Bayesian analysis

Class Notes

Manuel Arellano

March 8, 2016

1 Introduction

Bayesian methods have traditionally had limited influence in empirical economics, but they have become increasingly important with the popularization of computer-intensive stochastic simulation algorithms in the 1990s. This is particularly so in macroeconomics, where applications of Bayesian inference include vector autoregressions (VARs) and dynamic stochastic general equilibrium (DSGE) models. Bayesian approaches are also attractive in models with many parameters, such as panel data models with individual heterogeneity and flexible nonlinear regression models. Examples include discrete choice models of consumer demand in the fields of industrial organization and marketing.

An empirical study uses data to learn about quantities of interest (parameters). A likelihood function or some of its features specify the information in the data about those quantities. Such specification typically involves the use of a priori information in the form of parametric or functional restrictions. In the Bayesian approach to inference, one not only assigns a probability measure to the sample space but also to the parameter space. Specifying a probability distribution over potential parameter values is the conventional way of modelling uncertainty in decision-making, and offers a systematic way of incorporating uncertain prior information into statistical procedures.

Outline The following section introduces the Bayesian way of combining a prior distribution with the likelihood of the data to generate point and interval estimates. This is followed by some comments on the specification of prior distributions. Next we turn to discuss asymptotic approximations; the main result is that in regular cases there is a large-sample equivalence between Bayesian probability statements and frequentist confidence statements. As a result, frequentist and Bayesian inferences are often very similar and can be reinterpreted in each other's terms. Finally, we review Markov chain Monte Carlo methods (MCMC). The development of these methods has greatly reduced the computational difficulties that held back Bayesian applications in the past.

Bayesian methods are now not only generally feasible, but sometimes also a better practical alternative to frequentist methods. The upshot is an emerging Bayesian/frequentist synthesis around increasing agreement on what works for different kinds of problems. The shifting focus from philosophical debate to methodological considerations is a healthy state of affairs because both frequentist and Bayesian approaches have features that are appealing to most scientists.

2 Bayesian inference

Let us consider a data set $y = (y_1, ..., y_n)$ and a probability density (or mass) function of y conditional on an unknown parameter θ:

$$f(y_1, ..., y_n \mid \theta).$$

If y is an *iid* sample then $f(y_1, ..., y_n \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$ where $f(y_i \mid \theta)$ denotes the pdf of y_i. In survey sampling $f(y_i \mid \theta = \theta_0)$ is the pdf of the population, $(y_1, ..., y_n)$ are n independent draws from such population, and θ_0 denotes the true value of θ in the pdf that generated the data.

In general, for shortness we just write $f(y \mid \theta) = f(y_1, ..., y_n \mid \theta)$. As a function of the parameter this is called the likelihood function, also denoted $\mathcal{L}(\theta)$. We are interested in inference about the unknown parameter given the data. Any uncertain prior information about the value of θ is specified in a prior probability distribution for the parameter, $p(\theta)$. Both the likelihood and the prior are chosen by the researcher. We then combine the prior distribution and the sample information, using Bayes' theorem, to obtain the conditional distribution of the parameter given the data, also known as the posterior distribution:

$$p(\theta \mid y) = \frac{f(y, \theta)}{f(y)} = \frac{f(y \mid \theta) p(\theta)}{\int f(y \mid \theta^*) p(\theta^*) d\theta^*}.$$

Note that as a function of θ, the posterior density is proportional to

$$p(\theta \mid y) \propto f(y \mid \theta) p(\theta) = \mathcal{L}(\theta) p(\theta).$$

Once we calculate this product, all we have to do is to find the constant that makes this expression integrate to one as a function of the parameter. The posterior density describes how likely it is that a value of θ has generated the observed data.

Point estimation We can use the posterior density to form optimal point estimates. The notion of optimality is minimizing mean posterior loss for some loss function $\ell(r)$:

$$\min_c \int_\Theta \ell(c - \theta) p(\theta \mid y) d\theta$$

The posterior mean

$$\overline{\theta} = \int_\Theta \theta p(\theta \mid y) d\theta$$

is the point estimate that minimizes mean squared loss $\ell(r) = r^2$. The posterior median minimizes mean absolute loss $\ell(r) = |r|$. The posterior mode $\widetilde{\theta}$ is the maximizer of the posterior density and minimizes mean Dirac loss. When the prior density is flat, the posterior mode coincides with the maximum likelihood estimator.

Interval estimation The posterior quantiles characterize the posterior uncertainty about the parameter, and they can be used to obtain interval estimates. Any interval (θ_ℓ, θ_u) such that

$$\int_{\theta_\ell}^{\theta_u} p(\theta \mid y)\, d\theta = 1 - \alpha$$

is called a credible interval with coverage probability $1-\alpha$. If the posterior density is unimodal, a common choice is the shortest connected credible interval or the highest posterior density (HPD) interval. In practice, often an equal-tail-probability interval is favored because of its computational simplicity. In such case, θ_ℓ and θ_u are just the $\alpha/2$ and $1-\alpha/2$ posterior quantiles, respectively. Equal-tail-probability intervals tend to be longer than the others, except in the case of a symmetric posterior density. If the posterior is multi-modal then the HPD interval may consist of disjoint segments.[1]

Frequentist confidence intervals and Bayesian credible intervals are the two main interval estimation methods in statistics. In a confidence interval the coverage probability is calculated from a sampling density, whereas in a credible interval the coverage probability is calculated from a posterior density. As discussed in the next section, despite the differences in the two methods, they often provide similar interval estimates in large samples.

Bernoulli example Let us consider a random sample $(y_1, ..., y_n)$ of Bernoulli random variables. The likelihood of the sample is given by

$$\mathcal{L}(\theta) = \theta^m (1-\theta)^{n-m}$$

where $m = \sum_{i=1}^n y_i$. The maximum likelihood estimator is

$$\widehat{\theta} = \frac{m}{n}.$$

In general, given some prior $p(\theta)$, the posterior mode solves

$$\widetilde{\theta} = \arg\max_\theta \left[\ln \mathcal{L}(\theta) + \ln p(\theta)\right].$$

Since θ is a probability value, a suitable parameter space over which to specify a prior probability distribution is the $(0,1)$ interval. A flexible and convenient choice is the Beta distribution:

$$p(\theta; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

where $B(\alpha,\beta)$ is the beta function, which is constant with respect to θ:

$$B(\alpha,\beta) = \int_0^1 s^{\alpha-1} (1-s)^{\beta-1}\, ds.$$

The quantities (α, β) are parameters of the prior, to be set according to our a priori information about θ. These parameters are called prior hyperparameters.

[1] The minimum density of any point within the HPD interval exceeds the density of any point outside that interval.

The Beta distribution is a convenient prior because the posterior is also a Beta distribution:

$$p(\theta \mid y) \propto \mathcal{L}(\theta)\, p(\theta) \propto \theta^{m+\alpha-1} (1-\theta)^{n-m+\beta-1}$$

That is, if $\theta \sim Beta(\alpha, \beta)$ then $\theta \mid y \sim Beta(\alpha + m, \beta + n - m)$. This situation is described by saying that the Beta distribution is the conjugate prior to the Bernoulli.

The posterior mode is given by

$$\widetilde{\theta} = \arg\max_{\theta} \left[\theta^{m+\alpha-1} (1-\theta)^{n-m+\beta-1}\right] = \frac{m+\alpha-1}{n+\beta-1+\alpha-1}. \tag{1}$$

This result illustrates some interesting properties of the posterior mode in this example. The posterior mode is equivalent to the ML estimate of a data set with $\alpha-1$ additional ones and $\beta-1$ additional zeros. Such data augmentation interpretation provides guidance on how to choose α and β in describing a priori knowledge about the probability of success in Bernoulli trials. It also illustrates the vanishing effect of the prior in a large sample. Note for now that if n is large $\widetilde{\theta} \approx \widehat{\theta}$. However, maximum likelihood may not be a satisfactory estimator in a small sample that only contains zeros if the probability of success is known a priori to be greater than zero.

3 Specification of prior distribution

There is a diversity of considerations involved in the specification of a prior, not altogether different from those involved in the specification of a likelihood model.

Conjugate priors One consideration in selecting the form of both prior and likelihood is mathematical convenience. Conjugate prior distributions, such as the Beta density in the previous example, have traditionally played a central role in Bayesian inference for analytical and computational reasons. A prior is conjugate for a family of distributions if the prior and the posterior are of the same family. When a likelihood model is used together with its conjugate prior, the posterior is not only known to be from the same family of densities as the prior, but explicit formulas for the posterior hyperparameters are also available. In general, distributions in the exponential family have conjugate priors. Some likelihood models together with their conjugate priors are the following:

- Bernoulli – Beta
- Binomial – Beta
- Poisson – Gamma
- Normal with known variance – Normal
- Exponential – Gamma
- Uniform – Pareto
- Geometric – Beta

Conjugate priors not only have advantages in tractability but also in interpretation, since the prior can be interpreted in terms of a prior sample size or additional pseudo-data (as illustrated in the Bernoulli example).

Informative priors The argument for using a probability distribution to specify uncertain a priori information is more compelling when prior knowledge can be associated to past experience, or to a process of elicitation of consensus expert views. Other times, a parameter is a random realization drawn from some population, for example, in a model with individual effects for longitudinal survey data; a situation in which there exists an actual population prior distribution. In those cases one would like the prior to accurately express the information available about the parameters. However, often little is known a priori and one would like a prior density to just express lack of information, an issue that we consider next.

Flat priors For a scalar θ taking values on the entire real line a uniform, a flat prior distribution is typically employed as an uninformative prior, that is, one that sets $p(\theta) = 1$. A flat prior is non-informative in the sense of having little impact on the posterior, which is simply a renormalization of the likelihood into a density for θ.[2] A flat prior is therefore appealing from the point of view of seeking to summarize the likelihood.

Note that a flat prior is improper in the sense that $\int_{\Theta} p(\theta)\, d\theta = \infty$.[3] If an improper prior is combined with a likelihood that cannot be renormalized (due to lacking a finite integral with respect to θ), the result is an improper posterior that cannot be used for inference. Flat priors are often approximated by a proper prior with a large variance.

If $p(\theta)$ is uniform, then the prior of a transformation of θ is not uniform. If θ is a positive number, a standard reference prior is to assume a flat prior on $\ln\theta$, $p(\ln\theta) = 1$, which implies

$$p(\theta) = \frac{1}{\theta}. \tag{2}$$

Similarly, if θ lies in the $(0,1)$ interval, a flat prior on the logit transformation of $\theta, \ln\left(\frac{\theta}{1-\theta}\right)$, implies

$$p(\theta) = \frac{1}{\theta(1-\theta)}. \tag{3}$$

These priors are improper because $\int_0^{\infty} \frac{1}{\theta} d\theta$ and $\int_0^1 \frac{1}{\theta(1-\theta)} d\theta$ both diverge. They are easily dominated by the data, but (2) assigns most of the weight to values of θ that are either very large or very close to zero, and (3) puts most of the weight on values very near 0 and 1.

For example, if $(y_1, ..., y_n)$ is a random sample from a normal population $\mathcal{N}\left(\mu, \frac{1}{\tau}\right)$, the standard improper reference prior for (μ, τ) is to specify independent flat priors on μ and $\ln\tau$, so that

$$p(\mu, \tau) = p(\mu)\, p(\tau) = \frac{1}{\tau}.$$

[2] Though arguably a flat prior places a large weight on extreme parameter values.

[3] Any prior distribution with infinite mass is called improper.

Jeffreys prior It is a rule for choosing a non-informative prior that is invariant to transformation:

$$p(\theta) \propto [\det I(\theta)]^{1/2}$$

where $I(\theta)$ is the information matrix. If $\gamma = h(\theta)$ is one-to-one, applying Jeffreys' rule directly to γ we get the same prior as applying the rule to θ and then transforming to obtain $p[h(\theta)]$.

Bernoulli example continued Let us illustrate three standard candidates for non-informative prior in the Bernoulli example. The first possibility is to use a flat prior in the log-odds scale, leading to (3); this is the $Beta(0,0)$ distribution since it can be regarded as the limit of the numerator of the beta distribution as $\alpha, \beta \to 0$. The second is Jeffreys' prior, which in this case is proportional to

$$p(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}},$$

and corresponds to the $Beta(0.5, 0.5)$ distribution. Finally, the third candidate is the uniform prior $p(\theta) = 1$, which corresponds to the $Beta(1,1)$ distribution.

All three priors are data augmentation priors. The $Beta(0,0)$ prior adds no prior observations, Jeffreys' prior adds one observation with half a success and half a failure, and the uniform prior adds two observations with one success and one failure. The ML estimator coincides with the posterior mode for the $Beta(1,1)$ prior, and with the posterior mean for the $Beta(0,0)$ prior.

4 Large-sample Bayesian inference

To evaluate the performance of a point estimator we typically resort to large-sample approximations. The basic tools are consistency (convergence in probability of the estimation error to zero) and asymptotic normality (limiting sampling distribution of the scaled estimation error). Similarly, to obtain a (frequentist) confidence interval we usually rely on an asymptotic approximation. Here we wish to consider (i) asymptotic approximations to the posterior distribution, and (ii) the sampling properties of Bayesian estimators in large samples.

The main result of large-sample Bayesian inference is that as the sample size increases the posterior distribution of the parameter vector approaches a multivariate normal distribution, which is independent of the prior distribution. The convergence is in probability, where the probability is measured with respect to the true distribution of y. Posterior asymptotic results formalize the notion that the importance of the prior diminishes as n increases. Only when n is small, the prior choice is an important part of the specification of the model.

These results hold under suitable conditions on the prior distribution, the likelihood, and the parameter space. Conditions include a prior that assigns positive probability to a neighborhood about θ_0; a posterior distribution that is not improper; identification; a likelihood that is a continuous function of θ, and a true value θ_0 that is not on the boundary of Θ.

4.1 Consistency of the posterior distribution

If the population distribution of a random sample $y = (y_1, ..., y_n)$ is included in the parametric likelihood family, so that it equals $f(y_i \mid \theta_0)$ for some θ_0, the posterior is consistent in the sense that it converges to a point mass at the true parameter value θ_0 as $n \to \infty$. When the true distribution is not included in the parametric family, there is no longer a true value θ_0, except in the sense of the value θ_0 that makes the model distribution $f(y_i \mid \theta)$ closest to the true distribution $g(y_i)$ according to the Kullback-Leibler divergence:

$$KL(\theta) = \int \ln\left(\frac{g(y_i)}{f(y_i \mid \theta)}\right) g(y_i)\, dy_i,$$

so that[4]

$$\theta_0 = \arg\min_{\theta \in \Theta} KL(\theta).$$

Here is a consistency theorem of the posterior distribution for a discrete parameter space. The result is valid if $g(y_i)$ is not included in the $f(y_i \mid \theta)$ family, in which case we may refer to $\prod_{i=1}^{n} f(y_i \mid \theta)$ as a pseudo-likelihood and to $p(\theta \mid y)$ as a pseudo-posterior. The theorem and its proof are taken from Gelman et al (2014, p. 586).

Theorem (finite parameter space) If the parameter space Θ is finite and $\Pr(\theta = \theta_0) > 0$, then $\Pr(\theta = \theta_0 \mid y) \to 1$ as $n \to \infty$, where θ_0 is the value of θ that minimizes the Kullback-Leibler divergence.

Proof: For any $\theta \neq \theta_0$ let us consider the log posterior odds relative to θ_0:

$$\ln\left(\frac{p(\theta \mid y)}{p(\theta_0 \mid y)}\right) = \ln\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \ln\left(\frac{f(y_i \mid \theta)}{f(y_i \mid \theta_0)}\right) \tag{4}$$

For fixed values of θ and θ_0, if the y_i's are iid draws from $g(y_i)$, the second term on the right is a sum of n iid random variables with a mean given by

$$E\left[\ln\left(\frac{f(y_i \mid \theta)}{f(y_i \mid \theta_0)}\right)\right] = KL(\theta_0) - KL(\theta) \leq 0.$$

Thus, as long as θ_0 is the unique minimizer of $KL(\theta)$, for $\theta \neq \theta_0$ the second term on the right of (4) is the sum of n iid random variables with negative mean. By the LLN, the sum approaches $-\infty$ as $n \to \infty$. As long as the first term on the right is finite (provided $p(\theta_0) > 0$), the whole expression approaches $-\infty$ in the limit. Then $\frac{p(\theta|y)}{p(\theta_0|y)} \to 0$, and so $p(\theta \mid y) \to 0$. Moreover, since all probabilities add up to 1, $p(\theta_0 \mid y) \to 1$.

If θ has a continuous distribution, $p(\theta_0 \mid y)$ is always zero for any finite sample, and so the previous argument does not apply, but it can still be shown that $p(\theta \mid y)$ becomes more and more concentrated about θ_0 as n increases. A statement of the theorem for the continuous case in Gelman et al is as follows.

[4] Equivalently, $\theta_0 = \arg\max_{\theta \in \Theta} E[\ln f(y \mid \theta)]$.

Theorem (continuous parameter space) If θ is defined on a compact set and A is a neighborhood of θ_0 with nonzero prior probability, then $\Pr(\theta \in A \mid y) \to 1$ as $n \to \infty$, where θ_0 is the value of θ that minimizes $KL(\theta)$.

Bernouilli example Recall that the posterior distribution in this case is:

$$p(\theta \mid y) \propto \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1} \sim Beta(m+\alpha, n-m+\beta)$$

with mean and variance given by[5]

$$E(\theta \mid y) = \frac{m+\alpha}{n+\alpha+\beta} = \frac{m}{n} + O\left(\frac{1}{n}\right)$$

$$Var(\theta \mid y) = \frac{(m+\alpha)(n-m+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)} = O\left(\frac{1}{n}\right).$$

As n increases the posterior distribution becomes concentrated at a value that does not depend on the prior distribution.

By the strong LLN, for each $\varepsilon > 0$[6]

$$\Pr\left(\lim_{n\to\infty}\left|\frac{m}{n} - \theta_0\right| < \varepsilon \mid \theta_0\right) = 1.$$

Therefore, with probability 1, the sequence of posterior probability densities

$$\lim_{n\to\infty} p(\theta \mid y) = \lim_{n\to\infty} Beta(n\theta_0 + \alpha, n(1-\theta_0) + \beta)$$

has a limit distribution with mean θ_0 and variance 0, independent of α and β. Thus, under each conjugate Beta prior, with probability 1, the posterior probability for θ converges to the Dirac Delta distribution concentrated on the true parameter value.

4.2 Asymptotic normality of the posterior distribution

We have seen that as $n \to \infty$ the posterior distribution converges to a degenerate measure at the true value θ_0 (posterior consistency). To obtain a non-degenerate limit, we consider the sequence of posterior distributions of $\gamma = \sqrt{n}\left(\theta - \widehat{\theta}\right)$, whose densities are given by[7]

$$p^*(\gamma \mid y) = \frac{1}{\sqrt{n}} p\left(\widehat{\theta} + \frac{1}{\sqrt{n}}\gamma \mid y\right).$$

[5] The mean and variance of $X \sim Beta(\alpha, \beta)$ are:

$$E(X) = \frac{\alpha}{\alpha+\beta}, \qquad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

[6] That is, $\int \mathbf{1}\left(\lim_{n\to\infty}\left|\frac{m}{n} - \theta_0\right| < \varepsilon\right)\theta_0^m(1-\theta_0)^{n-m}\,dm = 1$, for any θ_0.

[7] We use the ML estimator $\widehat{\theta}$ as the centering quantity, but the limiting result is unaffected if the posterior mode $\widetilde{\theta}$ is used instead or if $\gamma = \sqrt{n}(\theta - T_n)$ with $T_n = \theta_0 + [nI(\theta_0)]^{-1}[\partial \ln \mathcal{L}(\theta_0)/\partial\theta]$.

The basic result of large-sample Bayesian inference is that as more and more data arrive, the posterior distribution approaches a normal distribution. This result is known as the Bernstein-von Mises Theorem. See, for example, Lehmann and Casella (1998, Theorem 8.2, p. 489), van der Vaart (1998, Theorem 10.1, p. 141), or Chernozhukov and Hong (2003, Theorem 1, p. 305).

A formal statement for iid data and a scalar parameter, under the standard regularity conditions of MLE asymptotics, the condition that the prior $p(\theta)$ is continuous and positive in an open neighborhood of θ_0, and some additional technical conditions, is as follows:

$$\int \left| p^*(\gamma \mid y) - \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{1}{2\sigma_\theta^2}\gamma^2\right) \right| d\gamma \xrightarrow{p} 0.$$

where $\sigma_\theta^2 = 1/I(\theta_0)$. That is, the L_1 distance between the scaled and centered posterior and a $\mathcal{N}\left(0, \sigma_\theta^2\right)$ density centered at the random quantity γ goes to zero in probability. Thus, for large n, $p(\theta \mid y)$ is approximately a random normal density with random mean parameter $\widehat{\theta}$ and a constant variance parameter $I(\theta_0)^{-1}/n$:

$$p(\theta \mid y) \approx \mathcal{N}\left(\widehat{\theta}, \frac{1}{n} I(\theta_0)^{-1}\right).$$

The result can be extended to a multidimensional parameter. To gain intuition for this result let us consider a Taylor expansion of $\ln p(\theta \mid y)$ about the posterior mode $\widetilde{\theta}$:

$$\begin{aligned} \ln p(\theta \mid y) &\approx \ln p\left(\widetilde{\theta} \mid y\right) + \frac{\partial \ln p\left(\widetilde{\theta} \mid y\right)}{\partial \theta'}\left(\theta - \widetilde{\theta}\right) + \frac{1}{2}\left(\theta - \widetilde{\theta}\right)' \frac{\partial^2 \ln p\left(\widetilde{\theta} \mid y\right)}{\partial \theta \partial \theta'}\left(\theta - \widetilde{\theta}\right) \\ &= c - \frac{1}{2}\sqrt{n}\left(\theta - \widetilde{\theta}\right)' \left[-\frac{1}{n}\frac{\partial^2 \ln p\left(\widetilde{\theta} \mid y\right)}{\partial \theta \partial \theta'}\right] \sqrt{n}\left(\theta - \widetilde{\theta}\right) \end{aligned}$$

Note that $\partial \ln p\left(\widetilde{\theta} \mid y\right)/\partial \theta' = 0$. Moreover,

$$\frac{1}{n}\frac{\partial^2 \ln p\left(\widetilde{\theta} \mid y\right)}{\partial \theta \partial \theta'} = \frac{1}{n}\frac{\partial^2 \ln p\left(\widetilde{\theta}\right)}{\partial \theta \partial \theta'} + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f\left(y_i \mid \widetilde{\theta}\right)}{\partial \theta \partial \theta'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f\left(y_i \mid \widetilde{\theta}\right)}{\partial \theta \partial \theta'} + O\left(\frac{1}{n}\right) \approx -I\left(\widetilde{\theta}\right)$$

Thus, for large n the curvature of the log posterior can be approximated by the Fisher information:

$$\ln p(\theta \mid y) \approx c - \frac{1}{2}\sqrt{n}\left(\theta - \widetilde{\theta}\right)' I\left(\widetilde{\theta}\right) \sqrt{n}\left(\theta - \widetilde{\theta}\right).$$

Dropping terms that do not include θ we get the approximation

$$p(\theta \mid y) \propto \exp\left[-\frac{1}{2}\left(\theta - \widetilde{\theta}\right)' nI\left(\widetilde{\theta}\right)\left(\theta - \widetilde{\theta}\right)\right],$$

which corresponds to the kernel of a multivariate normal density $\mathcal{N}\left(\widetilde{\theta}, \frac{1}{n} I\left(\widetilde{\theta}\right)^{-1}\right)$.

Often, convergence to normality of the posterior distribution for a parameter θ can be improved by transformation. If ϕ is a continuous transformation of θ, then both $p(\phi \mid y)$ and $p(\theta \mid y)$ approach normal distributions, but the accuracy of the approximation for finite n can vary substantially with the transformation chosen.

A Bernstein-von Mises Theorem states that under adequate conditions the posterior distribution is asymptotically normal, centered at the MLE with a variance equal to the asymptotic frequentist variance of the MLE. From a frequentist point of view, this implies that Bayesian methods can be used to obtain statistically efficient estimators and consistent confidence intervals. The limiting distribution does not depend on the Bayesian prior.

4.3 Asymptotic behavior of the posterior in pseudo-likelihood models

If $g(y_i) \neq f(y_i \mid \theta)$ for all $\theta \in \Theta$, then the fitted model $f(y_i \mid \theta)$ is misspecified. In such case the large-n sampling distribution of the pseudo-ML estimator is

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_S)$$

where Σ_S is the sandwich covariance matrix:

$$\Sigma_S = \Sigma_M V \Sigma_M,$$

with $\Sigma_M = [-E(H_i)]^{-1} \equiv [I(\theta_0)]^{-1}$, $V = E(q_i q_i')$ and

$$q_i = \frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta}, \qquad H_i = \frac{\partial^2 \ln f(y_i \mid \theta_0)}{\partial \theta \partial \theta'}$$

In a correctly specified model the information identity holds $V = \Sigma_M^{-1}$ but in general $V \neq \Sigma_M^{-1}$.

The large sample shape of a posterior distribution obtained from $\Pi_{i=1}^n f(y_i \mid \theta)$ becomes close to

$$\theta \mid y \sim \mathcal{N}\left(\widehat{\theta}, \frac{1}{n}\Sigma_M\right).$$

Thus, misspecification produces a discrepancy between the sampling distribution of $\widehat{\theta}$ and the shape of the (pseudo)-likelihood. That is, the pseudo likelihood does not correctly reflect the sample information about θ contained in $\widehat{\theta}$. So, for the purpose of Bayesian inference about θ_0 (in the knowledge that θ_0 is only a pseudo-true value) it makes sense to start from the correct large-sample approximation to the likelihood of $\widehat{\theta}$ instead of the (incorrect) approximate likelihood of $(y_1 ... y_n)$. That is, to consider a posterior distribution of the form:

$$p\left(\theta \mid \widehat{\theta}\right) \propto \exp\left[-\frac{1}{2} n \left(\theta - \widehat{\theta}\right)' \Sigma_S^{-1} \left(\theta - \widehat{\theta}\right)\right] \times p(\theta) \tag{5}$$

This approach is proposed in Müller (2013) who shows that Bayesian inference about θ_0 is of lower asymptotic frequentist risk when the standard pseudo-posterior

$$p\left(\theta \mid \widehat{\theta}\right) \propto \exp\left[\sum_{i=1}^{n} \ln f(y_i \mid \theta)\right] \times p(\theta) \tag{6}$$

is substituted by the pseudo-posterior (5) that relies on the asymptotic likelihood of $\widehat{\theta}$ (an "artificial" normal posterior centered at the MLE with sandwich covariance matrix).

4.4 Asymptotic frequentist properties of Bayesian inferences

The posterior mode is consistent in repeated sampling with fixed θ as $n \to \infty$. Moreover, the posterior mode is also asymptotically normal in repeated samples. So the large-sample Bayesian statement holds

$$\left[I\left(\widetilde{\theta}\right)\right]^{1/2}\left(\theta-\widetilde{\theta}\right) \mid y \sim \mathcal{N}(0, I)$$

alongside the large-sample frequentist statement

$$\left[I\left(\widetilde{\theta}\right)\right]^{1/2}\left(\theta-\widetilde{\theta}\right) \mid \theta \sim \mathcal{N}(0, I).$$

See for example Lehmann and Casella (1998, Theorem 8.3, p. 490).

These results imply that in regular estimation problems the posterior distribution is asymptotically the same as the repeated sample distribution. So, for example, a 95% central posterior interval for θ will cover the true value 95% of the time under repeated sampling with any fixed true θ. The frequentist statement speaks of probabilities of $\widetilde{\theta}(y)$ whereas the Bayesian statement speaks of probabilities of θ. Specifically,

$$\Pr(\theta \leq r \mid y) = \int 1(\theta \leq r)\, p(\theta \mid y)\, d\theta \propto \int 1(\theta \leq r)\, f(y \mid \theta)\, p(\theta)\, d\theta$$

$$\Pr\left[\widetilde{\theta}(y) \leq r \mid \theta_0\right] = \int 1\left(\widetilde{\theta}(y) \leq r\right) f(y \mid \theta_0)\, dy$$

These results require that the true data distribution is included in the parametric likelihood family.

Bernoulli example continued The posterior mode corresponding to the beta prior with parameters (α, β) in (1) and the maximum likelihood estimator $\widehat{\theta} = m/n$ satisfy

$$\sqrt{n}\left(\widetilde{\theta}-\theta\right) = \sqrt{n}\left(\widehat{\theta}-\theta\right) + R_n$$

where

$$R_n = \frac{\sqrt{n}}{n+k}\left(\alpha - 1 - k\frac{m}{n}\right).$$

and $k = \alpha + \beta - 2$. Since $R_n \xrightarrow{p} 0$, it follows that $\sqrt{n}\left(\widetilde{\theta}-\theta\right)$ has the same asymptotic distribution as $\sqrt{n}\left(\widehat{\theta}-\theta\right)$, namely $\mathcal{N}[0, \theta(1-\theta)]$. Therefore, the normalized posterior mode has an asymptotic normal distribution, which is independent of the prior parameters and has the same asymptotic variance as that of the MLE, so that the posterior mode is asymptotically efficient.

Robustness to statistical principle and its failures The dual frequentist/Bayesian interpretation of many textbook estimation procedures suggests that it is possible to aim for robustness to statistical philosophies in statistical methodology, at least in regular estimation problems.

Even for small samples, many statistical methods can be considered as approximations to Bayesian inferences based on particular prior distributions. As a way of understanding a statistical procedure, it is often useful to determine the implicit underlying prior distribution (Gelman et al 2014, p. 92).

In the case of units roots the symmetry of Bayesian probability statements and classical confidence statements breaks down. With normal errors and a flat prior the Bayesian posterior is normal even if the true data generating process is a random walk (Sims and Uhlig 1991). Kim (1998) studied conditions for asymptotic posterior normality, which cover much more general situations than the normal random walk with flat priors.

5 Markov chain Monte Carlo methods

A Markov Chain Monte Carlo method simulates a series of parameter draws such that the marginal distribution of the series is (approximately) the posterior distribution of the parameters.

The posterior density is proportional to

$$p(\theta \mid y) \propto f(y \mid \theta)\, p(\theta).$$

Usually $f(y \mid \theta)\, p(\theta)$ is easy to compute. However, computation of point estimates and credible intervals typically requires the evaluation of integrals of the form

$$\frac{\int_{\Theta} h(\theta)\, f(y \mid \theta)\, p(\theta)\, d\theta}{\int_{\Theta} f(y \mid \theta)\, p(\theta)\, d\theta}$$

for various functions $h(.)$. For problems for which no analytic solution exists, MCMC methods provide powerful tools for evaluating these integrals, especially when θ is high dimensional.

MCMC is a collection of computational methods that produce an ergodic Markov chain with the stationary distribution $p(\theta \mid y)$. A continuous-state Markov chain is a sequence $\theta^{(1)}, \theta^{(2)}, ...$, that satisfies the Markov property:

$$\Pr\left(\theta^{(j+1)} \mid \theta^{(j)}, ..., \theta^{(1)}\right) = \Pr\left(\theta^{(j+1)} \mid \theta^{(j)}\right).$$

The probability $\Pr\left(\theta' \mid \theta\right)$ of transitioning from state θ to state θ' is called the transition kernel and we denote it $K\left(\theta' \mid \theta\right)$. Our interest will be in the steady-state probability distribution of the process.

Given a starting value $\theta^{(0)}$, a chain $\left(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}\right)$ is generated using a transition kernel with stationary distribution $p(\theta \mid y)$, which ensures the convergence of the marginal distribution of $\theta^{(M)}$ to $p(\theta \mid y)$. For sufficiently large M, the MCMC methods produce a dependent sample

$\left(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}\right)$ whose empirical distribution approaches $p(\theta \mid y)$. The ergodicity and construction of the chains usually imply that as $M \to \infty$,

$$\widehat{\theta} = \frac{1}{M} \sum_{j=1}^{M} h\left(\theta^{(j)}\right) \xrightarrow{p} \int_{\Theta} h(\theta)\, p(\theta \mid y)\, d\theta.$$

Analogously, a 90% interval estimation is constructed simply by taking the 0.05th and 0.95th quantiles of the sequence $\left(h\left(\theta^{(1)}\right), ..., h\left(\theta^{(M)}\right)\right)$.

In the theory of Markov chains one looks for conditions under which there exists an invariant distribution, and conditions under which iterations of the transition kernel $K\left(\theta' \mid \theta\right)$ converge to the invariant distribution. In the context of MCMC methods the situation is the reverse: the invariant distribution is known and in order to generate samples from it the methods look for a transition kernel whose iterations converge to the invariant distribution. The problem is to find a suitable $K\left(\theta' \mid \theta\right)$ that satisfies the invariance property:

$$p\left(\theta' \mid y\right) = \int K\left(\theta' \mid \theta\right) p(\theta \mid y)\, d\theta. \tag{7}$$

Under the invariance property, if $\theta^{(j)}$ is a draw from $p(\theta \mid y)$ then $\theta^{(j+1)}$ is also a draw from $p(\theta \mid y)$.

A useful fact is that the steady-state distribution $p(\theta \mid y)$ satisfies the detailed balance condition:

$$K\left(\theta' \mid \theta\right) p(\theta \mid y) = K\left(\theta \mid \theta'\right) p\left(\theta' \mid y\right) \text{ for all } \theta, \theta'. \tag{8}$$

The interpretation of equation (8) is that the amount of mass transitioning from θ' to θ is the same as the amount of mass that transitions back from θ to θ'.

The invariance property is not enough to guarantee that an average of draws $h\left(\theta^{(j)}\right)$ from $K\left(\theta' \mid \theta\right)$ converges to the posterior mean. It has to be proved that $K\left(\theta' \mid \theta\right)$ has a unique invariant distribution, that repeatedly drawing from $K\left(\theta' \mid \theta\right)$ leads to convergence to the unique invariant distribution regardless of the initial condition, and that the dependence of the draws $\theta^{(j)}$ decays sufficiently fast such that Monte Carlo sample averages converge to population means. Robert and Casella (2004) provide a textbook treatment of the convergence theory for MCMC algorithms.

Two general methods of constructing transition kernels are the Metropolis-Hastings algorithm and the Gibbs sampler, which we discuss in turn.

5.1 Metropolis-Hastings method

The Metropolis–Hastings algorithm proceeds by generating candidates that are either accepted or rejected according to some probability, which is driven by a ratio of posterior evaluations. A description of the algorithm is as follows.

Given the posterior density $f(y \mid \theta)\, p(\theta)$, known up to a constant, and a prespecified conditional density $q\left(\theta' \mid \theta\right)$ called the "proposal distribution", generate $\left(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}\right)$ in the following way:

1. Choose a starting value $\theta^{(0)}$.

2. Draw a proposal θ^* from $q\left(\theta^* \mid \theta^{(j)}\right)$.

3. Update $\theta^{(j+1)}$ from $\theta^{(j)}$ for $j = 1, 2...$, using

$$\theta^{(j+1)} = \begin{cases} \theta^* \text{ with probability } \rho\left(\theta^* \mid \theta^{(j)}\right), \\ \theta^{(j)} \text{ with probability } 1 - \rho\left(\theta^* \mid \theta^{(j)}\right), \end{cases}$$

where

$$\rho\left(\theta^* \mid \theta^{(j)}\right) = \min\left(1, \frac{f\left(y \mid \theta^*\right) p\left(\theta^*\right) q\left(\theta^{(j)} \mid \theta^*\right)}{f\left(y \mid \theta^{(j)}\right) p\left(\theta^{(j)}\right) q\left(\theta^* \mid \theta^{(j)}\right)}\right)$$

Some intuition for how the algorithm deals with a candidate transition from θ to θ' is as follows (Letham and Rudin 2012). If $p\left(\theta' \mid y\right) > p\left(\theta \mid y\right)$, then for every accepted draw of θ, we should have at least as many accepted draws of θ' and so we always accept the transition $\theta \to \theta'$. If $p\left(\theta' \mid y\right) < p\left(\theta \mid y\right)$, then for every accepted draw θ, we should have on average $\frac{p(\theta'|y)}{p(\theta|y)}$ accepted draws of θ'. We thus accept the transition with probability $\frac{p(\theta'|y)}{p(\theta|y)}$. Thus, for any proposed transition, we accept it with probability $\min\left[1, \frac{p(\theta'|y)}{p(\theta|y)}\right]$, which corresponds to $\rho\left(\theta' \mid \theta\right)$ when the proposal distribution is symmetric: $q\left(\theta' \mid \theta\right) = q\left(\theta \mid \theta'\right)$, as is the case in the original Metropolis algorithm.

The chain of draws so produced spends a relatively high proportion of time in the higher density regions and a lower proportion in the lower density regions. Because such proportions of times are balanced in the right way, the generated sequence of parameter draws has the desired marginal distribution in the limit. A key practical aspect of this calculation is that the posterior constant of integration is not needed since $\rho\left(\theta' \mid \theta\right)$ only depends on a posterior ratio.

Choosing a proposal distribution To guarantee the existence of a stationary distribution, the proposal distribution $q\left(\theta' \mid \theta\right)$ should be such that there is a positive density of reaching any state from any other state. A popular implementation of the M-H algorithm is to use the random walk proposal distribution:

$$q\left(\theta' \mid \theta\right) = \mathcal{N}\left(\theta, \sigma^2\right)$$

for some variance σ^2. In practice, one will try several proposal distributions to find out which is most suitable in terms of rejection rates and coverage of the parameter space.[8]

Other practical considerations include discarding a certain number of the first draws to reduce the dependence on the starting point (burn-in), and only retaining every $d-$th iteration of the chain to reduce the dependence between draws (thinning).

[8] See Letham and Rudin (2012) for examples of the practice of MCMC simulation using the OpenBUGS software package.

Transition kernel and convergence of the M-H algorithm The M-H algorithm describes how to generate a parameter draw $\theta^{(j+1)}$ conditional on a parameter draw $\theta^{(j)}$. Since the proposal distribution $q\left(\theta' \mid \theta\right)$ and the acceptance probability $\rho\left(\theta' \mid \theta\right)$ depend only on the current state, the sequence of draws forms a Markov chain. The M-H transition kernel can be written as

$$K\left(\theta' \mid \theta\right) = q\left(\theta' \mid \theta\right)\rho\left(\theta' \mid \theta\right) + r\left(\theta\right)\delta_\theta\left(\theta'\right). \tag{9}$$

The first term $q\left(\theta' \mid \theta\right)\rho\left(\theta' \mid \theta\right)$ is the density that θ' is proposed given θ, times the probability that it is accepted. To this we add the term $r\left(\theta\right)\delta_\theta\left(\theta'\right)$, which gives the probability $r\left(\theta\right)$ that conditional on θ the proposal is rejected times the Dirac delta function $\delta_\theta\left(\theta'\right)$, equal to one if $\theta' = \theta$ and zero otherwise. Here

$$r\left(\theta\right) = 1 - \int q\left(\theta' \mid \theta\right)\rho\left(\theta' \mid \theta\right)d\theta'.$$

If the proposal is rejected, then the algorithm sets $\theta^{(j+1)} = \theta^{(j)}$, which means that conditional on the rejection, the transition density contains a point mass at $\theta = \theta'$, which is captured by the Dirac delta function.

For the M-H algorithm to generate a sequence of draws from $p\left(\theta \mid y\right)$ a necessary condition is that the posterior distribution is an invariant distribution under the transition kernel (9), namely that it satisfies condition (7). See Lancaster (2004, p. 213) or Herbst and Schorfheide (2015) for proofs that $K\left(\theta' \mid \theta\right)$ satisfies the invariance property.

5.2 Gibbs sampling

The Gibbs sampler is a fast sampling method that can be used in situations when we have access to conditional distributions.

The idea behind the Gibbs sampler is to partition the parameter vector into two components $\theta = (\theta_1, \theta_2)$. Instead of sampling $\theta^{(j+1)}$ directly from $K\left(\theta \mid \theta^{(j)}\right)$, one first samples $\theta_1^{(j+1)}$ from $p\left(\theta_1 \mid \theta_2^{(j)}\right)$ and then samples $\theta_2^{(j+1)}$ from $p\left(\theta_2 \mid \theta_1^{(j+1)}\right)$. Clearly if $\left(\theta_1^{(j)}, \theta_2^{(j)}\right)$ is a draw from the posterior distribution, so is $\left(\theta_1^{(j+1)}, \theta_2^{(j+1)}\right)$ generated as above, so that the Gibbs sampler kernel satisfies the invariance property; that is, it has $p\left(\theta_1, \theta_2 \mid y\right)$ as its stationary distribution (see Lancaster 2004, p. 209).

The Gibbs sampler kernel is

$$K\left(\theta_1, \theta_2 \mid \theta_1', \theta_2'\right) = p\left(\theta_1 \mid \theta_2'\right)p\left(\theta_2 \mid \theta_1\right).$$

It can be regarded as a special case of Metropolis-Hastings where the proposal distribution is taken to be the conditional posterior distribution.

The Gibbs sampler is related to data augmentation. A probit model nicely illustrates this aspect (Lancaster 2004, Example 4.17, p. 211).

Bibliographical note

• A good textbook source on applied Bayesian methods is Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2014). Textbook treatments of Bayesian econometrics include Koop (2003), Lancaster (2004), Geweke (2005), and Greenberg (2012).

• Rothenberg (1973)'s Cowles Foundation Monograph 23 provides a classic discussion of the use of a priori information in frequentist and Bayesian approaches to econometrics.

• Herbst and Schorfheide (2015) provide an up-to-date account of applications of Bayesian inference to DSGE macro models.

• Arellano and Bonhomme (2011) review nonlinear panel data models, drawing on the link between random-effects approaches and Bayesian computation.

• Ruppert, Wand, and Carroll (2003) and Rossi (2014) discuss likelihood-based inference of flexible nonlinear models.

• Fiorentini, Sentana, and Shephard (2004) develop simulation-based Bayesian estimation methods of time-series latent variable models of financial volatility.

• The initial work on Bayesian asymptotics is due to Laplace. Further early work was done by Bernstein (1917) and von Mises (1931). Textbook sources are Lehmann and Casella (1998) and van der Vaart (1998). Chernozhukov and Hong (2003) provide a review of the literature.

• The Metropolis-Hastings (M–H) algorithm was developed by Metropolis et al in 1953 and generalized by Hastings in 1970, but was unknown to statisticians until the early 1990s. Tierney (1994) and Chib and Greenberg (1995) created awareness about the algorithm and stimulated their use in statistics.

• The name Gibbs sampler was introduced by Geman and Geman (1984) after the statistical physicist Willard Gibbs.

• Chib (2001) and Robert and Casella (1999) provide excellent treatments of MCMC methods.

• In models with moment restrictions, Chernozhukov and Hong (2003) propose using a GMM-like criterion function in place of the unknown likelihood to calculate quasi-posterior distributions by MCMC methods.

References

[1] Arellano, Manuel, and Stéphane Bonhomme (2011): "Nonlinear Panel Data Analysis", *Annual Review of Economics*, 3, 395–424.

[2] Bernstein, S. (1917): *Theory of Probability*, 4th Edition 1946. Gostekhizdat, Moscow–Leningrad (in Russian).

[3] Chernozhukov, Victor, and Han Hong (2003): "An MCMC approach to classical estimation", *Journal of Econometrics*, 115, 293–346.

[4] Chib, Siddhartha (2001): "Markov chain Monte Carlo methods: computation and inference". In: Heckman, J.J., Leamer, E. (Eds.), *Handbook of Econometrics*, Vol. 5. North-Holland, Amsterdam, 3564–3634 (Chapter 5).

[5] Chib, Siddhartha and Edward Greenberg (1995): "Understanding the Metropolis–Hastings Algorithm", *The American Statistician*, 49, 327–335.

[6] Fiorentini, Gabriele, Enrique Sentana, and Neil Shephard (2004): "Likelihood-Based Estimation of Latent Generalized ARCH Structures", *Econometrica*, 72(5), 1481–1517.

[7] Gelman, Andrew, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin (2014): *Bayesian Data Analysis*, Third Edition, CRC Press.

[8] Geman, Stuart and Donald Geman (1984): "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721-741.

[9] Geweke, John (2005): *Contemporary Bayesian Econometrics and Statistics*, John Wiley & Sons.

[10] Greenberg, Edward (2012): *Introduction to Bayesian econometrics*, Cambridge University Press.

[11] Hastings, W. K. (1970): "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, 57(1), 97–109.

[12] Herbst, Edward, and Frank Schorfheide (2015): *Bayesian Estimation of DSGE Models*, Princeton.

[13] Kim, Jae-Young (1998): "Large Sample Properties of Posterior Densities, Bayesian Information Criterion and the Likelihood Principle in Nonstationary Time Series Models", *Econometrica*, 66, 359–380.

[14] Koop, Gary (2003): *Bayesian Econometrics*, John Wiley & Sons.

[15] Lancaster, Tony (2004): *An Introduction to Modern Bayesian Econometrics*, Blackwell.

[16] Lehmann, E. L. and George Casella (1998): *Theory of Point Estimation*, Second Edition, Springer.

[17] Letham, Ben and Cynthia Rudin (2012): "Probabilistic Modeling and Bayesian Analysis", Prediction, Machine Learning, and Statistics Lecture Notes, Sloan School of Management, MIT.

[18] Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953): "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*, 21, 1087–1092.

[19] Müller, Ulrich (2013): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix", *Econometrica*, 81(5), 1805–1849.

[20] Robert, C.P. and George Casella (1999): *Monte Carlo Statistical Methods*. Springer, Berlin.

[21] Rossi, Peter E. (2014): *Bayesian Non- and Semi-parametric Methods and Applications*, Princeton University Press.

[22] Rothenberg, Thomas (1973): *Efficient Estimation with A Priori Information*, Cowles Foundation Monograph 23, Yale University Press.

[23] Ruppert, David, M. P. Wand, and R. J. Carroll (2003): *Semiparametric Regression*, Cambridge University Press.

[24] Sims, Christopher A. and Harald Uhlig (1991): "Understanding Unit Rooters: A Helicopter Tour" *Econometrica*, 59(6), 1591–1599.

[25] Tierney, Luke (1994): "Markov Chains for Exploring Posterior Distributions" (with discussion), *Annals of Statistics*, 22, 1701–1762.

[26] van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press.

[27] von Mises, Richard (1931): *Wahrscheinlichkeitsrechnung*. Springer, Berlin (*Probability Theory*, in German).