayesian may investigate whether zero is a likely value for θ given the data, whereas a frequentist may analyze whether the data (summarized in the ML estimator $\hat{\theta}$) are likely under the hypothesis that the true value $\theta_0 = 0$.

Example: illustration of Bayes' rule in TV show game (Monty Hall problem)

We now illustrate the application of Bayes' rule in a simple example of a game that was played in the American TV show 'Let's make a deal'. It is known as the Monty Hall problem, after the show's host. In this TV show game, one could win a car by choosing among three doors the door behind which a car was parked. The candidate was faced with three closed doors: behind one door there was a car, behind the other two doors there was nothing.[1] The procedure of the game was as follows. First, the candidate chose one door, say door 1. Second, the TV show host - who knew behind which door the car could be found - opened one of the other two doors *with no car behind it*. So, if the car was behind door 2, the show's host opened door 3, and vice versa. If the car was behind door 1, the host would open either door 2 or 3 with probability 1/2. Suppose the presenter opened door 3. Finally, the candidate got the chance to switch his/her choice from his/her initial choice (door 1) to the other closed door (door 2). Throughout the episodes of the TV show there were many candidates who chose to stick with their initially chosen door. The question is now whether this was a wise decision; or stated otherwise, was this rationally an optimal decision? To answer this question, we will make use of Bayes' rule.

In order to be able to apply Bayes' rule in this example, we must formulate the TV show game as a model. In this model there is one parameter θ reflecting the door with the car behind it, $\theta \in \{1, 2, 3\}$. The data y are given by the door that the host opens, $y \in \{2, 3\}$. We assume that there is no prior preference for one of the three doors: $\mathsf{Pr}[\theta = i] = 1/3$ for $i = 1, 2, 3$. In the case in which the host opens the third door, the likelihood is given by: $\mathsf{Pr}[y = 3|\, \theta = 1] = 1/2$, $\mathsf{Pr}[y = 3|\, \theta = 2] = 1$, $\mathsf{Pr}[y = 3|\, \theta = 3] = 0$.

From the prior and the likelihood we can now obtain the posterior probability distribution of θ using Bayes' rule. First we obtain the marginal likelihood:[2]

$$\mathsf{Pr}[y = 3] = \sum_{i=1}^{3} \mathsf{Pr}[y = 3|\, \theta = i]\, \mathsf{Pr}[\theta = i] = \frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

Now we obtain the posterior probabilities:

$$\mathsf{Pr}[\theta = 1|\, y = 3] \;\; = \;\; \frac{\mathsf{Pr}[y = 3|\, \theta = 1]\, \mathsf{Pr}[\theta = 1]}{\mathsf{Pr}[y = 3]} = \frac{1/2 \cdot 1/3}{1/2} = \frac{1}{3}$$

[1] The Monty Hall problem is also described as the situation with a car behind one door and a goat behind the other two doors. Obviously, this does not intrinsically change the situation: the point is that behind one door there is something that is worth considerably more money than what is behind the other two doors.

[2] Note that we have a summation over the domain of θ here (instead of an integral), because this is a (quite rare) case in which the parameter θ has a discrete distribution.

$$\begin{aligned}
\Pr[\theta = 2 | \, y = 3] &= \frac{\Pr[y = 3 | \, \theta = 2] \Pr[\theta = 2]}{\Pr[y = 3]} = \frac{1 \cdot 1/3}{1/2} = \frac{2}{3} \\
\Pr[\theta = 3 | \, y = 3] &= \frac{\Pr[y = 3 | \, \theta = 3] \Pr[\theta = 3]}{\Pr[y = 3]} = \frac{0 \cdot 1/3}{1/2} = 0
\end{aligned}$$

We conclude that it would actually be the best rational decision to switch to door 2, having a (posterior) probability of $2/3$, whereas door 1 merely has a (posterior) probability of $1/3$. The problem is also called the Monty Hall *paradox*, as the solution may be counterintuitive: it may appear as if the *seemingly* equivalent doors 1 and 2 should have equal probability of $1/2$. However, the following reasoning explains intuitively why the probability that the car is behind door 1 is merely $1/3$, after door 3 has been opened. At the beginning the probability that the car is behind door 1 was $1/3$, and the fact that door 3 is opened does not change this: it is already known in advance that the host will open one of the *other* doors *with no car behind it*. In other words, the data do not affect the probability that the car is behind door 1. So after door 3 has been opened, door 1 still has $1/3$ probability, while door 2 now has the $2/3$ probability that doors 2 and 3 together had before door 3 had been opened.

It is interesting to see which decision would result from the maximum likelihood approach in this case. Here the ML approach would yield the same decision: $\hat{\theta}_{ML} = 2$. The likelihood $\Pr[y = 3 | \, \theta]$ is highest for $\theta = 2$: $\Pr[y = 3 | \, \theta = 2] = 1$. However, it should be noted that the ML approach does not immediately indicate what the probability is that the car is behind door 2; it does not immediately reveal the uncertainty about the decision. Moreover, if one would have the prior information that in 3 out of 5 TV shows the car is behind door 1, and in 1 out of 5 shows behind door 2 or door 3, then the Bayesian approach would yield a different choice than the ML approach. Then $\hat{\theta}_{ML}$ would still be $\hat{\theta}_{ML} = 2$, whereas the posterior probabilities would then be $\Pr[\theta = 1 | \, y = 3] = 3/5$ versus $\Pr[\theta = 2 | \, y = 3] = 2/5$. This illustrates how Bayes' rule provides us with a natural method to include prior information that is relevant for optimal decision making, and to assess the uncertainty about this decision.

2.3 Model evaluation and model selection

In this section, we discuss two Bayesian testing approaches for model selection. The first is based on the highest posterior density [HPD] region, which is the Bayesian counterpart of the classical confidence interval. The second is posterior odds analysis, comparing the probabilities of multiple considered models given the available data. An important difference between the two approaches is that tests using the

HPD region are based on finding evidence *against* the null model, whereas posterior odds analysis considers the evidence *in favor of* each of the models under scrutiny. So, the HPD approach treats models in an asymmetrical way, just like frequentist/classical testing procedures. The posterior odds approach treats models symmetrically.

2.3.1 The HPD region

The highest posterior density [HPD] region is defined such that any parameter point inside that region has a higher posterior density than any parameter point outside. Consequently, the usually considered 95% HPD region is the smallest region containing 95% of the posterior probability mass. We note that a HPD region does not necessarily consist of a single interval. For example, it might consist of two intervals if the posterior density is bimodal.

Figure 5 shows the 95% HPD region for the average real GNP growth rate θ in the normal model with known variance. The standard normal distribution has 2.5% probability mass both to the right of 1.96 and to the left of -1.96, so that the 95% HPD region for θ is $(2.92 - 1.96 \cdot 0.91,\ 2.92 + 1.96 \cdot 0.91) = (1.14, 4.70)$. It is seen from Figure 5 that a real GNP model imposing zero average growth is rejected, as $\theta = 0$ is located outside the HPD region.

Although the Bayesian HPD region has similarities with the classical confidence interval, the interpretations are very different. In the classical framework, the confidence interval (constructed from the data) is considered random and the postulated parameter value is given, so that one effectively tests whether the data are plausible for the assumed parameter value. On the other hand, a Bayesian considers the HPD region as given and the parameter outcome as random, so that it is effectively tested whether the parameter outcome is plausible given the available data.

2.3.2 Posterior odds analysis

A HPD region based test considers the amount of evidence against the null model, but it does not say anything about the amount of evidence in favor of the alternative model relative to the null model. So, the null model and the alternative model are treated asymmetrically. A testing approach in which models are directly compared is posterior odds analysis. Its formalization for two possibly *non-nested* competing models M_1 and M_2 is as follows. Given the available data y, the model probabilities are $Pr(M_1|y)$ and $Pr(M_2|y)$, where $Pr(M_1|y) + Pr(M_2|y) = 1$. Using Bayes' theorem, we can write these model probabilities as

$$Pr(M_1|y) = \frac{p(M_1, y)}{p(y)} = \frac{Pr(M_1)p(y|M_1)}{Pr(M_1)p(y|M_1) + Pr(M_2)p(y|M_2)}\,, \tag{11}$$

$$Pr(M_2|y) = \frac{p(M_2, y)}{p(y)} = \frac{Pr(M_2)p(y|M_2)}{Pr(M_1)p(y|M_1) + Pr(M_2)p(y|M_2)}\,. \tag{12}$$

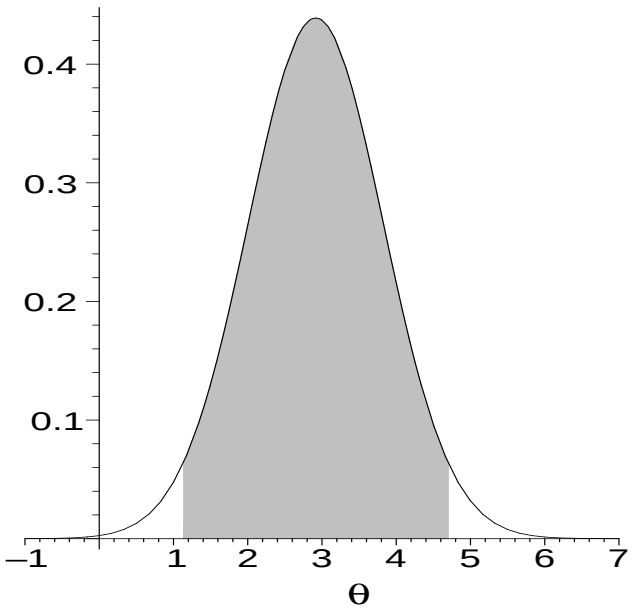

Figure 5: 95% HPD region for the average real GNP growth rate θ based on 24 quarterly observations from 1970 to 1975

The *posterior odds ratio* in favor of model 1, that is, the ratio of (11) and (12), is now defined by

$$K_{1,2} = \frac{Pr(M_1|y)}{Pr(M_2|y)} = \frac{Pr(M_1)}{Pr(M_2)} \frac{p(y|M_1)}{p(y|M_2)}\,. \tag{13}$$

Model 1 is preferred if $K_{1,2}$ is larger than 1, and model 2 is preferred in the opposite case. The relationship (13) states that the posterior odds ratio $K_{1,2}$ equals the prior odds ratio $\frac{Pr(M_1)}{Pr(M_2)}$, reflecting prior model beliefs, times the so-called *Bayes factor*

$$B_{1,2} = \frac{p(y|M_1)}{p(y|M_2)}\,, \tag{14}$$

accounting for the observed data y. We note that the posterior odds ratio equals the Bayes factor if the two models are a priori assumed to be equally likely, that is, $Pr(M_1) = Pr(M_2) = 0.5$. The subsequent discussion on Bayes factors is quite brief, but a more extensive treatment can be found in Kass and Raftery (1995).

The Bayes factor $B_{1,2}$ is the ratio of the marginal likelihoods

$$p(y|M_1) = \int p(y|\theta_1, M_1) p(\theta_1|M_1)\, d\theta_1, \tag{15}$$

$$p(y|M_2) = \int p(y|\theta_2, M_2) p(\theta_2|M_2)\, d\theta_2, \tag{16}$$

where θ_1 and θ_2 are the parameter vectors in the two models, and where the prior densities $p(\theta_1|M_1)$ and $p(\theta_2|M_2)$ and the likelihood functions $p(y|\theta_1, M_1)$ and $p(y|\theta_2, M_2)$ contain all scaling constants. It is interesting to note that the Bayes factor is closely related to the likelihood ratio. However, the latter maximizes over the model parameters, whereas the former integrates them out. Furthermore, if both the models M_1 and M_2 do not contain free parameters, then the Bayes factor is just the ratio of two likelihoods evaluated at fixed parameter values.

Example: growth of real GNP in the US (continued)

As an illustration, we consider the normal real GNP growth model with standard deviation 5. As before, the average growth parameter θ has a normal prior density with mean 4 and standard deviation 2. We use Bayes factors to compare the zero growth model M_1, imposing that $\theta = 0$, with the unrestricted model M_2. As model M_1 does not contain free parameters, the marginal likelihood for this model is just the likelihood function evaluated at $\theta = 0$, that is,

$$p(y|M_1) = p(y|\theta = 0) = (2\pi \cdot 25)^{-T/2} \exp\left(-\frac{\sum_{t=1}^{T} y_t^2}{2 \cdot 25} \right). \tag{17}$$

Furthermore, the marginal likelihood for model M_2 is

$$\begin{aligned} p(y|M_2) &= \int_{-\infty}^{\infty} p(y|\theta, M_2) p(\theta|M_2) d\theta \\ &= \int_{-\infty}^{\infty} (2\pi \cdot 25)^{-T/2} \exp\left(-\frac{\sum_{t=1}^{T} (y_t - \theta)^2}{2 \cdot 25} \right) (2\pi \cdot 4)^{-1/2} \exp\left(-\frac{(\theta - 4)^2}{2 \cdot 4} \right) d\theta \\ &= (2\pi \cdot 25)^{-T/2} \exp\left(-\frac{\sum_{t=1}^{T} y_t^2}{2 \cdot 25} \right) \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{4}{2} \right) \times \\ &\quad \times \int_{-\infty}^{\infty} \exp\left(-\frac{(T + 25/4)\theta^2 - 2\left(\sum_{t=1}^{T} y_t + 25 \right)\theta}{2 \cdot 25} \right) d\theta. \end{aligned} \tag{18}$$

As it can be shown that the integral in (18) is given by[3]

$$\int_{-\infty}^{\infty} \exp\left(-\frac{(T + 25/4)\theta^2 - 2\left(\sum_{t=1}^{T} y_t + 25 \right)\theta}{2 \cdot 25} \right) d\theta = \frac{\sqrt{2\pi}}{\sqrt{T + 25/4}} \exp\left(\frac{\left(\sum_{t=1}^{T} y_t + 25 \right)^2}{2 \cdot 25 \cdot (T + 25/4)} \right),$$

[3]Dividing the integrand by the right-hand side yields a normal density that obviously integrates to 1.

it follows from (17) and (18) that the Bayes factor $B_{1,2}$ becomes

$$\begin{aligned} B_{1,2} &= \frac{p(y|M_1)}{p(y|M_2)} \\ &= \frac{2\sqrt{2\pi}}{\sqrt{2\pi}} \exp\left(\frac{4}{2}\right) \sqrt{T+25/4}\ \exp\left(-\frac{\left(\sum_{t=1}^{T} y_t + 25\right)^2}{2 \cdot 25 \cdot (T+25/4)} \right) \\ &= \frac{(2\pi)^{-1/2} \left(\frac{T}{25}+\frac{1}{4}\right)^{1/2} \exp\left[-\frac{1}{2}\left(\frac{T}{25}+\frac{1}{4}\right) \left(0 - \frac{\sum_{t=1}^{T} y_t + 25}{T+25/4}\right)^2 \right]}{(2\pi)^{-1/2}\, 4^{-1/2}\, \exp\left(-\frac{(0-4)^2}{2\cdot 4}\right)} \\ &= \frac{p(\theta|y)\Big|_{\theta=0}}{p(\theta)\Big|_{\theta=0}} . \end{aligned} \tag{19}$$

the ratio of the posterior density and the prior density, both evaluated at the restricted parameter value $\theta = 0$.

Savage-Dickey density ratio

The remarkable result in the example above, that the Bayes factor is the ratio of the posterior density and the prior density, evaluated at the restricted parameter value, is not a coincidence. It is a special case of the Savage-Dickey density ratio (Dickey 1971). We note that the result above can also be derived immediately from Bayes' theorem (1) by evaluating it for $\theta = 0$ and rearranging it as

$$\frac{p(\theta|y)\Big|_{\theta=0}}{p(\theta)\Big|_{\theta=0}} = \frac{p(y|\theta=0)}{p(y)} = \frac{p(y|M_1)}{p(y|M_2)} . \tag{20}$$

Figure 6 provides a graphical illustration of the result. It shows that for $\theta = 0$ the unrestricted model M_2 is preferred over the restricted model M_1, as the Bayes factor $B_{1,2}$ is smaller than 1. Note that in the HPD approach the restricted model is also rejected (Figure 5). However, it is certainly possible that the HPD approach and the Bayes factor give different 'signals'. For example, the value $\theta = 4.5$ is not rejected by the HPD approach, whereas the Bayes factor favors the unrestricted model (Figure 6).

We note that the Savage-Dickey density ratio (20) implies that the restricted model M_1 would always be favored if the prior for the restricted parameters θ is

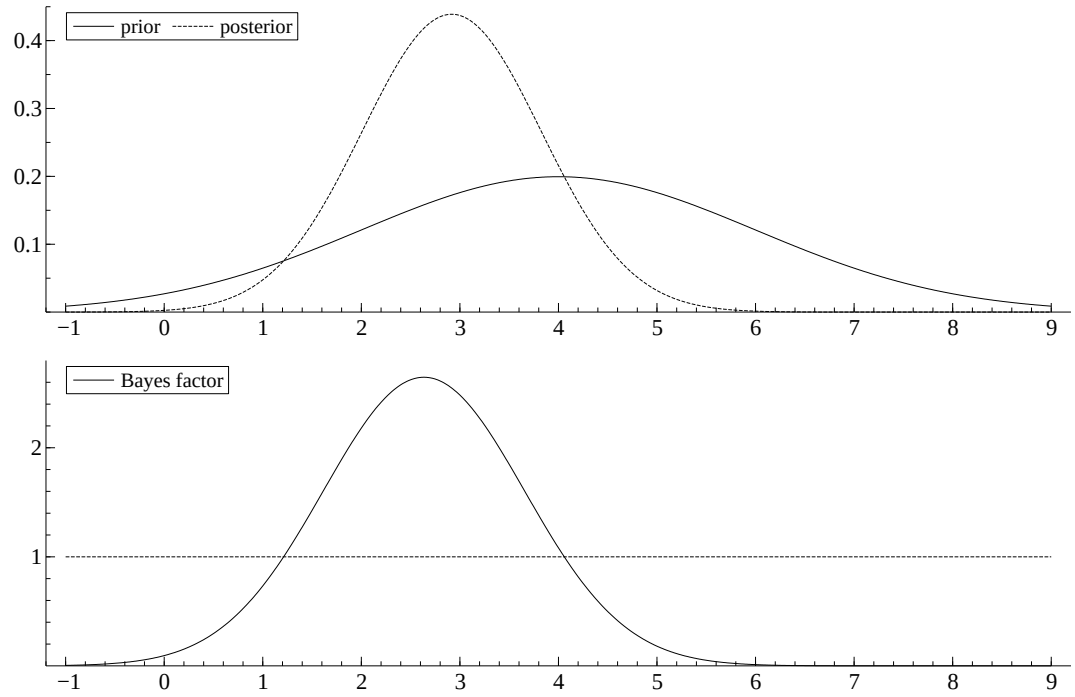

Figure 6: Prior and posterior densities for the average (annualized) real GNP growth rate, where the posterior involves 24 quarterly observations from 1970 to 1975 (above), and the Bayes factor to test that the average growth rate equals the value on the horizontal axis (below).

improper (= not integrating to a constant, "infinitely diffuse"), as the denominator in the Bayes factor $B_{1,2}$ would tend to zero. This phenomenon is the Bartlett paradox (Lindley 1957, Bartlett 1957). It demonstrates that, at least for the parameters being tested, improper priors should be avoided in posterior odds analysis.

Example: illustration of Bayes' rule, HPD region and posterior odds in World Series

Consider the following illustrative, simple model for the World Series 2004 between the Boston Red Sox and the St. Louis Cardinals. In this model we have data $y = \{y_1, \ldots, y_n\}$ with

$$y_i = \begin{cases} 1 & \text{Boston Red Sox win match i} \\ 0 & \text{St. Louis Cardinals win match i} \end{cases}, \quad i = 1, \ldots, n.$$

that are assumed independently Bernoulli(θ) distributed, i.e. the model contains only

one parameter θ, the probability that the Boston Red Sox beat the St. Louis Cardinals in match i ($i = 1, \ldots, n$). The probability distribution of y_i ($i = 1, \ldots, n$) is:

$$\mathsf{Pr}[y_i|\,\theta] = \theta^{y_i}(1-\theta)^{1-y_i}$$

leading to the likelihood:

$$\mathsf{Pr}[y|\,\theta] = \prod_{i=1}^{n} \mathsf{Pr}[y_i|\,\theta] = \theta^{n_1}(1-\theta)^{n_2}$$

with n_1 and n_2 the numbers of matches that have been won by the Boston Red Sox and the St. Louis cardinals, respectively. Suppose we have no a priori preference for the parameter θ, so we specify a uniform prior: $p(\theta) = 1$ for $\theta \in [0, 1]$, $p(\theta) = 0$ else.

In the year 2004 the World Series consisted of only 4 matches that were all won by the Boston Red Sox, so $y_i = 1$ for $i = 1, 2, 3, 4$. Hence, after n of these matches the likelihood is given by $\mathsf{Pr}[y_i|\,\theta] = \theta^n$, and the posterior density of θ is given by

$$p(\theta|\,y) \propto \mathrm{Pr}[y_i|\,\theta]\; p(\theta) = \begin{cases} \theta^n & 0 \leq \theta \leq 1 \\ 0 & \text{else} \end{cases}$$

for $n = 1, 2, 3, 4$. The scaling constant $\int \mathsf{Pr}[y_i|\,\theta]\; p(\theta) d\theta$ is

$$\int \mathrm{Pr}[y_i|\,\theta]\; p(\theta) d\theta = \int_0^1 \theta^n d\theta = \frac{1}{n+1}$$

so we have

$$p(\theta|\,y) = \frac{\mathrm{Pr}[y_i|\,\theta]\; p(\theta)}{\int \mathrm{Pr}[y_i|\,\theta]\; p(\theta) d\theta} = \begin{cases} (n+1)\theta^n & 0 \leq \theta \leq 1 \\ 0 & \text{else} \end{cases} .$$

Figure 7 shows the graphs of the prior and posterior density of θ after $n = 1, 2, 3, 4$ matches. Note that after each match - won by the Boston Red Sox - more density mass is located on the right side of $\theta = 0.5$. The posterior cumulative distribution function [CDF] of θ after $n = 1, 2, 3, 4$ matches is given by $Pr[\theta \leq \tilde{\theta}] = \tilde{\theta}^{n+1}$. So, the 95% HPD region is given by $[0.05^{1/(n+1)}, 1]$. The 95% HPD region is $[0.22, 1]$, $[0.37, 1]$, $[0.47, 1]$, $[0.55, 1]$ after $n = 1, 2, 3, 4$ observations, respectively.

We now consider a posterior odds analysis for the following two models M_1 and M_2: model M_1 in which $\theta \leq 1/2$ and model M_2 in which $\theta > 1/2$. Models 1 and 2 can be interpreted as the hypotheses that "the St. Louis Cardinals are at least as good as the Boston Red Sox" and "the Boston Red Sox are better than the St. Louis Cardinals", respectively. The prior distributions for θ under models 1 and 2 are assumed to be uniform on $[0, 1/2]$ and $(1/2, 1]$, respectively. Notice that the models M_1 and M_2 are non-nested. In the case in which the Boston Red Sox have won all matches, the

marginals likelihoods are given by:

$$\begin{aligned} p(y|M_1) &= \int p(y|\theta, M_1)p(\theta|M_1)d\theta = \int_0^{1/2} \theta^n 2 d\theta = \frac{2}{n+1}\left(\frac{1}{2}\right)^{n+1}, \\ p(y|M_2) &= \int p(y|\theta, M_2)p(\theta|M_2)d\theta = \int_{1/2}^{1} \theta^n 2 d\theta = \frac{2}{n+1}\left[1 - \left(\frac{1}{2}\right)^{n+1}\right]. \end{aligned}$$

So, if we assume equal prior probabilities $Pr[M_1] = Pr[M_2] = 0.5$, then the Bayes factor and posterior odds ratio $K_{1,2}$ are given by:

$$K_{1,2} \equiv \frac{Pr[M_1|y]}{Pr[M_2|y]} = \frac{(1/2)^{n+1}}{1-(1/2)^{n+1}}.$$

The posterior probabilities of models M_1 and M_2 are given by $Pr[M_1|y] = (1/2)^{n+1}$ and $Pr[M_2|y] = 1-(1/2)^{n+1}$. So, the probability that "the St. Louis Cardinals are at least as good as the Boston Red Sox" given n ($n = 1, 2, 3, 4$) observed matches (won by the Boston Red Sox) is $Pr[M_1|y] = (1/2)^{n+1}$, which equals 0.25, 0.125, 0.06 and 0.03 for $n = 1, 2, 3, 4$.

We now compare these conclusions of Bayesian methods with the frequentist/classical approach. In the frequentist/classical framework, a test of null hypothesis $H_0 : \theta \leq 0.5$ versus alternative hypothesis $H_1 : \theta > 0.5$ (using the number of matches won by the Boston Red Sox as a test statistic) has p-value $(1/2)^n$ after n ($n = 1, 2, 3, 4$) matches. After four matches we have a p-value of 0.06, so that at 5% size we can not even reject the null. Note that the posterior odds analysis already leads to a 'preference' of the Boston Red Sox over the St. Louis Cardinals after one match, whereas four matches are 'enough' to make the HPD region based approach lead to a rejection of $\theta = 0.5$.

2.4 Comparison of Bayesian inference and frequentist approach

In the previous subsections we have considered the principles of Bayesian inference. In order to gain more insight into the key elements of Bayesian inference, we now conclude this section with a brief comparison between Bayesian inference and the frequentist/classical approach. Table 1 provides an overview of four points at which these two approaches differ; for four elements of Bayesian inference the frequentist counterpart is given. Note that at some points the frequentist approach and Bayesian inference are each other's opposite. In the frequentist approach, the data are random and the parameters are fixed. Many realizations $\hat{\theta}$ are possible under the assumption $\theta = \theta_0$. Testing the hypothesis $\theta = \theta_0$ amounts to checking whether

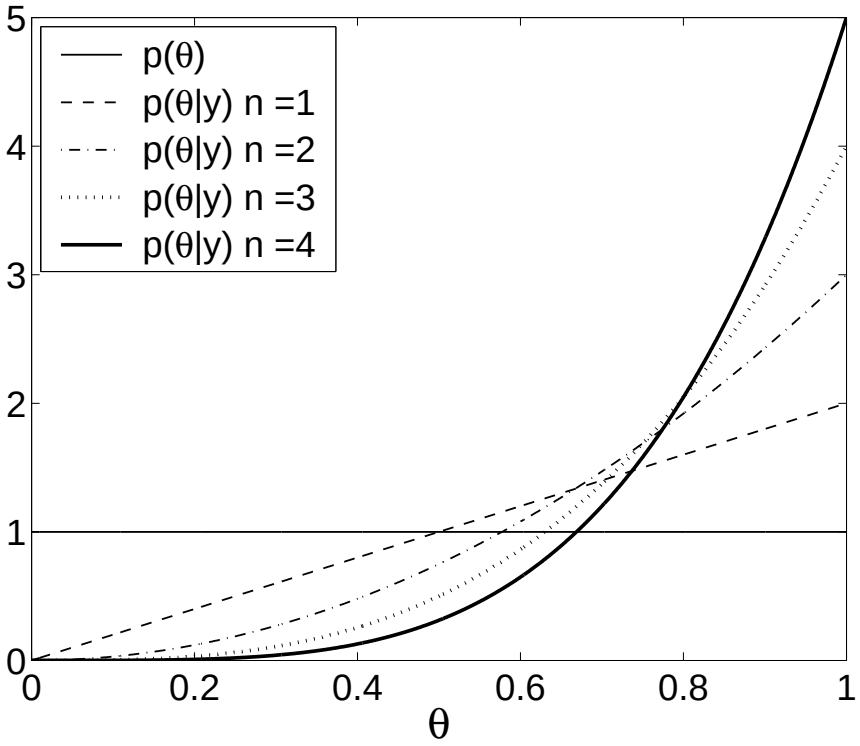

Figure 7: Prior density and posterior density of parameter θ (probability that Boston Red Sox win a match) after $n = 1, 2, 3, 4$ matches (that are won by the Boston Red Sox) in World Series of the year 2004.

the observed realization $\hat{\theta}_*$ is *plausible* under $\theta = \theta_0$ using the sampling density of $\hat{\theta}$. So, one checks whether the observed data realization is *plausible*, while (infinitely) many realizations are *possible*. On the other hand, in the Bayesian approach the parameters are random, whereas the data are given. Testing the hypothesis $\theta = \theta_0$ amounts to checking whether the value of θ_0 is *plausible* given the data. So, (infinitely) many values of θ are *possible*, but one checks whether $\theta = \theta_0$ is *plausible* under the *one* data realization.

Table 1: Comparison of frequentist (or classical) approach and Bayesian approach in a statistical/econometric model with parameter vector θ and data y

Frequentist approach	Bayesian approach
The parameters θ are fixed unknown constants. There is some unknown true value $\theta = \theta_0$.	The parameters θ are stochastic variables. One defines a prior distribution on the parameter space. All values in a certain region are possible with a certain probability density.
The data y are used to estimate and check the validity of the postulated model, by comparing data with an (infinitely large, hypothetical) data set from model.	The data y are used as evidence to update the state of the mind: data transform the prior into the posterior distribution by the likelihood.
Frequency concept of probability : a probability is the fraction of occurrences when a process is repeated infinitely often. It should be noted that, although the frequentist approach is often used in non-experimental sciences, repeating the process is only possible in experimental situations.	Subjective concept of probability : a probability is a degree of belief that an event occurs. This degree of belief is revised when new information becomes available.
One can use the maximum likelihood estimator $\hat{\theta}$ of θ as an estimator of θ.	One uses Bayes' theorem to obtain the posterior distribution of θ. One can use the posterior mean or mode as an estimator of θ.

3 Simulation Methods

3.1 Motivation for Using Simulation Techniques

The importance of integration in Bayesian inference can already be seen from the results in the previous section:

- In order to obtain the exact posterior density from Bayes' theorem one needs to evaluate the integral $p(y) = \int p(y|\theta)p(\theta)d\theta$ in the denominator of (1).
- In order to evaluate the posterior moments of (the elements of) θ, one requires additional integration of the numerator. For example, two integrals have to be evaluated for the posterior mean of θ:

$$E[\theta|y] = \int \theta p(\theta|y)d\theta = \int \theta \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}d\theta = \frac{\int \theta p(y|\theta)p(\theta)d\theta}{\int p(y|\theta)p(\theta)d\theta}.$$

- In order to evaluate the posterior odds ratio in favor of model 1 versus model 2, one needs to evaluate two marginal likelihoods, and hence two integrals.

Note that in linear and binomial models (for certain prior specifications) these integrals can be computed analytically. For more complicated models, it's usually impossible to find analytical solutions. In general, we need numerical integration methods for Bayesian inference. Basically there are two numerical integration methods: deterministic integration and Monte Carlo integration. Deterministic integration consists of evaluating the integrand at a set of many fixed points, and approximating the integral by a weighted average of the function evaluations. Monte Carlo integration is based on the idea that $E[g(\theta)|y]$, the mean of a certain function $g(\theta)$ under the posterior, can be approximated by its 'sample counterpart', the sample mean $\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)$, where $\theta_1, \ldots, \theta_n$ are drawn from the posterior distribution.

At a first glance, deterministic integration may always seem a better idea than Monte Carlo integration, as no extra uncertainty (caused by the required random variables) is added to the procedure. However, in deterministic integration the number of required function evaluations increases exponentially with the dimension of the integration problem, which is in our case the dimension k of the vector θ. Therefore, deterministic integration approaches like quadrature methods become unworkable if k exceeds, say, three. So, in many cases one has to make use of Monte Carlo integration. However, only for a very limited set of models and prior densities it is possible to directly draw random variables from the posterior distribution. Then one may use indirect sampling algorithms such as importance sampling or Markov chain Monte Carlo (MCMC) methods such as the Metropolis-Hastings algorithm. In the following subsections direct sampling methods, importance sampling and MCMC methods will be discussed.

3.2 Direct sampling methods

Only in the ideal case, Monte Carlo integration reduces to estimating the posterior expectation $E[g(\theta)|y]$ by the sample mean $g_{\mathrm{DS}} = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i)$, where $\theta_1, \ldots, \theta_n$ are directly sampled from the posterior. However, even when the posterior distribution is non-standard, direct sampling methods are useful, as they can serve as building blocks for more involved algorithms. For example, any sampling algorithm is based on collecting draws from the uniform $U(0,1)$ distribution, so that suitable methods to generate these "random numbers" are of utmost importance.

3.2.1 Uniform sampling

The most commonly used method to sample from the uniform distribution is the linear congruential random number generator [LCRNG], initially introduced by Lehmer (1951). This generator creates a sequence of "random numbers" $u_1, \ldots, u_n$ using the recursion

$$u_i = (a\, u_{i-1} + b) \bmod M, \qquad i = 1, \ldots n, \tag{21}$$

where mod M gives the remainder after division by M. The multiplier a and the modulus M are strictly positive integers, while the increment b is also allowed to be zero. The initial value u_0 of the sequence is called the seed. In order to map $u_1, \ldots, u_n$ to the unit interval, these values are divided by M. We note that the recursion (21) is completely deterministic, so that the generated "random numbers" are actually not random at all. For properly chosen a, b and M, it only seems *as if* they are random. In practice, multiplicative LCRNGs are frequently considered. These arise from (21) by setting $b = 0$, so that the increment is turned off. Two very popular multiplicative LCRNGs are the Lewis-Goodman-Miller generator (Lewis et al. 1969), obtained by setting $a = 16807$ and $M = 2^{31} - 1$, and the Payne-Rabung-Bogyo generator (Payne et al. 1969), obtained by setting $a = 630360016$ and $M = 2^{31} - 1$. This concludes our discussion on uniform sampling. For a more comprehensive text on generating pseudo-random numbers, the reader is referred to Law and Kelton (1991).

3.2.2 Inversion method

The inversion method is an approach which directly translates uniform $U(0,1)$ draws into draws from the (univariate) distribution of interest. The underlying idea is very simple. If the random variable θ follows a distribution with cumulative distribution function (CDF) denoted by F, then the corresponding CDF value $U = F(\theta)$ is uniformly distributed, as

$$\Pr(U \leq u) = \Pr(F(\theta) \leq u) = \Pr(\theta \leq F^{-1}(u)) = F(F^{-1}(u)) = u \tag{22}$$

with F^{-1} denoting the inverse CDF. By relying on this result, the inversion method consists of first collecting a uniform sample $u_1, \ldots, u_n$, and subsequently transform-

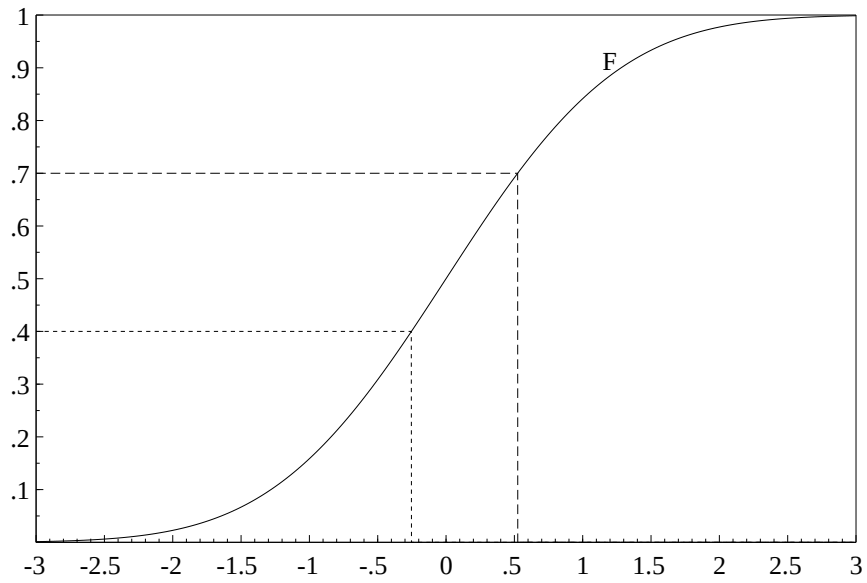

Figure 8: Illustration of the inversion method for the standard normal distribution. The uniform draws $u_1 = 0.4$ *and* $u_2 = 0.7$ *correspond to the standard normal realizations* $x_1 \approx -0.25$ *and* $x_2 \approx 0.52$*, respectively.*

ing this sample into realizations $\theta_1 = F^{-1}(u_1), \ldots, \theta_n = F^{-1}(u_n)$ from the distribution of interest. Figure 8 illustrates the inversion method for the standard normal distribution. Clearly, as the standard normal CDF is steepest around 0, that region is "hit" most frequently, so that most draws have values relatively close to 0. On the other hand, not many draws fall into regions far away from 0, as these regions are difficult to "hit". This mechanism causes that draws are assigned to regions in accordance with their probability mass. We note that the inversion method is particularly suited to sample from (univariate) truncated distributions. For example, if a distribution is truncated to the left of some value a and to the right of some value b, then all draws should fall into the region (a,b). This is easily achieved by sampling $u_1 \ldots, u_n$ uniformly on the interval $(F(a), F(b))$, instead of sampling them on the interval $(0, 1)$. All that has to be done is redefining

$$u_i \equiv F(a) + [\, F(b) - F(a) \,]\, u_i, \qquad i = 1, \ldots n. \tag{23}$$

For the inversion method, it is desirable that the inverse CDF F^{-1} can be evaluated easily. If F^{-1} has a closed-form expression, evaluation becomes trivial. For example, the exponential distribution with mean $\frac{1}{\lambda}$ has CDF

$$F(\theta) = 1 - \exp(-\lambda\theta), \qquad \theta > 0. \tag{24}$$

By solving

$$u = F(\theta) = 1 - \exp(-\lambda\theta) \tag{25}$$

for θ, it is seen that the inverse CDF is given by

$$\theta = F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u). \tag{26}$$

As the random variable $U = F(\theta)$ has the same uniform distribution as $1 - U$, it follows from (26) that a sample $\theta_1, \ldots, \theta_n$ from the exponential distribution is obtained by applying the algorithm

Generate $u_1, \ldots, u_n$ from $U(0,1)$.
Transform to $\theta_i = -\frac{1}{\lambda}\ln(u_i), \quad i = 1, \ldots, n.$

Although it is desirable that the inverse CDF F^{-1} has a closed form expression, this is not required. It is not even necessary that the CDF itself has a closed form expression. However, in such situations one has to resort to some numerical approximation. For example, an approximating CDF can be constructed by evaluating the probability density function (or some kernel) at many points to build a grid, and using linear interpolation. As the resulting approximation is piecewise linear, inversion is straightforward. This strategy underlies the griddy Gibbs sampling approach of Ritter and Tanner (1992), which will be discussed later on.

3.3 Indirect sampling methods that yield independent draws

If it is difficult to sample directly from the distribution of interest, hereafter referred to as the target distribution, indirect methods might be considered. Such methods aim to collect a representative sample for the target distribution by considering an alternative "candidate" distribution. This candidate distribution should be easy to sample from and it hopefully provides a reasonable approximation to the original target distribution. Indirect sampling methods involve some correction mechanism to account for the difference between the target density and the candidate density. In this section, we discuss two indirect sampling approaches resulting in independent draws, so that the Law of Large Numbers [LLN] and the Central Limit Theorem [CLT] still apply.

3.3.1 Rejection sampling

The first indirect method we discuss is rejection sampling. Following this approach, one collects a sample from the candidate distribution, and decides for each draw whether it is accepted or rejected. If a draw is accepted, it is included in the sample for the target distribution. Rejection means that the draw is thrown away. Note that the rejection step is the correction mechanism which is employed in rejection sampling.

In order to apply the rejection method to some target density P, one first needs to specify an appropriate candidate density Q. For example, one might consider some normal or Student-t density. Next, some constant c has to be found such that

$$P(\theta) \leq c\,Q(\theta) \tag{27}$$

for all θ, so that the graph of the kernel $c\,Q$ of the candidate density is entirely located above the graph of the target density P. We note that (27) implies that P is allowed to be a kernel of the target density, as the constant c can always adjust to P. However, the candidate density Q should be such that the ratio $\frac{P(\theta)}{Q(\theta)}$ is bounded for all θ, so that c is finite. Essentially, the rejection method consists of uniformly sampling points below the graph of $c\,Q$, and accepting the horizontal positions of the points falling below the graph of P. The remaining points are rejected. This idea is illustrated by Figure 9 for a bimodal target density. The coordinates of the points below the $c\,Q$ graph are sampled as follows. The horizontal position θ is obtained by drawing it from the candidate distribution with density Q. Next, the vertical position $\tilde{\theta}$ is uniformly sampled from the interval $(0, c\,Q(\theta))$. As the point $(\theta, \tilde{\theta})$ is accepted if and only if $\tilde{\theta}$ is located in the interval $(0, P(\theta))$, the acceptance probability for this point is given by $\frac{P(\theta)}{c\,Q(\theta)}$. The following rejection algorithm collects a sample of size n from the target distribution with density P:

Initialize the algorithm:

The set of accepted draws S is empty: $S = \emptyset$.

The number of accepted draws i is zero: $i = 0$.

Do while $i < n$:

Obtain θ from candidate distribution with density q.

Obtain u from uniform distribution $U(0, 1)$.

If $u < \frac{P(\theta)}{c\,Q(\theta)}$ then accept θ:

Add θ to the set of accepted draws: $S = S \cup \{\theta\}$.

Update the number of accepted draws: $i = i + 1$.

We note that although rejection sampling is based on using an *approximating* candidate distribution, the method yields an *exact* sample for the target distribution. However, the big drawback of the rejection approach is that many candidate draws might be required to obtain an accepted sample of moderate size, making the method inefficient. For example, in Figure 9 it is seen that most points are located above the P graph, so that many draws are thrown away. For large n, the fraction of accepted draws tends to the ratio of the area below the P graph and the area below the $c\,Q$

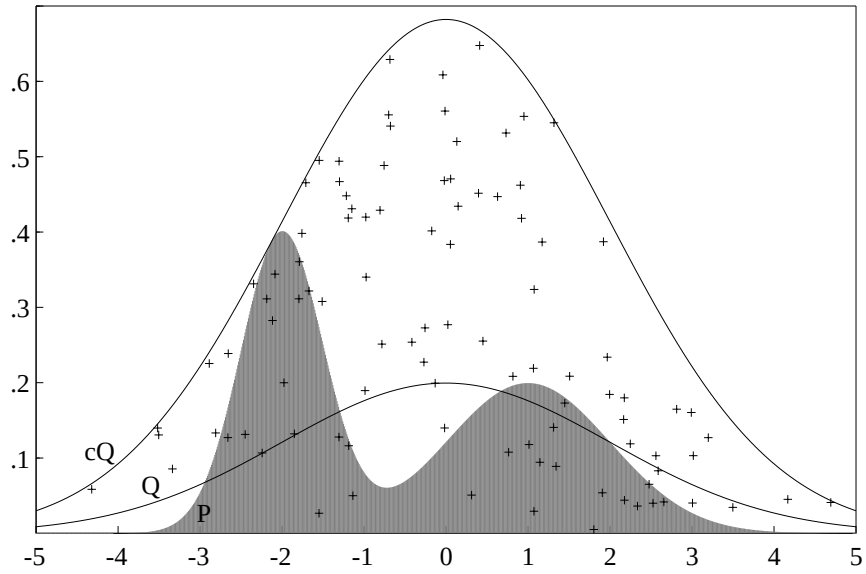

Figure 9: Illustration of rejection sampling. The candidate density Q is blown up by a factor c such that its graph is entirely located above the graph of the target density P. Next, points are uniformly sampled below the cQ graph, and the horizontal positions of the points falling into the shaded area below the P graph are accepted.

graph. As the candidate density Q integrates to one, this acceptance rate is given by $\int P(\theta)\, d\theta/c$, so that a smaller value for c results in more efficiency. Clearly, c is optimized by setting it at

$$c = \max_{\theta} \frac{P(\theta)}{Q(\theta)}\,, \tag{28}$$

implying that the optimal c is small if variation in the ratio $\frac{P(\theta)}{Q(\theta)}$ is small. This explains that a candidate density, providing a good approximation to the target density, is desirable.

3.3.2 Importance sampling

Importance sampling is another indirect approach to obtain an estimate for $E[g(\theta)]$, where θ is a random variable from the target distribution. It is initially discussed by Hammersley and Handscomb (1964) and introduced in econometrics by Kloek and Van Dijk (1978). The method is related to rejection sampling. The rejection method either accepts or rejects candidate draws, that is, draws either receive full weight or they do not get any weight at all. Importance sampling is based on this notion of assigning weights to draws. However, in contrast with the rejection method, these weights are not based on an all-or-nothing situation. Instead, they can take any possible value, representing the relative importance of draws. If Q is the candidate

density (= importance function) and P is a kernel of the target density, importance sampling is based on the relationship

$$E[g(\theta)] = \frac{\int g(\theta)P(\theta)\,d\theta}{\int P(\theta)\,d\theta} = \frac{\int g(\theta)w(\theta)Q(\theta)\,d\theta}{\int w(\theta)Q(\theta)\,d\theta} = \frac{E[w(\tilde{\theta})g(\tilde{\theta})]}{E[w(\tilde{\theta})]}\,, \tag{29}$$

where $\tilde{\theta}$ is a random variable from the candidate distribution, and $w(\tilde{\theta}) = \frac{P(\tilde{\theta})}{Q(\tilde{\theta})}$ is the weight function, which should be bounded. It follows from (29) that a consistent estimate of $E[g(\theta)]$ is given by the weighted mean

$$\widehat{E[g(\theta)]}_{IS} = \frac{\sum_{i=1}^{n} w(\tilde{\theta}_i)g(\tilde{\theta}_i)}{\sum_{j=1}^{n} w(\tilde{\theta}_j)}\,, \tag{30}$$

where $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$ are realizations from the candidate distribution and $w(\tilde{\theta}_1), \ldots, w(\tilde{\theta}_n)$ are the corresponding weights. As relationship (29) would still hold after redefining the weight function as $w(\tilde{\theta}) = \frac{P(\tilde{\theta})}{c\,Q(\tilde{\theta})}$, yielding the acceptance probability of $\tilde{\theta}$, there exists a clear link between rejection sampling and importance sampling, that is, the importance sampling method weights draws with the acceptance probabilities from the rejection approach. Figure 10 provides a graphical illustration of the method. Points for which the graph of the target density is located above the graph of the candidate density are not sampled often enough. In order to correct for this, such draws are assigned relatively large weights (weights larger than one). The reverse holds in the opposite case. We note that although importance sampling can be used to estimate characteristics of the target density (such as the mean), it does not provide a sample according to this density, as draws are generated from the candidate distribution. So, in a strict sense, importance sampling should not be called a sampling method but it should be called a pure integration method.

The performance of the importance sampler is greatly affected by the choice of the candidate distribution. If the importance function Q is inappropriate, the weight function $w(\tilde{\theta}) = \frac{P(\tilde{\theta})}{Q(\tilde{\theta})}$ varies a lot and it might happen that only a few draws with extreme weights almost completely determine the estimate $\widehat{E[g(\theta)]}_{IS}$. This estimate would be very unstable. In particular, a situation such that the tails of the target density are fatter than the tails of the candidate density is concerning, as this would imply that the weight function might even tend to infinity. In such a case, $E[g(\theta)]$ does not exist, see (29). It is for this reason that a fat-tailed Student-t importance function is usually preferred over a normal candidate density.

Using importance sampling to compute the marginal likelihood

As an application of importance sampling, we show how it can be used to com-

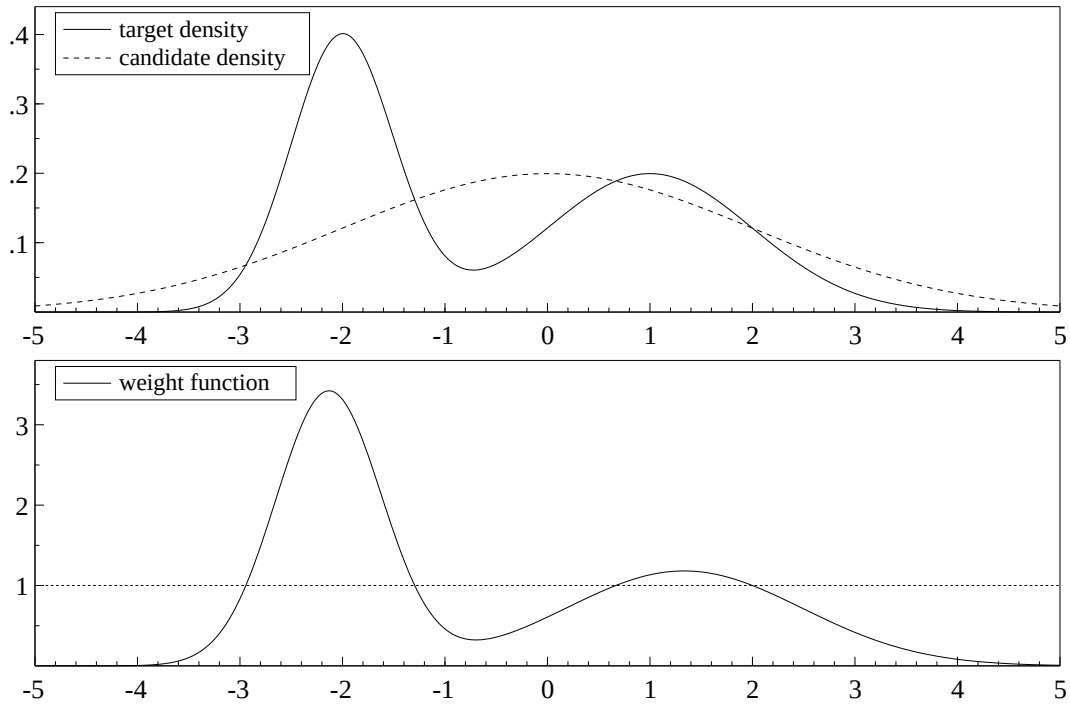

Figure 10: Illustration of importance sampling. The weight function reflects the importance of draws from the candidate density.

pute the marginal likelihood

$$p(y) = \int p(y|\theta)p(\theta)\, d\theta, \tag{31}$$

where y denotes the data and θ is the parameter vector. The most straightforward approach to estimate (31) is based on the interpretation $p(y) = E[p(y|\theta)]$, where the expectation is taken with respect to θ that obeys its prior distribution with density $p(\theta)$. The resulting estimate is given by

$$\hat{p}_A = \frac{1}{n}\sum_{i=1}^{n} p(y|\theta_i), \tag{32}$$

where $\theta_1, \ldots, \theta_n$ are sampled from the prior distribution. However, this approach is inefficient if the likelihood is much more concentrated than the prior, as most draws from the prior would correspond to extremely small likelihood values. Consequently, $\hat{p}_A$ would be determined by only a few draws with relatively large likelihood values. As an alternative, Newton and Raftery (1994) develop an estimate for $p(y)$ which is based on the importance sampling approach. Using the interpretation that the

marginal likelihood is $p(y) = E[p(y|\theta)]$, one is clearly interested in $E[g(\theta)]$, where $g(\theta) = p(y|\theta)$ is the likelihood value. Next, as the expectation is taken with respect to θ obeying its prior distribution, the target density is the prior, that is, $P(\theta) = p(\theta)$. Finally, by considering the posterior density as the candidate density, that is, $Q(\theta) = p(\theta|y) \propto p(y|\theta)p(\theta)$, the weight function becomes

$$w(\theta) = \frac{P(\theta)}{Q(\theta)} \propto \frac{p(\theta)}{p(y|\theta)p(\theta)} = p(y|\theta)^{-1}. \tag{33}$$

This results in the importance sampling estimate

$$\hat{p}_{NR} = \frac{\sum_{i=1}^{n} w(\theta_i)g(\theta_i)}{\sum_{j=1}^{n} w(\theta_j)} = \frac{\sum_{i=1}^{n} p(y|\theta_i)^{-1}p(y|\theta_i)}{\sum_{j=1}^{n} p(y|\theta_j)^{-1}} = \left(\frac{1}{n}\sum_{j=1}^{n} p(y|\theta_j)^{-1}\right)^{-1}, \tag{34}$$

where $\theta_1, \ldots, \theta_n$ are sampled from the posterior distribution. We note that the posterior density $p(\theta|y)$ (used in $\hat{p}_{NR}$) usually gives a much better approximation to the likelihood $p(y|\theta)$ than the prior $p(\theta)$ (used in $\hat{p}_A$). In particular, this holds if data information strongly dominates prior information, which is the case if many observations are used. However, a drawback of the harmonic mean $\hat{p}_{NR}$ is that it is consistent but also unstable, as the weight function $w(\theta) = p(y|\theta)^{-1}$ takes extreme values for occasionally sampled θ_j for which the likelihood value $p(y|\theta_j)$ is very small. In order to overcome this objection, several modifications and generalizations of $\hat{p}_{NR}$ are proposed, see for example Gelfand and Dey (1994), and Newton and Raftery (1994).

3.4 Markov chain Monte Carlo methods

Another approach to sample from non-standard distributions is the Markov Chain Monte Carlo [MCMC] approach. An MCMC method aims to collect a sample representative for the target distribution by construction of a Markov chain converging to that distribution. After a sufficiently long burn-in period, so that the influence of the initialization conditions has become negligible, draws from the Markov chain are regarded as draws from the target distribution itself. However, as Markov chain sampling naturally induces correlation, the resulting draws are not independent, so that the Law of Large Numbers [LLN] and the Central Limit Theorem [CLT] no longer apply. For ease of exposition, we only consider Markov chain theory for discrete state spaces, but the obtained results can be extended immediately to continuous distributions. The reader is referred to Norris (1997) and Ross (1997) for textbook discussions on Markov chain theory.

Elementary Markov chain theory

In order to make this section self-contained, we start with reviewing some elementary Markov chain theory. A *Markov chain* is a discrete-time stochastic process $\{\theta_0, \theta_1, \dots\}$ satisfying the Markov property, that is, the next state only depends on the current state and does not depend on the path of previous states. For a finite discrete state space S, the one-step transition probability from state θ to state $\tilde{\theta}$ is denoted by

$$P(\theta, \tilde{\theta}) = \Pr(\theta_{i+1} = \tilde{\theta} | \theta_i = \theta), \tag{35}$$

where $\theta, \tilde{\theta} \in S$. For example, we could specify a Markov chain process for a time series indicating whether an economy is in a recession or expansion; given that the current period is a recession, there is a certain probability $\tilde{p}$ of escaping the recession in the next period, and a probability $1 - \tilde{p}$ of staying in the recession.

By definition, it should hold that $P(\theta, \tilde{\theta}) \geq 0$ and $\sum_{\tilde{\theta} \in S} P(\theta, \tilde{\theta}) = 1$. Similarly, the j-step transition probability is denoted by

$$P^{(j)}(\theta, \tilde{\theta}) = \Pr(\theta_{i+j} = \tilde{\theta} | \theta_i = \theta), \tag{36}$$

where $\theta, \tilde{\theta} \in S$. We note that (36) can be computed by summing the probabilities of all paths moving from state θ to state $\tilde{\theta}$ in j steps. Under mild regularity conditions, it can be shown that the Markov chain converges to a unique distribution

$$P(\tilde{\theta}) = \lim_{j \to \infty} P^{(j)}(\theta, \tilde{\theta}), \tag{37}$$

not depending on the initial state and satisfying the *"invariance" condition*

$$P(\tilde{\theta}) = \sum_{\theta \in S} P(\theta) P(\theta, \tilde{\theta}) \tag{38}$$

for all $\tilde{\theta} \in S$. Intuitively, condition (38) says that the long-run proportion of states being $\tilde{\theta}$ is given by the limiting probability $P(\tilde{\theta})$. The regularity conditions which have to be satisfied are *irreducibility* and *aperiodicity*. The first requirement means that all states in the state space are accessible from each other, that is, for all $\theta, \tilde{\theta} \in S$, there exists a non-negative integer k such that $P^{(k)}(\theta, \tilde{\theta}) > 0$. The second requirement means that, for any state $\theta \in S$, the number of transitions necessary to return to state θ does not need to be a multiple of some integer ≥ 2. The two regularity conditions are, for example, satisfied if $P(\theta, \tilde{\theta}) > 0$ for all $\theta, \tilde{\theta} \in S$, that is, if it is possible to go from each state to any other state in one transition. Next, we note that an irreducible and aperiodic Markov chain running backward is again a Markov chain. After a sufficiently long burn-in period, the transition probabilities of the reversed process are given by

$$R(\theta, \tilde{\theta}) = \Pr(\theta_i = \tilde{\theta} | \theta_{i+1} = \theta) = \frac{\Pr(\theta_i = \tilde{\theta}) \Pr(\theta_{i+1} = \theta | \theta_i = \tilde{\theta})}{\Pr(\theta_{i+1} = \theta)} = \frac{P(\tilde{\theta}) P(\tilde{\theta}, \theta)}{P(\theta)}, \tag{39}$$

where $\theta, \tilde{\theta} \in S$. The Markov chain is called *time-reversible* if it has the same transition probabilities as its reversal, that is, if $P(\theta, \tilde{\theta}) = R(\theta, \tilde{\theta})$ for all $\theta, \tilde{\theta} \in S$. It is seen from (39) that this time-reversibility condition amounts to

$$P(\theta)P(\theta, \tilde{\theta}) = P(\tilde{\theta})P(\tilde{\theta}, \theta) \tag{40}$$

for all $\theta, \tilde{\theta} \in S$. Intuitively, condition (40) says that, in the long-run, the process moves as much from θ to $\tilde{\theta}$ as it moves from $\tilde{\theta}$ to θ.

3.4.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings [MH] algorithm, introduced by Metropolis et al. (1953) and generalized by Hastings (1970), samples from a time-reversible Markov chain converging to the target distribution. It has similarities with rejection sampling, as some rejection mechanism is involved. However, rejected draws are dealt with in a different way. An excellent exposition on the MH algorithm is Chib and Greenberg (1995) in which theory and intuition as well as application of the algorithm are discussed. An important survey on the broader class of MCMC methods is Tierney (1994).

The intuitive derivation of the MH algorithm starts from the time-reversibility condition

$$P(\theta)P_{MH}(\theta, \tilde{\theta}) = P(\tilde{\theta})P_{MH}(\tilde{\theta}, \theta) \qquad \forall \theta, \tilde{\theta} \in S, \tag{41}$$

where P is a kernel of the target probability function and P_{MH} is an appropriate but currently unknown transition density. So, the limiting distribution of the Markov chain is available, but the underlying process is not. Note that this is the opposite of the situation in which one knows the transition process and has to derive the limiting distribution, which is often encountered in Markov chain theory. The key idea is that if the transition probabilities $P(\theta, \tilde{\theta})$ and $P(\tilde{\theta}, \theta)$ satisfy the time-reversibility condition (41) for the given target probabilities $P(\theta)$ and $P(\tilde{\theta})$ for each $\theta, \tilde{\theta} \in S$, then this implies that the limiting distribution of the Markov chain is the desired target distribution with probability function $P(\theta)$. The reason is that the time-reversibility property implies that the invariance condition (38) is satisfied:

$$\sum_{\theta \in S} P(\theta)P(\theta, \tilde{\theta}) = \sum_{\theta \in S} P(\tilde{\theta})P(\tilde{\theta}, \theta) = P(\tilde{\theta}) \sum_{\theta \in S} P(\tilde{\theta}, \theta) = P(\tilde{\theta}), \tag{42}$$

where the first equality follows from the time-reversibility property, and the last equality obviously holds as the conditional probabilities of θ given $\tilde{\theta}$ have to sum to 1. Intuitively, it is clear that a Markov chain satisfying the time-reversibility condition for the given target probabilities must have this target distribution as its limiting distribution. When the Markov chain reaches the target distribution at a certain step, all following steps will have this target distribution: at each following iteration, each point $\theta \in S$ 'gets back' exactly the same probability mass that 'leaves' to any other point $\tilde{\theta} \in S$.

So, we are looking for a Markov chain with transition probabilities satisfying the time-reversibility condition (41). What may still seem to be an impossible task, that is, recovering such a Markov chain, can be done by considering the following approach. Suppose that the unknown transition density P_{MH} is replaced by some known but probably inappropriate candidate transition density Q satisfying irreducibility and aperiodicity. Unless Q satisfies the time-reversibility condition for all $\theta, \tilde{\theta} \in S$, which is extremely unlikely, there exist states θ and $\tilde{\theta}$ such that the probability of going from θ to $\tilde{\theta}$ is larger than the probability of going from $\tilde{\theta}$ to θ:

$$P(\theta)Q(\theta, \tilde{\theta}) > P(\tilde{\theta})Q(\tilde{\theta}, \theta), \tag{43}$$

where we note that only the "greater than" inequality $>$ is considered, as the "less than" inequality $<$ amounts to just interchanging the arbitrary states θ and $\tilde{\theta}$. In order to deal with the violation of the time-reversibility condition, a function $\alpha : S \times S \to [0, 1]$, indicating the probability of accepting a transition, is introduced such that

$$P(\theta)Q(\theta, \tilde{\theta})\alpha(\theta, \tilde{\theta}) = P(\tilde{\theta})Q(\tilde{\theta}, \theta)\alpha(\tilde{\theta}, \theta). \tag{44}$$

As the right-hand-side value of (43) is too small as compared with the left-hand-side, $\alpha(\tilde{\theta}, \theta)$ is set at its maximum value, which is 1 (since it is a probability):

$$\alpha(\tilde{\theta}, \theta) = 1. \tag{45}$$

Next, substituting (45) into (44) yields

$$\alpha(\theta, \tilde{\theta}) = \frac{P(\tilde{\theta})Q(\tilde{\theta}, \theta)}{P(\theta)Q(\theta, \tilde{\theta})} < 1. \tag{46}$$

It follows from (45) and (46) that the function α is defined by

$$\alpha(\theta, \tilde{\theta}) = \min\left\{ \frac{P(\tilde{\theta})Q(\tilde{\theta}, \theta)}{P(\theta)Q(\theta, \tilde{\theta})}, 1 \right\}, \tag{47}$$

where $\theta, \tilde{\theta} \in S$. Now, a first proposal for the unknown transition density P_{MH} might be such that $P_{MH}(\theta, \tilde{\theta}) = Q(\theta, \tilde{\theta})\alpha(\theta, \tilde{\theta})$ for all $\theta, \tilde{\theta} \in S$. However, as Q is already a transition density integrating to unity, and there exist θ and $\tilde{\theta}$ such that $\alpha(\theta, \tilde{\theta}) < 1$, this proposal cannot be a transition density itself. However, the "insufficient candidate probability mass problem" is easily fixed by adjusting $P_{MH}(\theta, \theta)$ for which the time-reversibility condition is satisfied by definition. For a discrete state space S, the adjusted transition density is defined by

$$P_{MH}(\theta, \tilde{\theta}) = Q(\theta, \tilde{\theta})\alpha(\theta, \tilde{\theta}), \qquad \tilde{\theta} \neq \theta, \tag{48}$$

$$P_{MH}(\theta, \theta) = 1 - \sum_{\tilde{\theta} \neq \theta} Q(\theta, \tilde{\theta})\alpha(\theta, \tilde{\theta}) = Q(\theta, \theta) + \sum_{\tilde{\theta} \neq \theta} Q(\theta, \tilde{\theta})(1 - \alpha(\theta, \tilde{\theta})), \tag{49}$$

where $\alpha(\theta, \tilde{\theta})$ is given by (47).

The MH algorithm is an interpretation of (48) and (49). For some current state θ, one can make a transition according to the transition density P_{MH} by drawing a candidate state $\tilde{\theta}$ from the density Q and accepting the transition, which is from θ to $\tilde{\theta}$, with probability $\alpha(\theta, \tilde{\theta})$. Acceptance implies that the move is made, that is, the next state is $\tilde{\theta}$. Rejection means that the move is not made, that is, the next state is again θ. By repeating this procedure many times, a Markov chain is constructed. After a burn-in period, draws from the Markov chain are regarded as draws from the target distribution. A sufficient condition for (long-run) convergence is that $Q(\theta, \tilde{\theta}) > 0$ for all θ and $\tilde{\theta}$ such that $P(\tilde{\theta}) > 0$. The MH algorithm constructs a Markov chain of length n as follows:

Initialize the algorithm:

Choose a feasible initial state θ_0.

Do for $i = 1, \ldots, n$:

Obtain $\tilde{\theta}$ from candidate transition density $Q(\theta_{i-1}, \cdot)$.

Obtain u from uniform distribution $U(0, 1)$.

Compute transition probability $\alpha(\theta_{i-1}, \tilde{\theta})$, defined by (47).

If $u < \alpha(\theta_{i-1}, \tilde{\theta})$ then accept transition:

$\theta_i = \tilde{\theta}$.

Else reject transition:

$\theta_i = \theta_{i-1}$.

Several approaches can be adopted to specify the candidate transition density Q, Frequently, Q is such that the resulting Markov chain is either an "independence chain" or a "random walk chain". An independence chain has the property that the candidate state $\tilde{\theta}$ is drawn independently of the current state θ, that is,

$$Q(\theta, \tilde{\theta}) = Q(\tilde{\theta}), \tag{50}$$

where $\theta, \tilde{\theta} \in S$. Typical choices for the candidate density $Q(\tilde{\theta})$ are normal or Student-t densities. It follows from (47) and (50) that the acceptance probability in an independence chain is given by

$$\alpha(\theta, \tilde{\theta}) = \min\left\{ \frac{P(\tilde{\theta})Q(\theta)}{P(\theta)Q(\tilde{\theta})}, 1 \right\} = \min\left\{ \frac{w(\tilde{\theta})}{w(\theta)}, 1 \right\}, \tag{51}$$

that is, the minimum of a ratio of importance weights and one. The interpretation of (51) is that a transition from θ to $\tilde{\theta}$ resulting in a larger importance weight is always made, whereas a transition resulting in a smaller importance weight is not always performed. We note that (51) establishes a link with importance sampling. As an alternative to the independence chain, we have also mentioned the random walk chain. A random walk chain draws the transition step $\tilde{\theta} - \theta$ instead of the state $\tilde{\theta}$ resulting from this transition, that is,

$$Q(\theta, \tilde{\theta}) = Q(\tilde{\theta} - \theta). \tag{52}$$

Typical choices for $Q(\tilde{\theta} - \theta)$ are normal or Student-t densities centered around 0, so that the expectation of the next state $\tilde{\theta} = \theta + (\tilde{\theta} - \theta)$ is the current state θ. Finally, we mention that if the transition density is symmetric, that is, $Q(\theta, \tilde{\theta}) = Q(\tilde{\theta}, \theta)$ for all $\theta, \tilde{\theta} \in S$, the acceptance probability $\alpha(\theta, \tilde{\theta})$ reduces to

$$\alpha(\theta, \tilde{\theta}) = \min\left\{ \frac{P(\tilde{\theta})}{P(\theta)}, 1 \right\}, \tag{53}$$

as in the original Metropolis algorithm (Metropolis et al. 1953). The acceptance probability (53) has a similar interpretation as (51). A transition from θ to $\tilde{\theta}$ implying an increase in the target density is always made, whereas a transition implying a decrease is not always performed.

3.4.2 Gibbs sampling

The MH algorithm is a very general MCMC approach; one can generally apply it – *given* that one has a good candidate density, of course. A more problem specific method within the MCMC class is the Gibbs sampling algorithm of Geman and Geman (1984). The Gibbs sampler is based on decomposing the multidimensional random variable θ into k components $\theta^1, \ldots, \theta^k$, which are not necessarily univariate. It constructs a Markov chain, converging to the target distribution, by iteratively drawing the k components of θ conditional on the values of all other components. Gibbs sampling may be seen as an application of the divide-and-conquer principle. For many seemingly intractable target densities, it is possible to derive a set of conditional densities for which sampling is straightforward. The Gibbs sampler exploits this notion, as it precisely considers these conditional densities. Its usefulness is, for example, demonstrated by Gelfand et al. (1990), Gelfand and Smith (1990), and Smith and Roberts (1993). Casella and George (1992) provide a tutorial on Gibbs sampling using simple examples to explain how and why the method works. As Gibbs sampling is an intuitively simple method that enables simulation from posterior distributions - and hence Bayesian inference - in many models that are useful for decision making and forecasting in practice, the Gibbs sampler has become enormously popular.

An obvious requirement for the Gibbs sampler is that all full conditional distributions can be sampled from. These conditional distributions are described by the densities $P(\theta^j|\theta^{-j})$, $j = 1, \ldots, k$, where $\theta^{-j} = (\theta^1, \ldots, \theta^{j-1}, \theta^{j+1}, \ldots \theta^k)$ denotes the set of $k-1$ components excluding the j-th component. The Gibbs sampling algorithm collects n draws $\theta_i = (\theta_i^1, \ldots, \theta_i^k)$, $i = 1, \ldots, n$, as follows. The components θ_i^j, $i = 1, \ldots, n$, $j = 1, \ldots, k$, are augmented into a single sequence $\theta_1^1, \ldots, \theta_1^k, \theta_2^1, \ldots, \theta_2^k, \ldots, \theta_n^1, \ldots, \theta_n^k$, and the elements of this Gibbs sequence are generated such that

$$\theta_i^j \text{ results from } P(\theta^j|\theta_{i-1}^{-j}), \qquad i = 1, \ldots, n, \quad j = 1, \ldots, k, \tag{54}$$

where $\theta_{i-1}^{-j} = (\theta_i^1, \ldots, \theta_i^{j-1}, \theta_{i-1}^{j+1}, \ldots \theta_{i-1}^k)$ denotes all components except θ^j at their *most recent* values. The complete algorithm is as follows:

> Initialize the algorithm:
>
> Choose a feasible initial state $\theta_0 = (\theta_0^1, \ldots, \theta_0^k)$.
>
> Do for draw $i = 1, \ldots, n$:
>
> Do for component $j = 1, \ldots, k$:
>
> Obtain θ_i^j from conditional target density $P(\theta^j|\theta_{i-1}^{-j})$.

Figure 11 illustrates how the Gibbs sampler works for two 2-dimensional target distributions involving correlation and bimodality. Clearly, as each time one of the two components (either θ^1 or θ^2) is fixed while the other component is sampled from its conditional distribution, a Gibbs path moves in orthogonal directions parallel to the coordinate axes. So, the horizontal position is updated given the current vertical position, and the vertical position is updated given the current horizontal position. The figure displays Gibbs paths after 10 iterations and after 1000 iterations, and it indicates that the orthogonal movement may cause the Gibbs sampler to break down. First, the two left-hand graphs demonstrate that high correlation results in a slowly moving Gibbs path, so that the Gibbs sampler might be stuck in a small region for quite a long time. This problem increases when the correlation between the two components becomes higher. Second, the two right-hand graphs demonstrate that if the target density has two modes located far away from each other, "mode hopping" does not occur often. This essentially induces the same problem as high correlation, that is, the Gibbs sampler might be stuck in a local region for a very long time. Consequently, an enormous number of draws might be needed to obtain a representative coverage of the entire target density. However, we note that in many cases a reparameterization of the sampling problem can be found to deal

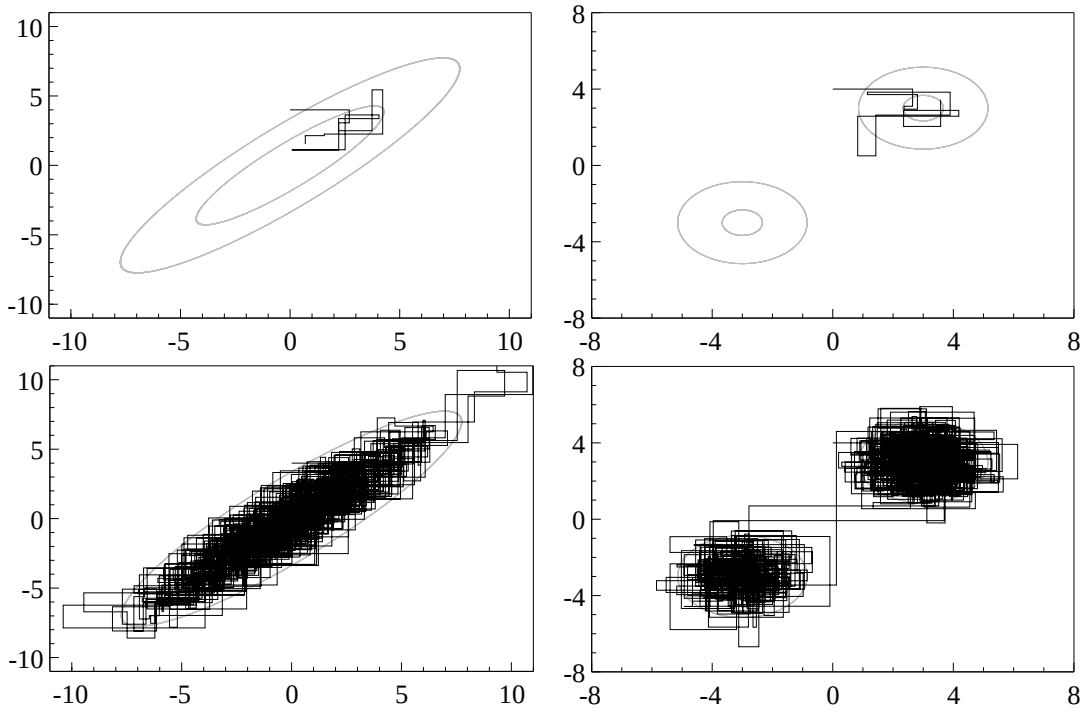

Figure 11: Illustration of the Gibbs sampler for a correlated target density (left) and a bimodal target density (right). The generated Gibbs paths are shown for 10 iterations (above) and 1000 iterations (below).

effectively with such high correlations, see for example Gilks and Roberts (1996).

Gibbs sampling is a special case of the Metropolis-Hastings algorithm

The Gibbs sampling algorithm is actually a special case of the MH algorithm. This can be understood as follows. First, it should be noted that an overall transition from state $\theta_{i-1} = (\theta_{i-1}^1, \ldots, \theta_{i-1}^k)$ to state $\theta_i = (\theta_i^1, \ldots, \theta_i^k)$ consists of k subsequent transitions from $(\theta_{i-1}^j, \theta_{i-1}^{-j})$ to $(\theta_i^j, \theta_{i-1}^{-j})$, where $j = 1, \ldots, k$. In each of these k transitions, one of the components of θ is updated given the most recent values of the other components. As the density for the j-th transition is given by

$$Q_j((\theta_{i-1}^j, \theta_{i-1}^{-j}), (\theta_i^j, \theta_{i-1}^{-j})) = P(\theta_i^j \,|\, \theta_{i-1}^{-j}), \tag{55}$$

where $j = 1, \ldots, k$, the density for the overall transition from state θ_{i-1} to state θ_i becomes

$$Q(\theta_{i-1}, \theta_i) = \prod_{j=1}^{k} P(\theta_i^j \,|\, \theta_{i-1}^{-j}). \tag{56}$$

By defining the candidate transition density of the MH algorithm by (55) and (56), the corresponding acceptance probabilities can be computed. The acceptance probability of the j-th transition from $(\theta_{i-1}^j, \theta_{i-1}^{-j})$ to $(\theta_i^j, \theta_{i-1}^{-j})$ is given by

$$\begin{aligned}
\alpha_j((\theta_{i-1}^j, \theta_{i-1}^{-j}), (\theta_i^j, \theta_{i-1}^{-j})) &= \min\left\{ \frac{P(\theta_i^j, \theta_{i-1}^{-j})\, Q_j((\theta_i^j, \theta_{i-1}^{-j}), (\theta_{i-1}^j, \theta_{i-1}^{-j}))}{P(\theta_{i-1}^j, \theta_{i-1}^{-j})\, Q_j((\theta_{i-1}^j, \theta_{i-1}^{-j}), (\theta_i^j, \theta_{i-1}^{-j}))}, 1 \right\} \\
&= \min\left\{ \frac{P(\theta_i^j, \theta_{i-1}^{-j})\, P(\theta_{i-1}^j |\, \theta_{i-1}^{-j})}{P(\theta_{i-1}^j, \theta_{i-1}^{-j})\, P(\theta_i^j |\, \theta_{i-1}^{-j})}, 1 \right\} \\
&= \min\left\{ \frac{P(\theta_i^j, \theta_{i-1}^{-j}) / P(\theta_i^j |\, \theta_{i-1}^{-j})}{P(\theta_{i-1}^j, \theta_{i-1}^{-j}) / P(\theta_{i-1}^j |\, \theta_{i-1}^{-j})}, 1 \right\} \\
&= \min\left\{ \frac{P(\theta_{i-1}^{-j})}{P(\theta_{i-1}^{-j})}, 1 \right\} = 1, \qquad (57)
\end{aligned}$$

where $i = 1, \ldots, k$. As all k transitions are accepted with probability 1, the overall transition from θ_{i-1} to θ_i is accepted with probability 1, that is,

$$\alpha(\theta_{i-1}, \theta_i) = 1. \qquad (58)$$

Thus, the Gibbs sampler is a special case of the MH algorithm in which rejections do not occur. This explains the enormous popularity of Gibbs sampling.

Griddy Gibbs sampling and the Metropolis-Hastings-within-Gibbs method

For application of the Gibbs sampling algorithm it is desirable but not necessary that all k conditional distributions can be directly sampled from. For example, if a "difficult" conditional distribution is one-dimensional, an approximating cumulative distribution function [CDF] can be constructed by building a density grid and using linear interpolation. Subsequently, the inversion method can be applied to the piecewise linear approximation. This is the griddy Gibbs sampling approach, proposed by Ritter and Tanner (1992). Alternatively, an MH step might be considered to sample from the (not necessarily univariate) "difficult" conditional distribution. This implies that each time a candidate transition is considered for the complicated component, which is either accepted or rejected in the Markov chain. Although this approach is just a special case of the MH algorithm, it is usually called the MH-within-Gibbs approach.

3.4.3 Gibbs sampling with data augmentation

For many models involving latent variables (such as the unobserved utilities in probit choice models), the parameters θ have a non-standard posterior distribution. Moreover, for such models, evaluation of the likelihood function and hence the posterior

density might be complicated and computationally intensive. This is for example the case in the conditional probit model of Hausman and Wise (1978), see also McCulloch and Rossi (1994). However, standard distributions would arise if the latent data z would be known. So, "observing" z would greatly facilitate the sampling procedure. The data augmentation algorithm of Tanner and Wong (1987) is a useful extension of the Gibbs sampler which is based on this notion. It extends the sampling space, as both the parameters θ and the latent data z are sampled. In the algorithm, z is drawn conditional on θ, and θ is drawn conditional on z. So, the latent data are imputed using the current parameter values, and subsequently the parameters are sampled *as if* the latent data are observed. By repeating this procedure many times, a Gibbs sequence is constructed involving both θ and z. Disregarding z, the process results in a Markov chain for the parameters θ converging to the posterior distribution. Data augmentation for the conditional probit model is discussed by Albert and Chib (1993), McCulloch and Rossi (1994) and McCulloch, Polson and Rossi (2000). Wei and Tanner (1990) and Chib (1992) consider data augmentation for the censored regression model.

Example: data augmentation in binary probit model for US recession indicator

In order to provide a simple illustration of the data augmentation approach, we apply it to a binary probit model with the purpose to explain and predict recessions in the United States using leading indicators. The data augmentation procedure follows Albert and Chib (1993). We define a recession indicator such that the economy is in a recession if the growth rate of U.S. real GDP is negative in at least the current period and either the preceding or next period. As leading indicators, we consider the growth rate of the Dow Jones Industrial Average, real consumption growth, the growth rate of the money stock M3, the term structure (the 10 year Treasury Bond yield minus the 1 year Treasury Bond yield), and the oil price. We use quarterly data running from the first quarter of 1968 to the fourth quarter of 2001. We find that the economy is in a recession for 12 of the 136 observed periods. A preliminary analysis indicates that a lag of two quarters between the leading indicators and the recession measure is appropriate. The binary probit model is given by

$$z_t = x_t'\beta + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1) \quad \text{i.i.d.}, \qquad t = 1, \ldots, T, \tag{59}$$

$$y_t = \begin{cases} 1 \text{ if } z_t > 0 \\ 0 \text{ if } z_t \leq 0 \end{cases}, \tag{60}$$

where y_t is the binary recession variable and x_t contains an intercept and the five leading indicators. The variable z_t is not observed. We consider the non-informative prior $p(\beta) \propto 1$ for the parameters β.

The conditional distributions for β and z are easily derived. First, if z_t would be

Table 2: Sampling results for the binary probit model with a U.S. recession indicator as the dependent variable. In the first three columns, the estimated mean, standard deviation, and autocorrelation (in the Gibbs sequence) are reported for each parameter. In the fourth and fifth column, the Maximum Likelihood parameter estimates and corresponding standard errors are shown.

	mean	s.d.	a.c.	ML	s.e.
intercept	0.335	0.762	0.869	0.183	0.719
Dow Jones	−0.144	0.048	0.946	−0.121	0.046
real consumption	−1.262	0.520	0.936	−1.070	0.491
money stock M3	−1.782	0.596	0.978	−1.454	0.551
term structure	−2.297	0.720	0.986	−1.860	0.669
oil price	0.062	0.031	0.918	0.053	0.030

observed, the model would reduce to a linear regression model with known variance σ^2. Using the symmetry with classical maximum likelihood (which holds in this case but does not hold in general), we obtain that

$$\beta|z, y \sim \mathcal{N}((X'X)^{-1}X'z, (X'X)^{-1}), \tag{61}$$

where x_t, $t = 1, \ldots, T$, are stacked in the matrix X. Second, given the parameters β and the observed data y, it holds that

$$\begin{cases} z_t|\beta, y \sim \mathcal{N}(x_t'\beta, 1)\, I\{z_t \leq 0\} & \text{if} \quad y_t = 0 \\ z_t|\beta, y \sim \mathcal{N}(x_t'\beta, 1)\, I\{z_t > 0\} & \text{if} \quad y_t = 1 \end{cases}, \tag{62}$$

for $t = 1, \ldots, T$. In sum, this gives the data augmentation algorithm:

Initialize the algorithm:

Choose initial β_0.

Do for draw $i = 1, \ldots, n$:

Sample z_t from $\begin{cases} \mathcal{N}(x_t'\beta_{i-1}, 1)\, I\{z_t \leq 0\} & \text{if} \quad y_t = 0 \\ \mathcal{N}(x_t'\beta_{i-1}, 1)\, I\{z_t > 0\} & \text{if} \quad y_t = 1 \end{cases}$.

Sample β_i from $\mathcal{N}((X'X)^{-1}X'z, (X'X)^{-1})$.

We take a burn-in period of 1000 draws and we consider 50000 effective draws with the zero vector as the initial location for the Markov chain. Alternatively, one might take the

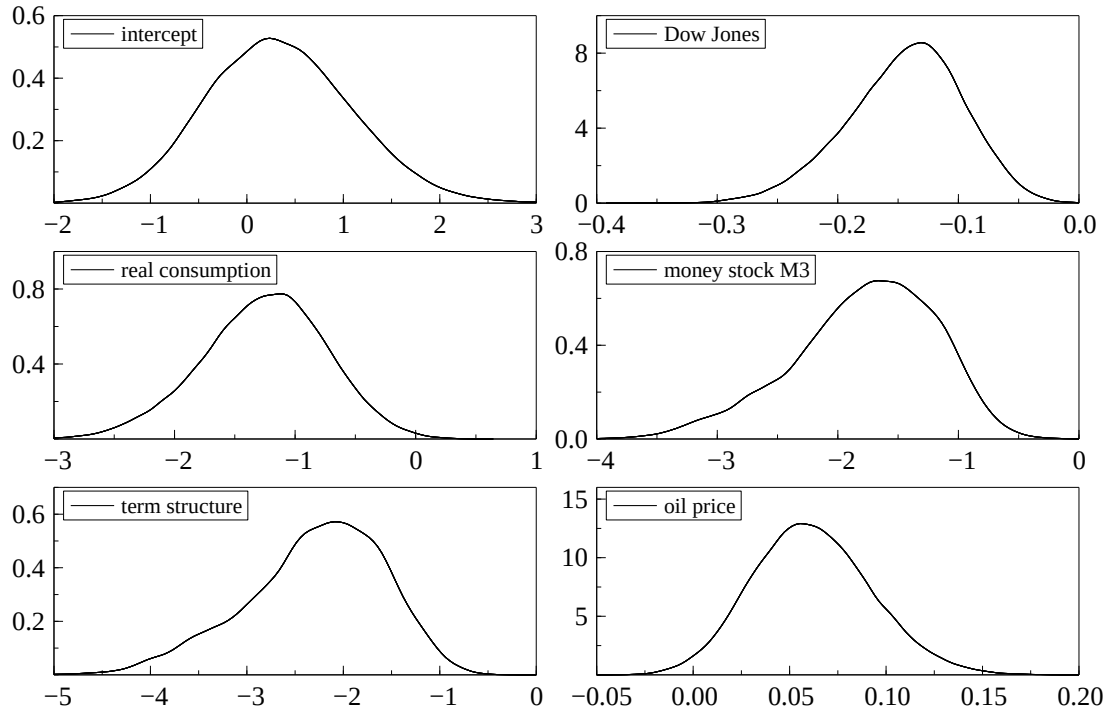

Figure 12: Marginal posterior densities for the binary probit model with a U.S. recession indicator as the dependent variable.

Maximum Likelihood [ML] parameter estimates as the initial values. In this illustration we use all draws after the burn-in, but other popular operationalizations are thinning (for example, only keeping every tenth draw) and independent runs (running many different chains from dispersed starting values and only keeping the final value), see for example Smith and Roberts (1993). The latter two approaches aim at reducing correlations at the expense of many (relatively uninformative) lost draws.

The obtained marginal densities for β are displayed in Figure 12. It is seen that the posterior densities have asymmetric tails and that, for all five leading indicators, nearly all posterior density mass is located such that the sign of the response parameter is as expected. Table 2 reports the estimated mean, standard deviation and autocorrelation for each parameter, together with the ML parameter estimates and the corresponding standard errors. All autocorrelations are larger than 0.85, and five of the six autocorrelations are larger than 0.90, indicating that the Markov chain only moves slowly through the parameter space. The table further shows that the estimated posterior standard deviations are larger than the corresponding ML standard errors and that the estimated posterior means are larger (in absolute values) than the corresponding ML parameter estimates. The explanation for this difference is that a Bayesian analysis allows for exact

inference when the number of observations is limited, whereas the ML results are based on asymptotic approximations.

3.4.4 Auxiliary variable Gibbs sampling

Auxiliary variable Gibbs sampling is a sampling approach developed by Damien et al. (1999), who extend the original work of Edwards and Sokal (1988). Similar to data augmentation (Tanner and Wong 1987), latent variables are incorporated in the sampling process in order to facilitate drawing from the full set of conditional distributions. However, contrary to data augmentation, the latent variables are not "missing data" from the model. Instead, the latent variables are introduced in an artificial way. The approach of Damien et al. (1999) might be interpreted as a reversion of the independence chain MH algorithm. We recall that the MH algorithm first draws a candidate state $\tilde{\theta}$, given the current state θ_{i-1}, and subsequently considers a uniform draw $u \in (0,1)$ to determine whether the candidate state is accepted. The sampling approach of Damien et al. (1999) turns this around, that is, first an auxiliary draw u from the uniform distribution is obtained and subsequently the state $\tilde{\theta}$ is sampled inside the acceptance region determined by u. The gain of this reversion is that the state $\tilde{\theta}$ is accepted by definition. However, the price to pay is that sampling inside the acceptance region amounts to drawing from some truncated distribution.

Auxiliary variable Gibbs sampling is based on a decomposition of the target density P. This decomposition is given by

$$P(\theta) \propto \pi(\theta) \prod_{j=1}^{\tilde{n}} l_j(\theta), \tag{63}$$

where π is a kernel of some density from which draws are easily obtained, and l_j $(j = 1, \ldots, \tilde{n})$ are non-negative functions. For practical implementation, l_j $(j = 1, \ldots, \tilde{n})$ should be invertible for univariate θ, i.e. invertible for any univariate component of θ when keeping all the other components constant. We note that decomposition (63) nests the prior-likelihood decomposition from Bayes' theorem, given by

$$p(\theta|y) \propto p(\theta) \prod_{j=1}^{N} p(y_j|\theta), \tag{64}$$

where $p(y_j|\theta)$ is the contribution to the likelihood by the j-th observation. Note that such a decomposition is possible, as long as the observations y_j $(j = 1, \ldots, N)$ are

independent (conditional on θ and exogenous/predetermined variables). In the approach of Damien et al. (1999), a set $U = (U^1, \ldots, U^{\tilde{n}})$ of uniform auxiliary variables is introduced such that

$$U^j|(\theta = \tilde{\theta}) \sim U(0, l_j(\tilde{\theta})), \qquad j = 1, \ldots, \tilde{n}, \tag{65}$$

resulting in the joint density

$$\begin{aligned} P(\theta, u) &= P(\theta)P(u|\theta) \\ &\propto \pi(\theta) \prod_{j=1}^{\tilde{n}} l_j(\theta) \prod_{j=1}^{\tilde{n}} \frac{I\{\, 0 < u^j < l_j(\theta)\}}{l_j(\theta)} \\ &= \pi(\theta) \prod_{j=1}^{\tilde{n}} I\{\, 0 < u^j < l_j(\theta)\} \end{aligned} \tag{66}$$

and the conditional density

$$P(\theta|u) \propto \pi(\theta)\, I\{\, l_j(\theta) > u^j, j = 1, \ldots, \tilde{n}\}. \tag{67}$$

Note that the marginal density of θ remains formula (63). Similar to data augmentation, the sampling space is extended, as both θ and U are sampled from their conditional distributions. We note that an iteration of this Gibbs procedure requires drawing from $\tilde{n}$ uniform distributions, and drawing from some truncated version of an "easy" distribution (by the assumption made on π). The complete algorithm is as follows:

Initialize the algorithm:

Choose a feasible initial state θ_0.

Do for draw $i = 1, \ldots, n$:

Obtain u_i^j from uniform distribution $U(0, l_j(\theta_{i-1}))$, $j = 1, \ldots, \tilde{n}$.

Obtain θ_i from $\pi(\theta)\, I\{\, l_j(\theta) > u_i^j, j = 1, \ldots, \tilde{n}\}$.

Collect $\theta_1, \ldots, \theta_n$.

Now, by setting $\tilde{n} = 1$ and suppressing the index j, the link with the independence chain MH algorithm becomes clear. First, it should be noted that $u_i = u\, l(\theta_{i-1})$, where u is some draw from $U(0, 1)$. Next, it should be noted that θ_i is such that $l(\theta_i) > u_i$. Consequently, θ_i satisfies the condition $u < \frac{l(\theta_i)}{l(\theta_{i-1})}$, which is equivalent to

$$u < \min\left\{ \frac{l(\theta_i)}{l(\theta_{i-1})}, 1 \right\}, \tag{68}$$

as $u \in (0,1)$. This shows that auxiliary variable Gibbs sampling is essentially a reversion of the independence chain MH algorithm with target density $P(\theta) \propto \pi(\theta)l(\theta)$ and candidate density $Q(\theta) \propto \pi(\theta)$.

Example: auxiliary variable Gibbs sampling in binary logit model for US recession indicator

Damien et al. (1999) demonstrate that their approach is useful for non-conjugate and hierarchical models by working out several examples. As an illustration of the method, we consider the binary logit model, given by

$$z_t = x_t'\beta + \varepsilon_t, \qquad \varepsilon_t \sim \text{Logistic} \quad \text{i.i.d.}, \qquad t = 1, \ldots, T, \tag{69}$$

$$y_t = \begin{cases} 1 \text{ if } z_t > 0 \\ 0 \text{ if } z_t \leq 0 \end{cases}, \tag{70}$$

where y_t is the U.S. recession variable from the binary probit example, and x_t contains an intercept and the five leading indicators (lag of two quarterly periods: growth rate of the Dow Jones Industrial Average, real consumption growth, growth of money stock M3, the term structure, and the oil price). We note that Dellaportas and Smith (1993) put forward an alternative procedure to sample the parameters, involving an adaptive rejection algorithm. The binary logit model has likelihood function

$$\begin{aligned} p(y|\beta) &= \prod_{t=1}^{T} \left(\frac{\exp(x_t'\beta)}{1+\exp(x_t'\beta)} \right)^{y_t} \left(\frac{1}{1+\exp(x_t'\beta)} \right)^{1-y_t} \\ &= \prod_{t=1}^{T} \frac{\exp(y_t x_t'\beta)}{1+\exp(x_t'\beta)} \,. \end{aligned} \tag{71}$$

We consider a (non-conjugate) normal prior for β with mean μ and covariance matrix Σ, so that

$$p(\beta) \propto \exp\!\left(-\frac{1}{2} (\beta-\mu)'\Sigma^{-1}(\beta-\mu) \right). \tag{72}$$

It follows from (71) and (72) that the posterior is given by

$$p(\beta|y) \propto \exp\!\left(-\frac{1}{2} (\beta-\mu)'\Sigma^{-1}(\beta-\mu) \right) \prod_{t=1}^{T} \frac{\exp(y_t x_t'\beta)}{1+\exp(x_t'\beta)} = \pi(\beta) \prod_{t=1}^{T} l_t(\beta). \tag{73}$$

The decomposition in (73) provides the basis for an auxiliary variable Gibbs algorithm. For the binary logit model, the truncation condition $l_t(\beta) > u_i^t$ amounts to

$$\begin{cases} x_t'\beta < \ln(1-u_i^t) - \ln(u_i^t) & \text{if} \quad y_t = 0 \\ x_t'\beta > \ln(u_i^t) - \ln(1-u_i^t) & \text{if} \quad y_t = 1 \end{cases}, \tag{74}$$

for $t = 1, \ldots, T$. An algorithm to sample the parameters β would be:

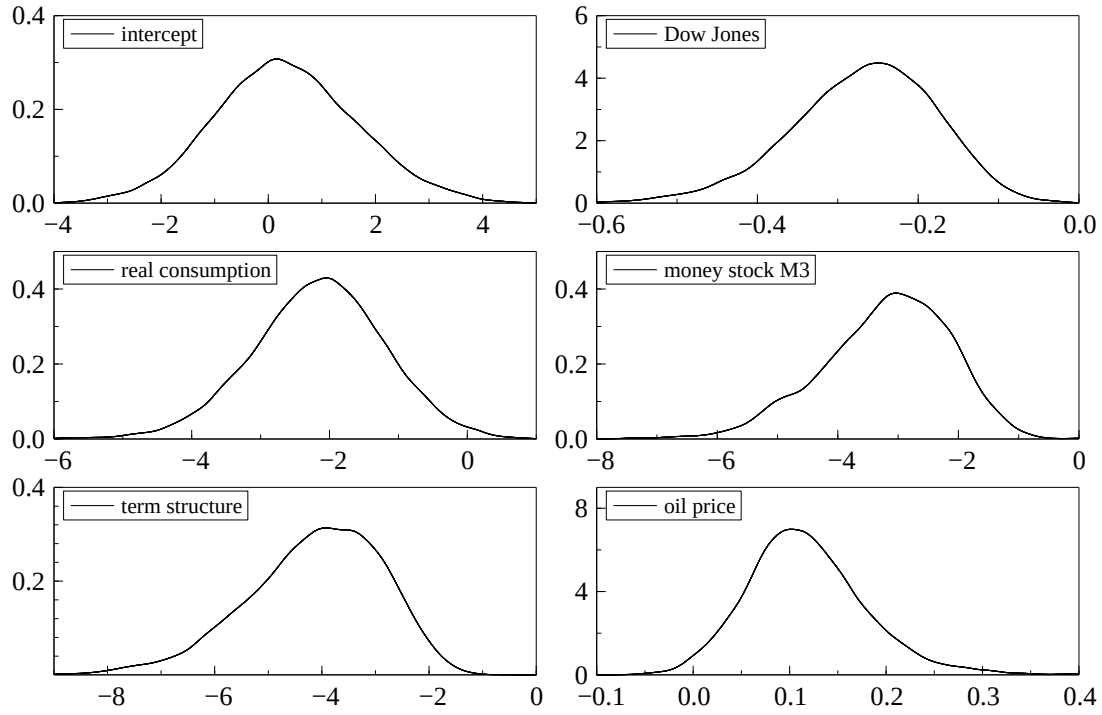

Figure 13: Marginal posterior densities for the binary logit model with a U.S. recession indicator as the dependent variable.

Initialize the algorithm:

Choose a feasible initial state β_0.

Do for draw $i = 1, \ldots, n$:

Obtain u_i^t from $U\Big(0, \frac{\exp(y_t x_t' \beta_{i-1})}{1+\exp(x_t' \beta_{i-1})}\Big)$, $t = 1, \ldots, T$.

Obtain β_i from $\mathcal{N}(\mu, \Sigma)\, I\{$ (74) holds for $t = 1, \ldots, T \}$.

Collect $\beta_1, \ldots, \beta_n$.

Unfortunately, a drawback of the algorithm above is that β has to be sampled from a *multivariate* truncated distribution for which rejection-based sampling might be very inefficient. A more efficient algorithm can be obtained by breaking up the parameter vector β and drawing its components separately from truncated *univariate* normal distributions using the inversion method. In order to determine the conditional distributions of the components β^j conditional on the remaining components β^{-j} and the auxiliary

Table 3: Sampling results for the binary logit model with a U.S. recession indicator as the dependent variable. In the first three columns, the estimated mean, standard deviation, and autocorrelation are reported for each parameter. In the fourth and fifth column, the Maximum Likelihood parameter estimates and corresponding standard errors are shown.

	mean	s.d.	a.c.	ML	s.e.
intercept	0.324	1.314	0.985	0.216	1.236
Dow Jones	−0.268	0.091	0.923	−0.216	0.081
real consumption	−2.159	0.966	0.935	−1.819	0.844
money stock M3	−3.190	1.044	0.991	−2.544	0.969
term structure	−4.124	1.254	0.972	−3.259	1.146
oil price	0.117	0.062	0.980	0.095	0.052

variables U, it should be noted that if

$$\begin{pmatrix} \beta^j \\ \beta^{-j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_j \\ \mu_{-j} \end{pmatrix}, \begin{pmatrix} \Sigma_{j,j} & \Sigma_{j,-j} \\ \Sigma_{-j,j} & \Sigma_{-j,-j} \end{pmatrix} \right), \tag{75}$$

then

$$\beta^j \Big| \beta^{-j} \sim \mathcal{N}\Big(\mu_j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}(\beta^{-j} - \mu_{-j}), \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} \Big). \tag{76}$$

For ease of exposition and since the extension to the general case is straightforward, we assume that Σ is a diagonal matrix, so that (76) boils down to

$$\beta^j \Big| \beta^{-j} \sim \mathcal{N}\Big(\mu_j, \Sigma_{j,j} \Big). \tag{77}$$

Using this result and rewriting the truncation condition (74) in terms of β^j, we obtain the final algorithm:

Initialize the algorithm:

Choose a feasible initial state β_0.

Do for draw $i = 1, \ldots, n$:

Obtain u_i^t from $U\Big(0, \frac{\exp(y_t x_t' \beta_{i-1})}{1+\exp(x_t' \beta_{i-1})}\Big)$, $t = 1, \ldots, T$.

Do for component $j = 1, \ldots, k$:

Obtain β_i^j from

$$\mathcal{N}\Big(\mu_j, \Sigma_{j,j}\Big)\, I\left\{\beta^j < \min_{t:y_t=0, x_{t,j}>0}\left\{\frac{\ln(1-u_i^t) - \ln(u_i^t) - \sum_{l\neq j} x_{t,l}\beta^l}{x_{t,j}}\right\}\right\}$$
$$I\left\{\beta^j > \max_{t:y_t=0, x_{t,j}<0}\left\{\frac{\ln(1-u_i^t) - \ln(u_i^t) - \sum_{l\neq j} x_{t,l}\beta^l}{x_{t,j}}\right\}\right\}$$
$$I\left\{\beta^j > \max_{t:y_t=1, x_{t,j}>0}\left\{\frac{\ln(u_i^t) - \ln(1-u_i^t) - \sum_{l\neq j} x_{t,l}\beta^l}{x_{t,j}}\right\}\right\}$$
$$I\left\{\beta^j < \min_{t:y_t=1, x_{t,j}<0}\left\{\frac{\ln(u_i^t) - \ln(1-u_i^t) - \sum_{l\neq j} x_{t,l}\beta^l}{x_{t,j}}\right\}\right\}.$$

Collect $\beta_1, \ldots, \beta_n$.

We apply the algorithm described above to the data set from the binary probit example. Again, we take a burn-in period of 1000 draws and we consider 50000 effective draws with the zero vector as the initial location for the Markov chain. Furthermore, we consider a (fairly non-informative) normal prior $p(\beta)$ with the density mass located around the origin and a covariance matrix which is 100 times the identity matrix. The estimated marginal densities for β are displayed in Figure 13. As for the binary probit example, we observe that the posterior densities have asymmetric tails and that, for all five leading indicators, nearly all posterior density mass is located such that the sign of the response parameter is as expected. Table 3 reports the estimated mean, standard deviation and autocorrelation for each parameter, together with the ML parameter estimates and the corresponding standard errors. As for the binary probit example, it can be seen that the Markov chain only moves slowly through the parameter space, and that the posterior densities are more spread out, away from zero, than the ML results would suggest.

3.5 Some recently developed simulation methods

The simulation methods that we discussed in the previous subsections are popular simulation algorithms that are applicable to many posterior densities, as long as these posterior densities are reasonably well-behaved. In this subsection we discuss two recently developed simulation methods that are flexible and robust in the sense that these methods also yield reliable results in the case of a posterior with highly non-elliptical shapes, e.g. multi-modality, extreme skewness, and/or heavy tails.

3.5.1 Adaptive Radial-based Direction Sampling

Adaptive radial-based direction sampling [ARDS] methods, due to Bauwens et al. (2004), constitute a class of Monte Carlo integration methods that involve a transformation from the usual Carthesian coordinates to radial coordinates. The ARDS algorithms can be especially useful for Bayesian inference in models with non-elliptical, possibly multi-modal target distributions. A key step is a radial-based transformation to directions and distances. After the transformation a Metropolis-Hastings or importance sampling method is applied to evaluate generated directions. Next, given these directions, distances are generated from the exact target distribution. An adaptive procedure is applied to update the initial location and covariance matrix in order to sample directions in an efficient way.

The main idea is that sampling from an ill-behaved distribution can be facilitated by slicing this target distribution in a clever way, that is, by drawing along one-dimensional lines. Suitable directions, defining these lines, are obtained through either an MH step or an importance sampling step. The MH variant is called Adaptive Radial-Based Metropolis-Hastings Sampling [ARMHS], and the importance sampling variant is called Adaptive Radial-Based Importance Sampling [ARIS].

The ARDS algorithms have three major advantages. First, the algorithms do not require much information on the shape of the target density: only approximate guesses of location and scale are required as initial values. Second, the ARDS algorithms are flexible and robust: they can handle highly non-elliptical target densities such as multi-modal, extremely skew or heavy-tailed target densities. Third, the ARDS algorithms can handle linear inequality conditions on the parameter space without any additional complications for the implementation.

The ARDS methods are inspired by other algorithms in which directions are generated in order to facilitate the integration or simulation process. The ARDS algorithms extend earlier methods like the algorithm of Box and Muller (1958) for generating normal variates, the adaptive direction sampling [ADS] algorithms due to Gilks et al. (1994), the mixed integration method by Van Dijk et al. (1985), and the spherical integration method by Monahan and Genz (1997).

The radial transformation

Since the radial transformation is the key step of the ARDS algorithms, we start by describing the transformation from Cartesian coordinates to radial coordinates. The original m-dimensional parameter space is transformed into a $(m-1)$-dimensional space of directions and a unidimensional complementary space of distances. In our notation, $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_m)$ denotes the Cartesian coordinates of a point, and (ρ, η) denotes the corresponding radial coordinates. Here $\eta = (\eta_1, \ldots, \eta_{m-1})$ indicates the direction of the point relative to the origin, and ρ is related to the Euclidean distance. The m-dimensional transformation from $(\tilde{\theta}_1, \ldots, \tilde{\theta}_m) \in \mathbb{R}^m$ to $(\rho, \eta) = (\rho, \eta_1, \ldots, \eta_{m-1}) \in \mathbb{R} \times \{\eta \in \mathbb{R}^{m-1} : \eta'\eta < 1\}$ is given by

$$\rho = \text{sgn}(\tilde{\theta}_m)\sqrt{\tilde{\theta}'\tilde{\theta}}, \tag{78}$$

$$\eta_j = \frac{\tilde{\theta}_j}{\rho}, \qquad j = 1, \ldots, m-1, \tag{79}$$

with inverse transformation

$$\tilde{\theta}_j = \rho\,\eta_j, \qquad j = 1, \ldots m-1, \tag{80}$$

$$\tilde{\theta}_m = \rho\sqrt{1-\eta'\eta}\,. \tag{81}$$

By defining $\tilde{\theta}^* = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{m-1})$, the Jacobian of the transformation is

$$\begin{aligned} J_{\tilde{\theta}}(\rho, \eta) &= \det\begin{pmatrix} \frac{\partial \tilde{\theta}^*(\rho,\eta)}{\partial \eta'} & \frac{\partial \tilde{\theta}^*(\rho,\eta)}{\partial \rho} \\ \frac{\partial \tilde{\theta}_m(\rho,\eta)}{\partial \eta'} & \frac{\partial \tilde{\theta}_m(\rho,\eta)}{\partial \rho} \end{pmatrix} = \det\begin{pmatrix} \rho\, I_{m-1} & \eta \\ -\frac{\rho\,\eta'}{\sqrt{1-\eta'\eta}} & \sqrt{1-\eta'\eta} \end{pmatrix} \\ &= \frac{\rho^{m-1}}{\sqrt{1-\eta'\eta}} = J_{\tilde{\theta}}(\rho) J_{\tilde{\theta}}(\eta), \end{aligned} \tag{82}$$

where

$$J_{\tilde{\theta}}(\rho) = \rho^{m-1}, \tag{83}$$

$$J_{\tilde{\theta}}(\eta) = (1-\eta'\eta)^{-1/2}. \tag{84}$$

The form of this Jacobian has some important implications which are used in ARDS. It is shown by Bauwens et al. (2004) that the implementation of the ARDS algorithms is only based on the Jacobian factor $J_{\tilde{\theta}}(\rho)$, and does not depend on $J_{\tilde{\theta}}(\eta)$.

Basically, $\tilde{\theta}$ is transformed to $m-1$ Cartesian coordinates on the unit circle and a stretching factor ρ. This is illustrated in Figure 14 for $m = 2$ dimensions. Here we note that the sign of ρ determines whether $\tilde{\theta}$ is located above or below the $\tilde{\theta}_1$ axis.

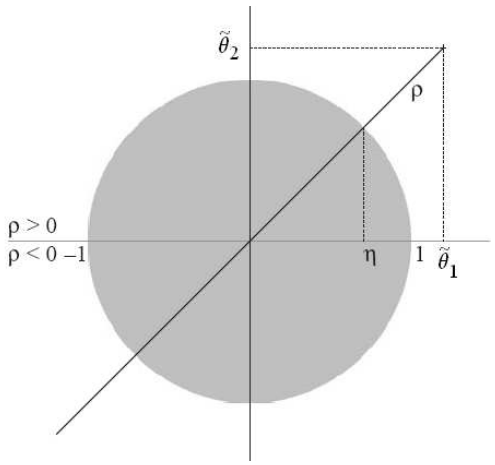

Figure 14: The relationship between Cartesian coordinates and radial coordinates in the two-dimensional case

Radial-based Metropolis-Hastings sampling

We now define the radial-based Metropolis-Hastings algorithm [RMHS], which is based on a candidate density that is taken to be multivariate normal with parameters μ and Σ. However, Bauwens et al. (2004) show that actually any elliptically contoured candidate distribution can be considered without affecting the sampling results. After defining RMHS, we will define the adaptive RMHS algorithm [ARMHS], where μ and Σ are iteratively updated using the sample of draws from a previous round of the RMHS algorithm.

RMHS is based on an independence chain MH algorithm. It uses draws from a $\mathcal{N}(\mu, \Sigma)$ candidate where hopefully μ and Σ provide good approximations to the unknown mean and covariance matrix of the target distribution. In contrast with the MH algorithm, the draws are not used for construction of a Markov chain in the original parameter space. Instead, a composite transformation is made. For expository purpose we treat this transformation explicitly in two steps. The first step is a location-scale transformation of a realization θ to a realization $\tilde{\theta}$. This transformation aims at standardizing the candidate density with respect to the location, scale, and correlations of the target (posterior) density, denoted by $p(\theta)$. It is defined by the affine transformation

$$\tilde{\theta} = \tilde{\theta}(\theta|\mu, \Sigma) = \Sigma^{-1/2}(\theta - \mu) \tag{85}$$

with inverse transformation

$$\theta = \theta(\tilde{\theta}|\mu, \Sigma) = \mu + \Sigma^{1/2}\tilde{\theta} \tag{86}$$

and Jacobian

$$J_\theta(\tilde{\theta}) = \det(\Sigma^{1/2}). \tag{87}$$

The second step is the radial transformation, which is defined by (78) and (79), with inverse transformation given by (80) and (81), and Jacobian (82).

Combining the two transformations, one obtains the composite transformation

$$\begin{pmatrix} \rho \\ \eta \end{pmatrix} = \begin{pmatrix} \rho(\tilde{\theta}(\theta|\mu,\Sigma)) \\ \eta(\tilde{\theta}(\theta|\mu,\Sigma) \end{pmatrix} \tag{88}$$

with inverse transformation

$$\theta = \theta(\tilde{\theta}(\rho,\eta)|\mu,\Sigma) \tag{89}$$

and Jacobian

$$J_\theta(\rho,\eta) = J_{\tilde{\theta}}(\rho,\eta)J_\theta(\tilde{\theta}) = J_{\tilde{\theta}}(\rho)J_{\tilde{\theta}}(\eta)\det(\Sigma^{1/2}). \tag{90}$$

Applying the two transformations to a candidate realization θ_i from $\mathcal{N}(\mu,\Sigma)$ yields a distance ρ_i^* and a vector of directions η_i^*. Ignoring the distance, the candidate direction is either accepted or rejected in an MH step: the direction becomes

$$\eta_i = \begin{cases} \eta_i^* & \text{with probability } \alpha(\eta_{i-1},\eta_i^*) \\ \eta_{i-1} & \text{with probability 1-}\alpha(\eta_{i-1},\eta_i^*) \end{cases} \tag{91}$$

for some acceptance probability $\alpha(\eta_{i-1},\eta_i^*)$ that will be given below. An iteration of RMHS is completed by drawing from the target distribution on the line defined by the direction η_i. This can be done as follows. First, one draws a distance ρ_i from the transformed target distribution for given direction η_i using the numerical inverse transformation method. Next, η_i and ρ_i are transformed to the original space by inverting the radial transformation and the location-scale transformation. Therefore, the steps of one iteration of RMHS are as follows:

1. *Obtain candidate*: Get realization θ_i^* from $N(\mu,\Sigma)$.
2. *Standardization*: Transform θ_i^* to $\tilde{\theta}_i^* = \Sigma^{-1/2}(\theta_i^* - \mu)$.
3. *Radialization*: Transform $\tilde{\theta}_i^*$ to (ρ_i^*,η_i^*) using (78) and (79).
4. *MH step*: $\eta_i = \begin{cases} \eta_i^* & \text{with probability } \alpha(\eta_{i-1},\eta_i^*) \\ \eta_{i-1} & \text{with probability } 1-\alpha(\eta_{i-1},\eta_i^*) \end{cases}$.
5. *Inversion step*: Sample ρ_i from its conditional density $p(\rho|\eta_i)$ by applying the inversion method to the density grid obtained in step 4.
6. *Deradialization*: Transform (ρ_i,η_i) to $\tilde{\theta}_i$ using (80) and (81).
7. *Destandardization*: Transform $\tilde{\theta}_i$ to $\theta_i = \mu + \Sigma^{1/2}\tilde{\theta}_i$.

Note that step 4 of an RMHS iteration requires the acceptance probability $\alpha(\eta_{i-1}, \eta_i^*)$, while step 5 requires the distribution of the distance ρ_i conditional on the direction η_i. Bauwens et al. (2004) show that $\alpha(\eta_{i-1}, \eta_i^*)$ is given by

$$\alpha(\eta_{i-1}, \eta_i^*) = \min \left\{ \frac{I(\eta_i^*)}{I(\eta_{i-1})}, 1 \right\} \tag{92}$$

where

$$I(\eta) = \int_{-\infty}^{\infty} \kappa(\rho|\eta) d\rho, \tag{93}$$

and where $\kappa(\rho|\eta)$ is a kernel of the conditional density $p(\rho|\eta)$ of step 5, defined by

$$p(\rho|\eta) \propto \kappa(\rho|\eta) = P(\theta(\rho, \eta|\mu, \Sigma))|J_y(\rho)|, \tag{94}$$

where $P(\theta)$ is (a kernel of) the target density. Note that in order to obtain the acceptance probability $\alpha(\eta_{i-1}, \eta_i^*)$, the *one-dimensional* integral $I(\eta)$ defined by (93) is computed by a deterministic integration rule. Since the density of ρ conditional on η is proportional to the integrand of $I(\eta)$, evaluations of the integrand, gathered during the deterministic integration phase, can be used in order to construct a grid for $\kappa(\rho|\eta)$. Using the numerical inverse transformation method, sampling the distance ρ conditional on the direction η – step 5 of RMHS – is straightforward. We can further reduce the computational effort by generating several draws of ρ for each draw of η, thereby capitalizing on the construction of the grid for $\kappa(\rho|\eta)$.

Further note that the integral $I(\eta)$ has infinite integration bounds. However, in practice we use finite integration bounds for its numerical evaluation. In order to obtain bounds for the distance ρ we impose minimum and maximum values for each element of θ in the original space. It is often possible to find sensible bounds by either theory and/or common sense. Bauwens et al. (2004) show that, as these bounds on the elements of θ can be considered as linear restrictions, additional linear restrictions do not cause any additional complications for the algorithm.

Convergence of radial-based Metropolis-Hastings sampling

RMHS is a combination of a Metropolis-Hastings sampler for the directions and direct sampling of the distance ρ. Hence, the transition kernel of RMHS is the transition kernel of the MH step, and we can rely on known convergence results for the MH algorithm, see e.g. Smith and Roberts (1993). As long as the covariance matrix Σ is non-singular, these convergence results are preserved after applying the location-scale transformation. Moreover, they are also preserved after applying the radial transformation given that this transformation does not induce singularities, which is the case if $\eta \neq \pm 1$ and $\rho \neq 0$. As these singularities have Lebesgue measure zero, the radial transformation does not affect convergence properties. So, the sampled RMHS chain converges in distribution to the target distribution. Nevertheless,

in practice convergence after a finite number of draws should obviously be monitored by the usual tools, see e.g. Van Dijk and Kloek (1980) and Oh and Berger (1992). But at least, since only the direction η, and not the distance ρ, is generated from a possibly 'wrong' candidate distribution, the risk of collecting a 'wrong' sample is substantially reduced. In other words, ARMHS is quite robust, as the distance ρ conditional on the direction η immediately comes from the target distribution, that is, sampling on a given line mimics exactly the target density.

Adaptive radial-based Metropolis-Hastings sampling

For implementation of RMHS, the mean μ and the covariance matrix Σ of the normal candidate distribution have to be specified. Good enough initial approximations are usually the posterior mode and minus the inverse Hessian of the log posterior evaluated at the mode. Heuristically, convergence of RMHS should improve if μ and Σ are taken closer to the target mean and covariance matrix. Adaptive radial-based Metropolis-Hastings sampling [ARMHS] considers a sequential adaptive approach. Given a generated sample $\theta_1, \theta_2, \ldots, \theta_n$ from a previous run of the algorithm, μ and Σ are replaced by the Monte Carlo estimates of the posterior mean and covariance matrix, which are given by:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\theta_i \tag{95}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\theta_i - \hat{\mu})(\theta_i - \hat{\mu})' \tag{96}$$

Using these estimates, one can proceed with a new sampling round. This process can be repeated any number of times. In order to monitor convergence over sampling rounds, we find the Mahalanobis distance particularly useful. It is defined as $\text{Mah}_j = (\hat{\mu}(j) - \hat{\mu}(j-1))'[\hat{\Sigma}^{(j)}]^{-1}(\hat{\mu}(j) - \hat{\mu}(j-1))$, where j indicates the sampling round. The Mahalanobis distance measures the extent to which the estimated posterior mean changes between successive sampling rounds, while accounting for parameter uncertainty and the underlying correlation structure.

Adaptive radial-based importance sampling

Radial-based importance sampling (RIS) replaces the MH step of RMHS for the direction η by an importance sampling step. So, step 4 of an RMHS iteration changes. In RIS, every sampled direction η_i is kept, a distance ρ_i is sampled conditional on it, and the resulting radial coordinates are transformed to a draw θ_i in the

original space, which is weighted according to the appropriate importance weight

$$w(\eta_i) = \frac{p(\eta_i)}{q(\eta_i)} \propto I(\eta_i), \tag{97}$$

where $I(\eta)$ is defined by (93). As RIS can be interpreted as a special case of importance sampling, convergence properties of RIS follow directly from those for the latter method. Important diagnostics are given by the distribution of the weights $w(\eta_i)$. For details, see Geweke (1989).

In a similar fashion to ARMHS, the parameters μ and Σ of the location-scale transformation can be updated by replacing them by their Monte Carlo estimates. We will refer to this adaptive extension of RIS as adaptive RIS (ARIS).

Example: ARDS in two-regime mixture model for the US GNP growth rate

In order to illustrate the advantages of the ARDS methods, we investigate a mixture model for the analysis of economic growth in the USA, which is also considered by Bauwens et al. (2004). Bauwens et al. (2004) compare the performance of the ARDS methods with the (independence chain) Metropolis- Hastings algorithm and importance sampling with a Student-t candidate distribution (with 5 degrees of freedom). They compare estimation results after a given computing time with the 'true' results - estimation results after many more draws - and inspect the graphs of estimated marginal densities resulting from different sampling methods. Here we take another approach to investigate the accuracy of different simulation methods given the same computing time. For each simulation method, we repeat the simulation process ten times with different random seeds, after which we compute the standard deviations of the ten estimates of the posterior means. We note that in these empirical examples the mixture process refers to the data space. However, such mixture processes may give rise to bimodality and skewness in the parameter space.

In models for the growth rate of the gross national product, great advances have been made by allowing for separate regimes in periods of recession and expansion. However, these models may give rise to difficulties with respect to convergence of sampling methods due to multiple modes. Here we consider a mixture model with two AR(1) regimes for real GNP growth:

$$\begin{aligned} y_t &= \begin{cases} \beta_{11} + \beta_{12} y_{t-1} + \varepsilon_t & \text{with probability } p, \\ \beta_{21} + \beta_{22} y_{t-1} + \varepsilon_t & \text{with probability } 1-p, \end{cases} \\ \varepsilon_t &\sim N(0, \sigma^2), \end{aligned} \tag{98}$$

where y_t denotes the quarterly growth rate. The data (source: Economagic) consist of observations from the first quarter of 1959 to the last quarter of 2001. Note that we have

a 6-dimensional vector $\theta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \sigma, p)'$. The priors for β_{11}, β_{12}, β_{21}, β_{22} and p are chosen uniform, and the prior for σ is taken proportional to $1/\sigma$, which amounts to specifying a uniform prior for $\log(\sigma)$. So, we have $p(\theta) = p(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \sigma, p) = 1/\sigma$. For identification, it is imposed that $\beta_{11} < \beta_{21}$. In order to numerically evaluate the integral $I(\eta)$ in (93), parameter bounds are specified; see Table 4.

We choose the same sampling setup as Bauwens et al. (2004). In our adaptive approach, additional sampling rounds are considered as long as the Mahalanobis distance is larger than 0.02. However, we allow for at most 8 rounds. In any round, ARMHS and ARIS draw 5000 directions and 5 distances per direction, resulting in a sample of size 25 000. In order to make the computing times comparable, the MH and importance sampling algorithms are allowed to collect a larger sample of size 250 000. The scale of the initial candidate distribution is taken sufficiently large, so that MH and importance sampling can initially cover the whole density mass.

Sampling results are given by Table 4, which also gives the 'large sample' values (computed from 250 000 ARMHS draws). Notice that each standard deviation of the 10 estimates of the posterior means is smaller for the ARDS methods than for the MH and IS approach with a Student t candidate, where the ARIS somewhat outperforms the ARMHS method. Even though 10 times less draws have been generated, the 'quality' of those draws is much higher. This can be seen from the acceptance rate that is much higher for ARMHS than for MH, and from the weight of the 5% most influential points that is much smaller for ARIS than for IS. It should be noted that it is also possible to apply the data augmentation algorithm to this model. However, this approach requires more 'inputs' than the ARDS methods. For the data augmentation method, the conditional posterior distribution of each parameter has to be derived, whereas the ARDS methods only require a kernel of the posterior density (and approximate guesses of the location and scale).

In this model we define the latent variables Z_t $(t = 1, \ldots, T)$ as:

$$Z_t = \begin{cases} 0 & \text{if period } t \text{ is a period of regime 1} \\ 1 & \text{if period } t \text{ is a period of regime 2} \end{cases} \qquad t = 1, 2, \ldots, T. \tag{99}$$

Conditionally on the values of the parameters, the latent variables Z_t $(t = 1, \ldots, T)$ have a Bernoulli distribution. Conditionally on the latent variables Z (and each other), (β_{11}, β_{12}), and β_{21}, β_{22} are normally distributed, while σ^2 and p have an inverted gamma and a beta distribution, respectively. The results of the data augmentation method are given by Table 5. The number of draws has been chosen in order to make the computing time comparable with the ARIS method. Notice that each standard deviation of the 10 estimates of the posterior means is smaller for the data augmentation than for the ARDS methods. Estimates of the marginal densities are given by Figure 15. Note the bimodality in the marginal posterior of p and the skewness for the β parameters. These shapes can be explained by inspecting the scatter plots of parameter draws. Figure 16 shows draws of (p, β_{11}) and (p, β_{21}). If $p \to 0$ $(p \to 1)$, then β_{11} and β_{12} (β_{21} and β_{22}) become unidentified, so that a wide range of values is possible for these parameters.

Table 4: Sampling results for the two-regime mixture model for US real GNP growth

	Bounds ARDS		ARMHS		ARIS		MH		IS		Large sample	
	min.	max.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
β_{11}	-4.00	4.00	0.11	0.64	0.10	0.59	-0.14	0.88	0.01	0.72	0.07	0.70
(s.d. 10x)			(0.06)		(0.04)		(0.12)		(0.09)			
β_{12}	-1.00	1.00	0.45	0.24	0.45	0.25	0.42	0.28	0.40	0.28	0.41	0.27
(s.d. 10x)			(0.03)		(0.03)		(0.04)		(0.04)			
β_{21}	-4.00	4.00	1.32	0.74	1.27	0.78	1.22	0.83	1.28	0.85	1.30	0.79
(s.d. 10x)			(0.07)		(0.05)		(0.11)		(0.10)			
β_{22}	-1.00	1.00	-0.07	0.39	-0.02	0.38	0.05	0.39	0.01	0.40	-0.04	0.41
(s.d. 10x)			(0.03)		(0.02)		(0.04)		(0.04)			
σ	0.00	2.00	0.82	0.05	0.82	0.06	0.82	0.06	0.82	0.06	0.82	0.06
(s.d. 10x)			(0.00)		(0.00)		(0.00)		(0.00)			
p	0.00	1.00	0.59	0.38	0.53	0.38	0.48	0.39	0.52	0.39	0.55	0.38
(s.d. 10x)			(0.03)		(0.02)		(0.04)		(0.04)			

Draws per iteration ($\eta \times \rho$)	5000×5	5000×5	250 000	250 000
Number of iterations	8	5	8	8
Average time per iteration (in s)	23.7	23.5	25.1	24.8
Mahalanobis distance	0.04	0.02	0.20	0.15
Acceptance rate (in %)	17.6		1.2	
5% most influential weights (in %)		57.9		99.7

Further, note the two modes in the distributions of (p, β_{11}) and (p, β_{21}): one mode for p close to 0 and one mode for p close to 1. In fact, the data augmentation algorithm hardly moves from one mode to the other. This can be seen from the high serial correlation in the data augmentation sequence for the parameter p, which is 0.996. For other models and data sets, the probability of "mode hopping" can be even smaller than in this example. In that case, the data augmentation may require billions of draws in order to obtain reliable estimation results. In such situations, the ARDS methods are much more reliable (and quicker) alternatives.

Table 5: Data augmentation sampling results for the two-regime mixture model for US real GNP growth

	mean	s.d.
β_{11}	0.019	0.748
(s.d. 10x)	(0.018)	
β_{12}	0.407	0.287
(s.d. 10x)	(0.002)	
β_{21}	1.237	0.737
(s.d. 10x)	(0.017)	
β_{22}	-0.012	0.393
(s.d. 10x)	(0.008)	
σ	0.820	0.056
(s.d. 10x)	(0.000)	
p	0.525	0.377
(s.d. 10x)	(0.012)	
Draws	600 000	
Computing time (in s)	119.4	

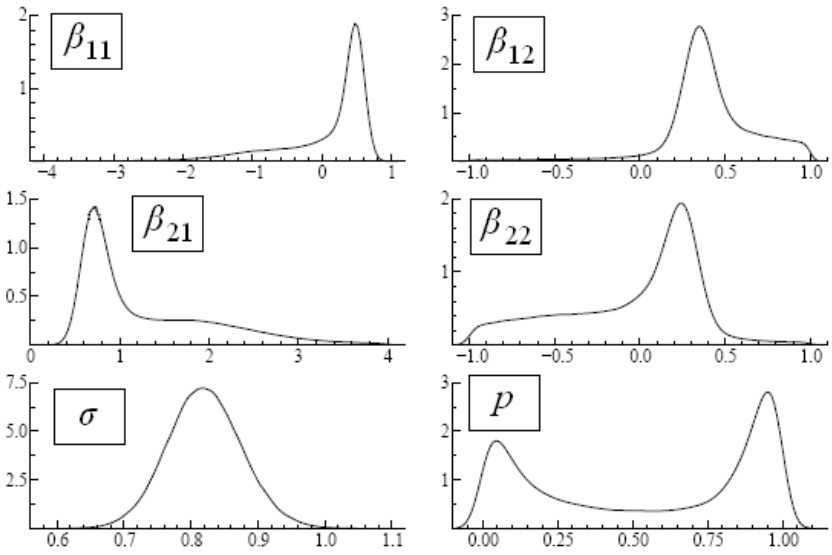

Figure 15: Estimates of marginal posterior densities in model (98) for US real GNP, based on draws generated by the data augmentation algorithm

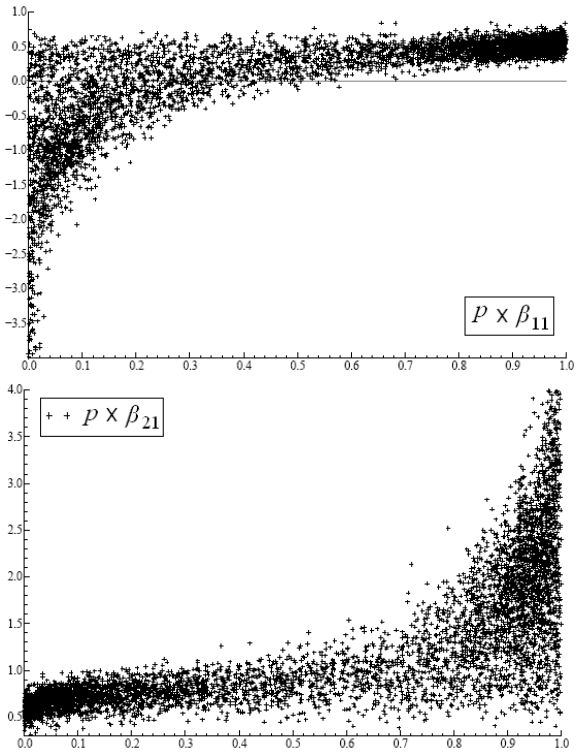

Figure 16: Model (98) for US real GNP: scatter plots of draws generated by the data augmentation algorithm

3.5.2 Neural network sampling

Neural network [NN] sampling methods, due to Hoogerheide et al. (2007), constitute a class of Monte Carlo integration methods that involve a neural network approximation to (a kernel of) the target density. Just like the ARDS algorithms, the NN algorithms may be especially useful for Bayesian inference in models with non-elliptical, possibly multi-modal posterior distributions. A key step is the construction of a NN function that provides a 'reasonably good' approximation to the target density. After a NN approximation to the target density has been obtained, this NN function is used as a candidate density in the Metropolis-Hastings or importance sampling method.

Hoogerheide et al. (2007) show examples of highly non-elliptical, bimodal posterior distributions that may occur in the instrumental variables [IV] regression model with weak instruments. In these cases a sampling method based on an approximation by a mixture of Student-t densities (which is a specific type of NN function) outperforms several competing algorithms – the Gibbs sampler, importance sampling and the Metropolis-Hastings algorithm with a Student-t candidate distribution – in the sense of yielding more accurate estimates of posterior moments in the same computing time. Hoogerheide et al. (2007) propose a quick, iterative method for constructing such an approximation to a target density by a mixture of t densities, the Adaptive Mixture of t [AdMit] method that will be discussed below.

The NN sampling algorithms share two advantages with the ARDS methods. First, the NN sampling algorithms also require little information on the shape of the target density. Again, only approximate guesses of location and scale are required as initial values. Second, the NN sampling algorithms are also flexible and robust. NN sampling methods can also handle highly non-elliptical target densities such as multi-modal, extremely skew or heavy-tailed target densities.

Neural network sampling methods provide estimates of characteristics of a posterior distribution with density kernel $p(\theta)$ with $\theta \in \mathbb{R}^m$ by the following steps:

1. Construct a neural network approximation $nn : \mathbb{R}^m \rightarrow \mathbb{R}$ to the target density kernel $p(\theta)$.
2. Obtain a sample of draws from the density (kernel) $nn(\theta)$.
3. Perform importance sampling or the (independence chain) Metropolis-Hastings algorithm,
   using this sample of draws from $nn(\theta)$ in order to obtain estimates of the characteristics of $p(\theta)$.

Hoogerheide et al. (2007) consider three types of neural networks that are members of the class of four-layer feed-forward neural networks. Here we only consider the

type that performs best in their examples, the mixture of Student-t distributions:

$$nn(\theta) = \sum_{h=1}^{H} p_h \; t(\theta|\mu_h, \Sigma_h, \nu), \tag{100}$$

where p_h $(h = 1, \ldots, H)$ are the probabilities of the Student-t components and where $t(\theta|\mu_h, \Sigma_h, \nu)$ is a multivariate t density with mode vector μ_h, scaling matrix Σ_h, and ν degrees of freedom:

$$t(\theta|\mu_h, \Sigma_h, \nu) = \frac{\Gamma((\nu+m)/2)}{\Gamma(\nu/2)(\pi\nu)^{m/2}} \, |\Sigma_h|^{-1/2} \left(1 + \frac{(\theta-\mu_h)'\Sigma_h^{-1}(\theta-\mu_h)}{\nu}\right)^{-(\nu+m)/2}. \tag{101}$$

Note that this mixture of t densities is a four-layer feed-forward neural network function

$$nn(\theta) = eG_2\left(CG_1(A\theta + b) + d\right) + f, \qquad \theta \in \mathbb{R}^m, \tag{102}$$

in which the vector functions $G_1(.)$ and $G_2(.)$ apply the following scalar functions $g_1(.)$ and $g_2(.)$ to each element of their argument vector, respectively:

$$g_1(x) = x^2 \quad \text{and} \quad g_2(x) = x^{-(\nu+m)/2} \frac{\Gamma((\nu+m)/2)}{\Gamma(\nu/2)(\pi\nu)^{m/2}}, \qquad x \in \mathbb{R},$$

and with weights $e_h = p_h \, |\Sigma_h|^{-1/2}$ $(h = 1, \ldots, H)$, $f = 0$ and:

$$A = \begin{pmatrix} \Sigma_1^{-1/2} \\ \vdots \\ \Sigma_H^{-1/2} \end{pmatrix}, \; b = \begin{pmatrix} -\Sigma_1^{-1/2}\mu_1 \\ \vdots \\ -\Sigma_H^{-1/2}\mu_H \end{pmatrix}, \; C = \begin{pmatrix} \iota_m'/\nu & 0 & \cdots & 0 \\ 0 & \iota_m'/\nu & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \iota_m'/\nu \end{pmatrix}, \; d = \iota_H;$$

ι_k denotes a $k \times 1$ vector of ones. Notice that $(\theta-\mu_h)'\Sigma_h^{-1}(\theta-\mu_h)$ is the sum of the squared elements of $\Sigma_h^{-1/2}(\theta-\mu_h)$. The reason for this choice is that a mixture of t distributions allows for easy and quick sampling, and that the Student t distribution has fatter tails than the normal distribution. This property causes that these NN sampling methods can cope with fat-tailed target distributions. Note that the p_h $(h = 1, \ldots, H)$ in (100) have to satisfy $\sum_{h=1}^{H} p_h = 1$. Zeevi and Meir (1997) show that under certain conditions any density function can be approximated to arbitrary accuracy by a convex combination of 'basis' densities; the mixture of Student t densities in (100) falls within their framework. This makes these NN sampling methods flexible and robust, as a wide variety of target density functions can be well approximated by mixtures of t distributions.

The Adaptive Mixture of t [AdMit] method

The Adaptive Mixture of t [AdMit] method of Hoogerheide et al. (2007) constructs a mixture-of-t approximation to a certain target density with kernel $P(\theta)$ by the following steps.

First, compute the mode μ_1 and scale Σ_1 of the first Student t distribution in the mixture as $\mu_1 = \text{argmax}_\theta\, P(\theta)$, the mode of the target distribution, and Σ_1 as minus the inverse Hessian of $\log P(\theta)$ evaluated at its mode μ_1. Then draw a set of points θ_i $(i = 1, \ldots, n)$ from the 'first stage candidate density' $nn(\theta) = t(\theta|\mu_1, \Sigma_1, \nu)$, with small ν to allow for fat tails; for example, $\nu = 1$. Next, add components to the mixture, iteratively, by performing the following steps:

Step 1: Compute the importance sampling weights $w(\theta_i) = P(\theta_i)/nn(\theta_i)$ $(i = 1, \ldots, n)$. In order to determine the number of components H of the mixture we make use of a simple diagnostic criterion: the coefficient of variation, i.e. the standard deviation divided by the mean, of the IS weights $w(\theta_i)$ $(i = 1, \ldots, n)$. If the relative decrease in the coefficient of variation of the IS weights caused by adding one new Student-t component to the candidate mixture is small, e.g. less than 10%, then stop: the current $nn(\theta)$ is our approximation to the target density. Otherwise, go to step 2. Notice that $nn(\theta)$ is a proper density, whereas $P(\theta)$ is merely a density kernel. So, the neural network does not provide an approximation to the target density kernel $P(\theta)$ in the sense that $nn(\theta) \approx P(\theta)$. Instead, $nn(\theta)$ provides an approximation to the density of which $P(\theta)$ is a kernel in the sense that the ratio $P(\theta)/nn(\theta)$ has relatively little variation.

Step 2: Add another Student t distribution with density $t(\theta|\mu_h, \Sigma_h, \nu)$ to the mixture with $\mu_h = \text{argmax}_\theta\, w(\theta) = \text{argmax}_\theta\{P(\theta)/nn(\theta)\}$ and Σ_h equal to minus the inverse Hessian of $\log w(\theta) = \log P(\theta) - \log nn(\theta)$ evaluated at its mode μ_h. Here $nn(\theta)$ denotes the mixture of $(h-1)$ Student t densities obtained in the previous iteration of the procedure. An obvious initial value for the maximization procedure for computing $\mu_h = \text{argmax}_\theta\, w(\theta)$ is the point θ_i with the highest weight $w(\theta_i)$ in the sample $\{\theta_i | i = 1, \ldots, n\}$. The idea behind this choice of μ_h and Σ_h is that the new t component should 'cover' a region where the weights $w(\theta)$ are relatively large: the point where the weight function $w(\theta)$ attains its maximum is an obvious choice for the mode μ_h, while the scale Σ_h is the covariance matrix of the local normal approximation to the distribution with density kernel $w(\theta)$ around the point μ_h.

If the region of integration of the parameters θ is bounded, it may occur that $w(\theta)$ attains its maximum at the boundary of the integration region; in this case minus the inverse Hessian of $\log w(\theta)$ evaluated at its mode μ_h may be a very poor scale matrix; in fact this matrix may not even be positive definite.

In that case μ_h and Σ_h are obtained as estimates of the mean and covariance matrix of a certain 'residual distribution' with density kernel:

$$res(\theta) = \max\{p(\theta) - \tilde{c}\, nn(\theta), 0\}, \tag{103}$$

where $\tilde{c}$ is a constant; we take $\max\{.,0\}$ to make it a (non-negative) density kernel. These estimates of the mean and covariance matrix of the 'residual distribution' are easily obtained by importance sampling with the current $nn(\theta)$ as the candidate density, using the sample θ_i $(i = 1, \ldots, n)$ from $nn(\theta)$ that we already have. The weights $w_{res}(\theta_i)$ and scaled weights $\tilde{w}_{res}(\theta_i)$ $(i = 1, \ldots, n)$ are:

$$w_{res}(\theta_i) = \frac{res(\theta_i)}{nn(\theta_i)} = \max\{w(\theta_i) - \tilde{c}, 0\} \quad \text{and} \quad \tilde{w}_{res}(\theta_i) = \frac{w_{res}(\theta_i)}{\sum_{i=1}^{n} w_{res}(\theta_i)}, \tag{104}$$

and μ_h and Σ_h are obtained as:

$$\mu_h = \sum_{i=1}^{n} \tilde{w}_{res}(\theta_i)\theta_i \qquad \Sigma_h = \sum_{i=1}^{n} \tilde{w}_{res}(\theta_i)(\theta_i - \mu_h)(\theta_i - \mu_h)'. \tag{105}$$

There are two issues relevant for the choice of $\tilde{c}$ in (103) and (104). First, the new t density should appear exactly at places where $nn(\theta)$ is too small (relative to $P(\theta)$), i.e. the scale should not be too large. Second, there should be enough points θ^i with $w(\theta^i) > \tilde{c}$ in order to make Σ_h nonsingular. A procedure is to calculate Σ_h for $\tilde{c}$ equal to 100 times the average value of $w(\theta_i)$ $(i = 1, \ldots, n)$; if Σ_h in (105) is nonsingular, accept $\tilde{c}$; otherwise lower $\tilde{c}$.

Step 3: Choose the probabilities p_h $(h = 1, \ldots, H)$ in the mixture $nn(\theta) = \sum_{h=1}^{H} p_h\, t(\theta|\mu_h, \Sigma_h, \nu)$ by minimizing the (squared) coefficient of variation of the importance sampling weights. First, draw n points θ_i^h from each component $t(\theta|\mu_h, \Sigma_h, \nu)$ $(h = 1, \ldots, H)$. Then minimize $E[w(\theta)^2]/E[w(\theta)]^2$, where:

$$E[w(\theta)^k] = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} p_h\, w\big(\theta_i^h\big)^k \quad (k = 1, 2), \qquad w\big(\theta_i^h\big) = \frac{P(\theta_i^h)}{\sum_{l=1}^{H} p_l\, t\big(\theta_i^h|\mu_l, \Sigma_l, \nu\big)}. \tag{106}$$

Step 4: Draw a sample of n points θ_i $(i = 1, \ldots, n)$ from our new mixture of t distributions, $nn(\theta) = \sum_{h=1}^{H} p_h\, t(\theta|\mu_h, \Sigma_h, \nu)$, and go to step 1; in order to draw a point from the density $nn(\theta)$ first use a draw from the $U(0, 1)$ distribution to determine which component $t(\theta|\mu_h, \Sigma_h, \nu)$ is chosen, and then draw from this multivariate t distribution.

It may occur that one is dissatisfied with diagnostics like the coefficient of variation of the IS weights corresponding to the final candidate density resulting from the

procedure above. In that case one may start all over again with a larger number of points n. The idea behind this is that the larger n is, the easier it is for the method to 'feel' the shape of the target density kernel, and to specify the t distributions of the mixture adequately.

Note that an advantage of the AdMit approach is that it does not require the specification of a certain bounded region where the random variable $\theta \in \mathbb{R}^m$ takes its values.

After the construction of the NN approximation to the target density, one can simply use the NN candidate density in importance sampling or the Metropolis-Hastings algorithm. Here an advantage is that it is very easy to sample from a mixture of t distributions. Convergence properties of the NN sampling methods follow directly from those for the importance sampling and Metropolis-Hastings algorithms.

Example: NN sampling in two-regime mixture model for the US GNP growth rate

In order to illustrate the advantages of the AdMit methods, we investigate a mixture model for the analysis of economic growth in the USA, which is also considered by Bauwens et al. (2004) and the previous subsection of this paper. We consider a mixture model with two AR(1) regimes for real GNP growth:

$$\begin{aligned} y_t &= \begin{cases} \beta_{11} + \beta_{12} y_{t-1} + \varepsilon_t & \text{with probability } p, \\ \beta_{21} + \beta_{22} y_{t-1} + \varepsilon_t & \text{with probability } 1-p, \end{cases} \\ \varepsilon_t &\sim N(0, \sigma^2), \end{aligned} \tag{107}$$

where y_t denotes the quarterly growth rate. The data (source: Economagic) consist of observations from the first quarter of 1959 to the last quarter of 2001. We specify the prior $p(\theta) = p(\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \sigma, p) = 1/\sigma$. For identification, it is imposed that $\beta_{11} < \beta_{21}$.

First, the AdMit approach constructs a candidate distribution; in this case it yields a mixture of 10 Student t distributions. Next, we use this candidate distribution in the (independence chain) MH algorithm and IS. Sampling results are given by Table 6. The number of draws has been chosen in order to make the computing time comparable with the methods in the previous subsection (ARDS methods, MH, IS, and data augmentation). For both AdMit methods, we repeat the simulation process ten times with different random seeds, after which we compute the standard deviations of the ten estimates of the posterior means. Notice that except for the parameter β_{12}, for which the data augmentation algorithm is somewhat more precise, the AdMit methods outper-

form the competing approaches. This is remarkable, as the AdMit methods only require a kernel of the posterior density (and approximate guesses of the location and scale), whereas the data augmentation method requires that the conditional posterior distribution of each parameter is derived. The serial correlation in the AdMit-MH sequence for the parameter p is 0.914, which is much lower than the serial correlation of 0.996 in the data augmentation approach.

In this example, the ARDS methods have a lower precision than the AdMit methods, given the same computing time. This is caused by the much smaller number of draws in the ARDS algorithms. The process of evaluating a one-dimensional integral over distances given a direction and sampling from the exact conditional target distribution given a direction is relatively quite time consuming. However, because of this sampling from the exact target distribution given a direction, the ARDS methods may be more robust and reliable than the AdMit methods in other cases of highly non-elliptical posterior distributions. Furthermore, an interesting topic for further research is to combine these algorithms in a clever way.

Table 6: NN Sampling results for the two-regime mixture model for US real GNP growth

	AdMit-IS		AdMit-MH	
	mean	s.d.	mean	s.d.
β_{11}	0.052	0.743	0.053	0.716
(s.d. 10x)	(0.011)		(0.013)	
β_{12}	0.409	0.284	0.410	0.282
(s.d. 10x)	(0.005)		(0.006)	
β_{21}	1.276	0.762	1.278	0.762
(s.d. 10x)	(0.004)		(0.009)	
β_{22}	-0.026	0.399	-0.025	0.400
(s.d. 10x)	(0.002)		(0.002)	
σ	0.820	0.055	0.820	0.055
(s.d. 10x)	(0.000)		(0.000)	
p	0.547	0.374	0.548	0.374
(s.d. 10x)	(0.002)		(0.005)	
Draws	500 000		500 000	
Computing time: NN construction (in s)	73.3		73.3	
Computing time: NN sampling (in s)	40.5		40.9	
Computing time: total (in s)	113.8		114.2	
Acceptance rate (in %)			9.5	
5% most influential weights (in %)	67.7			

4 Concluding remarks

In this paper we discussed several aspects of simulation based Bayesian econometric inference [SBBEI]. First, we showed that the Bayesian framework provides a natural learning rule, that allows for optimal learning and (hence) optimal decision making under uncertainty. The Bayesian framework provides a proper way to consider the sensitivity of estimates and to use probability statements that indicate a 'degree of confidence'. We discussed the basic principles of Bayesian inference (prior & posterior density, Bayes' rule, Highest Posterior Density [HPD] region, posterior odds) and described some substantial differences between Bayesian inference and the frequentist/classical approach. We showed that evaluating integrals by simulation methods is a crucial ingredient in Bayesian inference.

After that, we discussed some of the most popular and well-known simulation techniques, plus two recently developed sampling methods: adaptive radial based direction sampling [ARDS], which makes use of a transformation to radial coordinates, and neural network sampling, which makes use of a neural network approximation

to the posterior distribution of interest. Both methods are especially useful in cases where the posterior distribution is not well-behaved, in the sense of having highly non-elliptical shapes. We illustrated the simulation techniques with several example models, such as a model for the real US GNP and models for binary data of a US recession indicator.

The development of advanced sampling methods, that perform the evaluation of integrals efficiently, makes Bayesian inference possible in an ever increasing number of complex models. This allows for more realistic descriptions of many processes in several fields of research, for example in finance and macro-economics, leading to more accurate forecasts, a better quantification of uncertainty, and hence better policies.

It should be noted that we did not attempt to provide a complete survey of simulation methods. For further reading we refer to the textbooks by, in alphabetical order, Bauwens et al. (1999), Geweke (2005), Koop (2003), Lancaster (2004) and Rossi et al. (2005).

Acknowledgements

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.

References

[1] Albert J.H. and S. Chib (1993), "Bayesian Analysis of Binary and Polychotomous Response Data", *Journal of the American Statistical Association*, 88, 669−679.

[2] Bartlett M.S. (1957), "A Comment on D.V. Lindley's Statistical Paradox", *Biometrika*, 44, 533−534.

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

[4] Bauwens L., C.S. Bos, H.K. van Dijk and R.D. van Oest (2004), "Adaptive Radial-Based Direction Sampling: Some Flexible and Robust Monte Carlo Integration Methods", *Journal of Econometrics*, 123, 201−225.

[5] Box G.E.P. and M.E. Muller (1958), "A note on the generation of random normal deviates", *Annals of Mathematical Statistics*, 29, 610611.

[6] Casella G. and E. George (1992), "Explaining the Gibbs Sampler", *The American Statistician*, 46, 167−174.

[7] Chib S. (1992), "Bayesian Inference in the Tobit Censored Regression Model", *Journal of Econometrics*, 51, 79−99.

[8] Chib S. and E. Greenberg (1995), "Understanding the Metropolis-Hastings algorithm", *The American Statistician*, 49, 327−335.

[9] Damien P., J. Wakefield and S. Walker (1999), "Gibbs Sampling for Bayesian Non-conjugate and Hierarchical Models by using Auxiliary Variables", *Journal of the Royal Statistical Society B*, 61, 331−344.

[10] De Finetti B. (1974), *Theory of Probability*, John Wiley & Sons, Chichester.

[11] Dellaportas P. and A.F.M. Smith (1993), "Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling", *Applied Statistics*, 42, 443−459.

[12] Dickey J. (1971), "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters", *The Annals of Statistics*, 42, 204−223.

[13] Edwards R.G. and A.D. Sokal (1988), "Generalization of the Fortuin-Kasteleyn-Swendsen-Wang Representation and Monte Carlo Algorithm", *Physical Review D*, 38, 2009−2012.

[14] Frühwirth-Schnatter S. (2001), "Markov chain Monte Carlo estimation of classical and dynamic switching models", *Journal of the American Statistical Association*, 96, 194−209.

[15] Gelfand A.E. and D.K. Dey (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations", *Journal of the Royal Statistical Society B*, 56, 501−514.

[16] Gelfand A.E., S.E. Hills, A. Racine-Poon and A.F.M. Smith (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling", *Journal of the American Statistical Association*, 85, 972−985.

[17] Gelfand A.E. and A.F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association*, 85, 398−409.

[18] Geman S. and D. Geman (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721−741.

[19] Geweke J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, 1317−1339.

[20] Geweke J. (1999), "Using simulation methods for Bayesian econometric models: inference, development, and communication", *Econometric Reviews*, 18, 173.

[21] Geweke J. (2005), *Contemporary Bayesian Econometrics and Statistics*, John Wiley & Sons, New Jersey.

[22] Gilks W.R. and G.O. Roberts (1996), "Strategies for Improving MCMC", in *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 89−114, Chapman and Hall.

[23] Gilks W.R., G.O. Roberts and E.I. George (1994), "Adaptive direction sampling". *The Statistician* 43, 179−189.

[24] Hamilton J.D. (2006), "Computing Power and the Power of Econometrics", Medium Econometrische Toepassingen, 14(2), 32−38.

[25] Hammersley J.M. and D.C. Handscomb (1964), *Monte Carlo Methods*, first edition, Methuen, London.

[26] Hastings W.K. (1970), "Monte Carlo Sampling Methods using Markov Chains and their Applications", *Biometrika*, 57, 97−109.

[27] Hausman J.A. and D.A. Wise (1978), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences", *Econometrica*, 46, 403−426.

[28] Hoogerheide L.F., J.F. Kaashoek and H.K. van Dijk (2007), "On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks", *Journal of Econometrics*, forthcoming, 27 pages.

[29] Jeffreys H. (1961), *The Theory of Probability*, third edition, Oxford: Clarendon Press.

[30] Kass R.E. and A.E. Raftery (1995), "Bayes Factors", Journal of the American Statistical Association, 90, 773−795.

[31] Kloek T. and H.K. van Dijk (1978), "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo", *Econometrica*, 46, 1−20.

[32] Koop G. (2003), *Bayesian Econometrics*, Wiley, New Jersey.

[33] Lancaster T. (2004) *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing, Oxford.

[34] Law A.M. and W.D. Kelton (1991), *Simulation Modeling and Analysis*, second edition, McGraw-Hill.

[35] Lehmer D.H. (1951), "Mathematical Methods in Large-Scale Computing Units", *Ann. Comput. Lab. Harvard University*, 26, 141−146.

[36] Lewis P.A.W., A.S. Goodman and J.M. Miller (1969), "A Pseudo-Random Number Generator for the System/360", *IBM Syst. Journal*, 8, 136−146.

[37] Lindley D.V. (1957), "A Statistical Paradox", *Biometrika*, 44, 187-192.

[38] McCulloch R., N.G. Polson and P.E. Rossi (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters", *Journal of Econometrics*, 99, 173−193.

[39] McCulloch R. and P.E. Rossi (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model", *Journal of Econometrics*, 64, 207−240.

[40] Metropolis N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller (1953), "Equation of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, 21, 1087−1092.

[41] Monahan J. and A. Genz (1997), "Spherical-radial integration rules for Bayesian computation", *Journal of the American Statistical Association*, 92, 664−674.

[42] Newton M.A. and A.E. Raftery (1994), "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap", *Journal of the Royal Statistical Society B*, 56, 3−48.

[43] Norris J.R. (1997), *Markov Chains*, Cambridge University Press.

[44] Oh M.S. and J.O. Berger (1992), "Adaptive importance sampling in Monte Carlo integration", *Journal of Statistical Computation and Simulation* 41, 143−168.

[45] Payne W.H., J.R. Rabung and T.P. Bogyo (1969), "Coding the Lehmer Pseudorandom Number Generator", *Commun. Assoc. Comput. Mach.*, 12, 85−86.

[46] Ritter C. and M.A. Tanner (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler", *Journal of the American Statistical Association*, 87, 861−868.

[47] Ross S.M. (1997), *Introduction to Probability Models*, sixth edition, Academic Press.

[48] Rossi P.E., G.M. Allenby and R. McCulloch (2005), *Bayesian Statistics and Marketing*, Wiley, Chichester.

[49] Rubinstein R.Y. (1981), *Simulation and the Monte Carlo Method*, John Wiley & Sons.

[50] Smith A.F.M. and G.O. Roberts (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods", *Journal of the Royal Statistical Society B*, 55, 3−23.

[51] Stoer J. and R. Bulirsch (1993), *Introduction to Numerical Analysis*, second edition, Springer-Verlag.

[52] Tanner M.A. and W.H. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528−550.

[53] Tierney L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701−1762.

[54] Van Dijk H.K. and T. Kloek (1980), "Further experience in Bayesian analysis using Monte Carlo integration", *Journal of Econometrics*, 14, 307−328.

[55] Van Dijk H.K. and T. Kloek (1984), "Experiments with some alternatives for simple importance sampling in Monte Carlo integration". In: Bernardo, J.M., Degroot, M., Lindley, D., Smith, A.F.M. (Eds.), *Bayesian Statistics*, Vol. 2. Amsterdam, North Holland.

[56] Van Dijk H.K., T. Kloek and C.G.E. Boender (1985), "Posterior moments computed by mixed integration", *Journal of Econometrics*, 29, 3−18.

[57] Van Dijk H.K. (1999), "Some Remarks on the simulation revolution in Bayesian econometric inference", Econometric Reviews, 18(1), 105−112.

[58] 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, 614−618.

[59] Wei G.C.G. and M.A. Tanner (1990), "Posterior Computations for Censored Regression Data", *Journal of the American Statistical Association*, 85, 829−839.

[60] Zeevi A.J. and R. Meir (1997), "Density estimation through convex combinations of densities; approximation and estimation bounds", *Neural Networks*, 10, 99−106.