H

Hierarchical Bayes Models

Siddhartha Chib and Edward Greenberg

Abstract

The standard Bayesian model is defined in terms of an outcome model and the prior density of the parameters. The latter depends on parameters called hyperparameters. A hierarchical Bayes model results when one or more of the hyperparameters are assumed to be random and modelled probabilistically. We discuss canonical versions of these models for the case when both the parameters and the hyperparameters are modelled in groups or blocks, provide relevant examples, and discuss how inference by Markov chain Monte Carlo methods makes even the fitting of complex hierarchical models practical and simple. The problem of model comparisons is also addressed.

Keywords

Bayes' th; Component densities; Exchangeability; Hierarchical Bayes models; Hyperparameters; Marginal likelihood; Markov chain Monte Carlo methods

JEL Classifications

C11

Suppose that y is a univariate random variable or multivariate random vector and θ is a d-dimensional parameter vector that lies in $\mathscr{D}$, a subset of $\mathscr{R}^d$. The standard Bayesian model is then defined in terms of the density of y given θ (the outcome model) and the prior density of θ (the prior model). Specifically, the Bayesian model is specified as

$$y|\theta \sim p(y|\theta) \text{ (outcome model : stage 1)} \quad (1)$$

$$\theta|\gamma \sim \pi(\theta|\gamma) \text{ (prior model : stage 2)} \quad (2)$$

where γ is the vector of parameters in the prior density. These are called hyperparameters. We can assume that γ is g-dimensional and lies in $\mathscr{G}$, a subset $\mathscr{R}^g$. The labelling of the outcome model as stage 1 and the prior model as stage 2 is arbitrary, and the numbering can be reversed. The outcome model may be called the top or bottom level of the model because this difference in nomenclature has no significance.

Suppose that the researcher is not able to specify one or more of the hyperparameters in γ. In that case, the unknown hyperparameters can be

This chapter was originally published in The New Palgrave Dictionary of Economics, 2nd edition, 2008. Edited by Steven N. Durlauf and Lawrence E. Blume

Palgrave Macmillan (ed.), *The New Palgrave Dictionary of Economics*,
DOI 10.1057/978-1-349-95121-5_2230-1

assumed to be random and modelled probabilistically. This modelling of the hyperparameters leads to what is called a Bayesian hierarchical model (Berger 1985; Lehmann and Casella 1998). The simplest version of a Bayesian hierarchical model is defined in terms of the ingredients

$$y|\theta \sim p(y|\theta) \text{ (outcome model : stage 1)} \quad (3)$$

$$\theta|\gamma \sim \pi(\theta|\gamma) \text{ (prior model : stage 2)} \quad (4)$$

$$\gamma|\lambda \sim \psi(\gamma|\lambda) \text{ (hyperparameter model : stage 3)}, \quad (5)$$

where $\psi(\gamma|\lambda)$ is the prior density of γ. The hyperparameters λ in the stage 3 model are assumed known. In effect, a hierarchical model is a way of modelling the outcomes and the parameters through a sequence of easily interpretable steps.

In practice, it is often helpful to divide θ into natural groups or blocks $(\theta_1; \theta_2, \ldots, \theta_p)$, where, for instance, θ_1 consists of the regression coefficients, θ_2 the scale parameters and θ_p the covariance parameters. Each of these separate blocks may then be modelled independently in terms of prior densities $\pi(\theta_j|\gamma)$. In turn, γ may also be grouped into blocks $(\gamma_1,\ldots, \gamma_q)$ and, in the third stage, modelled independently through the densities $\psi(\gamma_j|\lambda)$. The resulting three-stage hierarchical model then has the form

$$y|\theta \sim p(y|\theta) \text{ (outcome model : stage 1)} \quad (6)$$

$$\theta|\gamma \sim \prod_{j=1}^{p} \pi\left(\theta_j|\gamma\right) \text{ (prior model : stage 2)} \quad (7)$$

$$\gamma|\lambda \sim \prod_{j=1}^{q} \psi\left(\gamma_j|\lambda\right) \text{ (hyperparameter model : stage 3)}. \quad (8)$$

This specification may be considered as the canonical hierarchical Bayes model.

Example 1 (Gaussian linear regression model). *Suppose that $y = (y_1,\ldots,y_n)$ is a vector of observations and θ consists of the two blocks (β, σ^2), where β is a k-vector of regression parameters. Now let*

$$y|\theta \sim N_n(y|X\beta, \sigma^2 I_n)$$
$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0)IG\left(\sigma^2|\frac{v_0}{2}, \frac{\delta_0}{2}\right),$$

where

$$N_k(\beta|\beta_0, B_0) = (2\pi)^{-k/2}\exp\left\{-\frac{1}{2}(\beta-\beta_0)B_0^{-1}(\beta-\beta_0)\right\}$$

is the k-variate normal density, X is the $n \times k$ matrix of covariates and

$$IG\left(\sigma^2|\frac{v_0}{2}, \frac{\delta_0}{2}\right) = \frac{(\delta_0/2)^{(v_0/2)}}{\Gamma(v_0/2)}\left(\frac{1}{\sigma^2}\right)^{(v_0/2)+1} \exp\left(-\frac{\delta_0}{2\sigma^2}\right), \sigma^2 > 0$$

is the inverse-gamma density. In this case, the hyperparameters γ consist of the four blocks of parameters $(\beta_0, B_0, v_0, \delta_0)$. The top level of the model is the model of the outcome and the bottom level the model of θ. If it is not possible to fix the value of β_0, for example, one may specify a prior, $\beta_0|\lambda \sim N_k(\beta_0|\beta_{00}, B_{00})$, where the hyperparameters of the third stage $\lambda = (\beta_{00}, B_{00})$ are pre-specified. Further discussion along these lines is provided by Lindley and Smith (1972).

Since the difficulty of specifying hyperparameters in the second stage model of the model arises in almost all applications, hierarchical Bayes modelling is of special interest and importance in Bayesian analysis. To further fix the ideas, the following example, which we develop further below, is instructive and should be studied carefully.

Example 2 *(Gaussian clustered data model). Clustered data arise when n observations are available for each subject i ($i \leq n$) in the sample. For example, in the panel or longitudinal set-up, there are observations across time for each subject. Let the observations on the ith subject be denoted by $y_i = (y_{i1}, \ldots, y_{in_i})$. Assume that the*

observations are continuous. Binary or ordinal responses can be dealt with in much the same way by adopting the framework of Albert and Chib (1993). The data for all n subjects are collected in the vector $y = (y_1, \ldots, y_n)$. *It is common in this context to allow for unique cluster-specific effects. Let* $W_i = (w_{i1}, \ldots, w_{in_i})'$ *be a* $n_t \times q$ *matrix of observations on q covariates* w_{ij} *whose effect on y is assumed to be cluster-specific. Also suppose that* X_{1i} *is an additional* $n_i \times k_1$ *matrix of observations on* k_1 *covariates whose effect on y is assumed to be non-cluster-specific (fixed effect). Then under the assumption that the observations across clusters are independent, a model for the outcomes is*

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}\left(y_i|X_{i1}\beta_1 + W_i\beta_{2i}, \sigma^2 I_{n_i}\right),$$

where the β_{2i} *are the cluster-specific effects. If the numbers of clusters is large, as is usual in practice, it is useful to assume that the effects* β_{2i} *have some structure. One possibility is to assume that the* β_{2i} *are drawn from a common distribution*

$$\beta_{2i}|\gamma \sim N_q(\beta_2, D)$$

independently across i. This is called the exchangeability assumption since the joint distribution of the β_{2i} *is invariant to permutation of the indices. Another possibility is the assumption that the* β_{2i} *are determined by a set of r cluster-specific covariates* a_i*:*

$$\beta_{2i}|\gamma \sim N_q(A_i\beta_2, D)$$

where

$$A_i = \begin{pmatrix} a_i' & 0' & \ldots & \ldots & 0' \\ 0' & a_i' & \ldots & \ldots & 0' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0' & 0' & \ldots & \ldots & a_i' \end{pmatrix},$$

$\beta_2 = \beta_{21}, \beta_{22}, \ldots, \beta_{2q})$ *is a* $k_2 = r \times q$*-dimensional vector and D, as in the first example, is a* $q \times q$ *matrix. Writing the second stage model in equivalent form as* $\beta_{2i} = A_i\beta_2 + b_i$, *where* $b_i|D \sim N_q(0, D)$, *and substituting this into the outcome model, it follows that the outcome model can be expressed as*

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}\left(y_i|X_i\beta + W_ib_i, \sigma^2 I_{n_i}\right),$$

where $\theta = (\beta, \sigma^2, b_1, \ldots, b_n)$, $X_i = (X_{1i} : W_iA_i)$ *is a* $n_i \times k$ *matrix* $(k = k_1 + k_2)$ *and* $\beta = (\beta_1, \beta_2)$. *The second stage of the model could now be specified as*

$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0)IG\left(\sigma^2|v_0/2, \delta_0/2\right)\prod_{i=1}^{n} N_q(b_i|0, D).$$

Next suppose that there is enough prior information to fix $(\beta_0, B_0, v_0, \delta_0)$, *but that D (equivalently* D^{-1}*) cannot be fixed directly. Then* $\gamma = D^{-1}$. *A convenient assumption is*

$$\gamma|\lambda \sim \text{Wishart}_q\left(D^{-1}|\rho_0, R_0\right),$$

where

$$\text{Wishart}_q\left(D^{-1}|\rho_0, R_0\right) = c\frac{\left|D^{-1}\right|^{(\rho_0-q-1)/2}}{|R_0|^{\rho_0/2}} \exp\left\{-\frac{1}{2}\text{trace}\ \left(R^{-1}D^{-1}\right)\right\}, |D^{-1}| > 0,$$

is the q-variate Wishart density,

$$c = \left(2^{\rho_0 q/2}\pi^{q(q-1)/4}\prod_{i=1}^{T}\Gamma\left(\frac{\rho_0 + 1 - i}{2}\right)\right)^{-1}$$

is its normalizing constant, and the stage 3 hyperparameters $\lambda = (\rho_0, R_0)$ *are known. Under these assumptions the full model is given by*

$$y|\theta \sim \prod_{i=1}^{n} N_{n_i}\left(y_i|X_i\beta + W_ib_i, \sigma^2 I_{n_i}\right) \tag{9}$$

$$\theta|\gamma \sim N_k(\beta|\beta_0, B_0)IG\left(\sigma^2|v_0/2, \delta_0/2\right) \prod_{i=1}^{n} N_q(b_i|0, D) \tag{10}$$

$$\gamma|\lambda \sim \text{Wishart}_q\left(D^{-1}|\rho_0, R_0\right). \tag{11}$$

Putting a prior distribution on the

hyperparameters γ in this way has several advantages. For one, it produces a prior distribution on θ that is less dogmatic than a prior based on specified hyperparameters since the resulting prior distribution of θ is averaged over the possible values of γ as dictated by the density $\psi(\gamma|\lambda)$:

$$\pi(\theta|\lambda) = \int \pi(\theta|\gamma)\psi(\gamma|\lambda)\mathrm{d}\gamma.$$

If the hyperparameter γ is a scalar discrete quantity with support on the set $\{\gamma_1,\ldots,\gamma_G\}$, where G is potentially infinite, then the mixing density $\psi(\gamma \mid \lambda)$ is a probability mass function of the type $\sum_{j=1}^{G} p_j \delta_{\gamma_j}$, where δ_{γ_j} is the indicator function of γ_j, $0 \le p_j \le 1$ and $\sum_{j=1}^{G} p_j = 1$. The resulting conditional density $\pi(\theta|\lambda)$ is then a mixture of densities of the form

$$\pi(\theta|\lambda) = \sum_{j=1}^{G} p_j \pi\left(\theta|\gamma_j\right).$$

In this context, $\pi(\theta|\gamma_j)$ are called the component densities and p_j are the component weights. Such mixtures of component densities provide a simple mechanism for modelling θ in a flexible way.

Of course, one could have started at the outset with the prior $\pi(\theta|\ \lambda)$ by combining stages 2 and 3, leading to the collapsed model

$$y|\theta \sim p(y|\theta) \tag{12}$$

$$\theta|\lambda \sim \int \pi(\theta|\gamma)\psi(\gamma|\lambda)\ \mathrm{d}\gamma, \tag{13}$$

which has the same structure as the standard 2-stage Bayesian model. This is not done, however, because the density of $\theta|\ \lambda$, even if tractable, is generally less easy to manage.

Example 3 *(Gaussian linear regression model and Student-t prior). Suppose that $y = (y_1, \ldots, y_n)$ is a vector of observations and $\theta = (\beta, \sigma^2)$, where β is a scalar regression parameter. Assume that*

$$y|\theta \sim N_n(y|X\beta, \sigma^2 I_n)$$
$$\theta|\gamma \sim N(\beta|\beta_0, B_0) IG\left(\sigma^2|\frac{v_0}{2}, \frac{\delta_0}{2}\right)$$
$$B_0^{-1} \sim G\left(B_0^{-1}|\frac{v}{2}, \frac{v}{2}\right)$$

where $G(\cdot \mid \cdot, \cdot)$ is the gamma density and the quantities (β_0, v_0, δ_0) and v are known. Then the density of β marginalized over B_0^{-1} is Student-t, T ($\beta|\beta_0$, 1, v), with location β_0, dispersion 1 and v degrees of freedom. This Student-t prior density is not conjugate with the outcome model and therefore cumbersome to deal with.

Bayesian hierarchical models can have additional stages. For instance, a further stage can be added by placing a prior density on λ, which leads to the model

$$y|\theta \sim p(y|\theta)\ \ (\text{outcome}\ \ \text{model}\ :\ \text{stage}\ \ 1) \tag{14}$$

$$\theta|\gamma \sim \prod_{j=1}^{p} \pi\left(\theta_j|\gamma\right)\ (\text{prior model} : \text{stage}\ \ 2) \tag{15}$$

$$\gamma|\lambda \sim \prod_{j=1}^{q} \psi\left(\gamma_j|\lambda\right)\ (\text{hyperparameter model} : \text{stage}\ 3). \tag{16}$$

$$\lambda \sim \delta(\lambda)\ \ (\text{hyperparameter}\ \ \text{model}\ \ 2\ :\ \ \text{stage}\ \ 4), \tag{17}$$

where δ is the density of λ. Models with more than four stages are rare.

Posterior Distributions

In a Bayesian analysis one is interested in deriving and summarizing the posterior distribution of θ given y. One obvious question concerns the form of this posterior distribution. Another question concerns the posterior distribution of the hyperparameters γ. Consider the canonical three-stage hierarchical model in (6)–(8). By Bayes's theorem,

$$\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{m(y)},$$

where $n(\theta) = \int \pi(\theta|\gamma)\pi(\gamma|\lambda)\ \mathrm{d}\gamma$ and $m(y) = \int p(y|\theta)\pi(\theta)\ \mathrm{d}\theta$, called the marginal likelihood, is the normalizing constant. Similarly, the posterior distribution of γ is

$$\pi(\gamma|y) = \frac{p(y|\gamma)\pi(\gamma|\lambda)}{m(y)},$$

where $p(y|\gamma) = \int p(y|\theta)\pi(\theta|\lambda)\ \mathrm{d}\theta$. Before we discuss the tractability of these distributions we state a general result about how much information the data y supply about θ and γ beyond what is introduced by the prior densities $\pi(\theta)$ and $\pi(\gamma|\lambda)$. To measure this information we can use the Kullback–Leibler (KL) divergence measure, which, for any two densities f and g, is defined as

$$K(f,g) = E^f \log\frac{f}{g},$$

where E^f is the expectation with respect to the density f. The following result was proved by Goel and Degroot (1981). The result and proof can also be found in Lehmann and Casella (1998).

Theorem 1 *For the three-stage hierarchical model,*

$$K[\pi(\gamma|y), \pi(\gamma)] < K[\pi(\theta|y),\ \pi(\theta)].$$

This result states that the KL divergence between $\pi(\theta|y)$ and $\pi(\theta)$ is greater than between $\pi(\gamma|y)$ and $\pi(\gamma)$. In other words, the data supply more information about θ than they do about γ. Equivalently, the prior and the posterior of γ are closer than the prior and the posterior of θ. This implies that less learning is possible about the hyperparameters γ than about the parameters θ.

Much less can be said about the form of the posterior densities. In general, the posterior densities $\pi(\theta|y)$ and $\pi(\gamma|y)$ are not tractable. But if we consider the density of θj given (y,γ) and $\theta_{-j} = (\theta_1,\ldots, \theta_{j-1}, \theta_{j+1},\ldots,\theta_p)$, we have

$$\pi(\theta_j|y,\gamma,\theta_{-j}) \propto p(y|\theta)\pi(\theta_j|\gamma),$$

which is in closed form provided the prior density $\pi(\theta_j|\gamma)$ is conjugate with $p(y|\theta)$. The density $\pi(\theta_j|y, \gamma, \theta_{-j})$ is called the full conditional density of θ_j. Of course, the marginal density,

$$\pi(\theta_j|y) = \int \pi(\theta_j|y,\gamma,\theta_{-j}|y)\ \mathrm{d}\gamma\ \ \mathrm{d}\theta_{-j},$$

where the mixing distribution is the marginal posterior distribution of (γ, θ_{-j}), is almost never available in closed form.

The same sort of difficulty arises in finding $\pi(\gamma|y)$. The problem is that the prior $\pi(\gamma|\lambda)$ generally does not combine with $p(y|\gamma)$ to produce a recognizable density. Nonetheless, just as in the case of θj, the calculations are easier if one considers the full conditional density of γ_j. To see this, note that

$$\begin{aligned}\pi(\gamma_j|y,\theta,\gamma_{-j}) &\propto p(y|\theta)\pi(\theta|\gamma)\pi(\theta_j|\lambda)\\ &\propto \pi(\theta|\gamma)\pi(\theta_j|\lambda),\end{aligned}$$

where the second line follows from the fact that the outcome model in stage 1 is free of γ. Thus, provided $\pi(\theta_j|\lambda)$ is conjugate with $\pi(\theta|\gamma)$, the full conditional density of γ_j can be derived in closed form.

Example 4 *Consider again the clustered data model given in (9)–(11). The full conditional density of b_i is obtained as*

$$\begin{aligned}\pi(b_i|y,\theta_{-b_i},\gamma) &\propto p(y|\theta)\pi(b_i|\gamma)\\ &\propto N_{n_i}(y_i|X_i\beta + W_i b_i, \sigma^2 I_{n_i})N_q(b_i|0,D),\end{aligned}$$

which, by standard Bayesian manipulations, is seen to be a $N_q\left(b_i|\widehat{b}_i, B_i\right)$ density, where

$$\widehat{b_i} = B_i(\sigma^{-2}W_i'(y_i - X_i\beta) \quad \text{and}$$
$$B_i = \left(D^{-1} + \sigma^{-2}W_i'\,W_i\right)^{-1}.$$

Turning now to the full conditional density of D^{-1}*, we obtain*

$$\pi\left(D^{-1}|y,\theta\right) = \pi\left(D^{-1}|\{b_i\}\right) \propto \pi\left(\{b\}|D^{-1}\right)$$
$$\pi\left(D^{-1}|\lambda\right) \propto \prod_{i=1}^{n} Nq(b_i|0,D)\ \text{Wishart}_q\left(D^{-1}|v_0,R_0\right)$$
$$= \text{Wishart}_q\left(D^{-1}|\rho_0+n,\left(R_0^{-1}+\sum_{i=1}^{n} b_i b_i'\right)^{-1}\right),$$

where in the first line we have used the fact that the full conditional density of D^{-1} *depends neither on y nor on β; in the second line, Bayes's theorem; in the third line, substitutions for the needed densities; and in the fourth line, by observation that the product of the normal and Wishart prior densities is an updated Wishart distribution with the stated parameters.*

Computational Issues

Difficulties in the computation of the marginal posterior densities of θ_j and γ_j were previously an impediment to the development and application of hierarchical Bayesian models. These difficulties have largely been resolved through the use of Markov chain Monte Carlo (MCMC) methods. These methods typically proceed by simulating the full conditional distributions, $\pi(\theta_j|y,\ \gamma,\ \theta_{-j})$ and $\pi(\gamma_j|y,\ \theta,\ \gamma_{-j})$. Under general conditions, the recursive simulation of these distributions produces a Markov chain whose limiting invariant distribution is the posterior density of interest, $\pi(\theta;\gamma\,|y)$.

Although it is not possible in this discussion to provide the theory behind MCMC methods, as outlined in Tierney (1994), and Chib and Greenberg (1995), or the range of hierarchical Bayes models that have been thus processed, it is useful to illustrate the computations with the help of the simplest MCMC method, the so-called Gibbs sampling algorithm. This algorithm was introduced by Geman and Geman (1984) in the context of image processing, but the papers of Tanner and Wong (1987) and Gelfand and Smith (1990) brought it into the limelight.

Suppose that the various blocks $\{\theta_j\}$ and $\{\gamma_j\}$ are chosen to ensure that the associated set of full conditional densities $\{\pi(\theta_j|y,\ \theta_{-j},\ \gamma)\}$ and $\{\pi(\gamma_j|y,\ \theta,\ \gamma_{-j})\}$ are all tractable. Then one cycle of the Gibbs sampling algorithm is completed by simulating $\{\theta_j\}$ and $\{\gamma_j\}$ from each full conditional distribution, recursively updating the conditioning variables while moving through the set of distributions. The Gibbs sampler in which each block is revised in fixed order is defined as follows.

Algorithm: Gibbs Sampling

1. Specify an initial value $\theta^{(0)} = \left(\theta_1^{(0)},\ldots,\theta_p^{(0)}\right)$ and $\gamma^{(0)} = \left(\gamma_1^{(0)},\ldots,\gamma_q^{(0)}\right)$.
2. Repeat for $j = 1,2,\ldots,n_0 + M$:
   Generate $\theta_1^{(j)}$ from $\pi\left(\theta_1|y,\theta_2^{(j-1)},\ldots,\theta_p^{(j-1)},\gamma^{(j-1)}\right)$
   Generate $\theta_2^{(j)}$ from $\pi\left(\theta_2|y,\theta_1^{(j)},\theta_3^{(j-1)}\ldots,\theta_p^{(j-1)},\gamma^{(j-1)}\right)$
   $\vdots$
   Generate $\theta_p^{(j)}$ from $\pi\left(\theta_p|y,\theta_1^{(j)},\theta_2^{(j)}\ldots,\theta_{p-1}^{(j)},\gamma^{(j-1)}\right)$
   Generate $\gamma_1^{(j)}$ from $\pi\left(\gamma_1|y,\theta^{(j)},\gamma_2^{(j-1)},\ldots,\gamma_q^{(j-1)}\right)$
   Generate $\gamma_2^{(j)}$ from $\pi\left(\gamma_2|y,\theta^{(j)},\gamma_1^{(j)},\gamma_3^{(j-1)}\ldots,\gamma_q^{(j-1)}\right)$
   $\vdots$
   Generate $\gamma_q^{(j)}$ from $\pi\left(\gamma_q|y,\theta^{(j)},\gamma_1^{(j)},\ldots,\gamma_{q-1}^{(j)}\right)$.
3. Return the values $\left\{\theta^{(n_0+1)},\gamma^{(n_0+1)},\ \theta^{(n_0+2)},\ \gamma^{(n_0+2)},\ldots,\theta^{(n_0+M)},\gamma^{(n_0+M)}\right\}$.

Thus, in this algorithm, block θ_j is generated from the full conditional distribution

$$\pi\left(\theta_j|y,\theta_1^{(j)},\ldots,\theta_{j-1}^{(j)},\theta_{j+1}^{(j-1)},\ldots,\theta_p^{(j-1)},\gamma^{(j-1)}\right),$$

where the conditioning elements for the jth block reflect the fact that the previous $(j-1)$ blocks of θ have already been updated, but the rest have not been. Note that the output from the first n_0 cycles (the burn-in phase) is ignored to allow the effect of the initial values to wear off. One additional point about MCMC methods is that the blocks must be carefully chosen. Sampling over unnecessary blocks can worsen the quality of the output produced by the algorithm, where quality is measured by how quickly the serial correlations of the sampled draws decline to zero. Chains whose serial correlations decline quickly are preferred because they are closer to the ideal of independent sampling.

Example 5 *Consider again the hierarchical Bayesian model for clustered data given in (9)–(11). The joint distribution of the data and the unknowns is given by*

$$\begin{aligned} p\left(y,\theta,D^{-1}\right) &= \pi\left(\beta,\sigma^2,\{b_i\},D^{-1}\right)p(y|\theta) \\ &= \pi(\beta)\pi\left(\sigma^2\right)\pi\left(D^{-1}\right)\sum_{i=1}^{n}p(y_i|\theta)\pi(b_i|D). \end{aligned} \tag{18}$$

Wakefield et al. (1994) propose a Gibbs MCMC approach for joint distribution that is based on full blocking (that is, sampling each block of parameters from its full conditional distribution). Chib and Carlin (1999) suggest a number of reduced blocking schemes. One of the simplest proceeds by first sampling β marginalized over $\{b_i\}$ and then sampling $\{b_i\}$ conditioned on β. This reduced blocking is possible because b_i in (18) can be marginalized out leaving a normal distribution that can be combined with the assumed normal prior on β. In particular,

$$\begin{aligned} p\left(y_i|\beta,\sigma^2,D\right) &= \int p(y_i|\theta)\pi(b_i|D)\mathrm{d}b_i \propto |V_i|^{-1/2} \\ &\exp\left\{(-1/2)(y_i-X_i\beta)'V_i^{-1}(y_i-X_i\beta)\right\}, \end{aligned}$$

where $V_i = \sigma^2 I_{n_i} + W_i D W_i'$. The reduced conditional posterior of β is therefore

$$\begin{aligned} \pi(\beta|y,\sigma^2,D) &\propto \pi(\beta)\prod_{i=1}^{n}|V_i|^{-1/2} \\ &\exp\left\{-\frac{1}{2}(y_i-X_i\beta)'V_i^{-1}(y_i-X_i\beta)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\beta-\widehat{\beta}\right)'B^{-1}\left(\beta-\widehat{\beta}\right)\right\}, \end{aligned}$$

where

$$\widehat{\beta} = B\left(B_0^{-1}\beta_0 + \sum_{i=1}^{n}X_i'\ V_i^{-1}y_i\right) \quad \text{and}$$

$$B = \left(B_0^{-1} + \sum_{i=1}^{n}X_i'\ V_i^{-1}X_i\right)^{-1}.$$

The rest of the MCMC algorithm follows the steps of Wakefield et al. (1994). In full, we sequentially sample the following distributions many times:

$$\beta \sim N_k\left(\widehat{\beta},B\right)$$

$$b_i \sim N_q\left(D_i(\sigma^{-2}W_i'\ (y_i - X_i\beta),\right.$$

$$\left.D_i = \left(D^{-1}+\sigma^{-2}W_i'\ W_i\right)^{-1},\right),\quad i \le n$$

$$\sigma^2 \sim IG\left(\frac{v_0+\Sigma n_i}{2},\frac{\delta_0+\Sigma_{i=1}^{n}||y_i-X_i\beta-W_ib_i||^2}{2}\right)$$

$$D^{-1} \sim \text{Wishart}_q\left\{\rho_0+n,\left(R_0^{-1}+\sum_{i=1}^{n}b_ib_i'\right)^{-1}\right\},$$

where the second and fourth of these distributions were derived in Example 4.

Model Choice

Another inferential concern in practice is the comparison of several hierarchical Bayesian models in order to judge the extent to which the various models are supported by the data. In the context of a hierarchical model for clustered data, for instance, one may be interested in determining the support for an additional cluster-specific effect or of an additional fixed effect. Questions of this type can be answered via *Bayes factors,* or ratios

of *marginal likelihoods.* The marginal likelihood of a particular model $\mathscr{M}$ is the normalizing constant of the posterior density,

$$m(y|\mathscr{M}) = \int p(y|\mathscr{M},\theta)\pi(\theta|\mathscr{M},\gamma)\pi(\gamma|\mathscr{M},\lambda)\mathrm{d}\theta\ \ \mathrm{d}\gamma, \tag{19}$$

the integral of the first stage outcome density function with respect to the prior density of θ and the prior density of the hyperparameters γ. If there are two models $\mathscr{M}_k$ and $\mathscr{M}_l$, the Bayes factor is the ratio

$$\mathrm{B}_{kl} = \frac{m(y|\mathscr{M}_k)}{m(y|\mathscr{M}_l)}. \tag{20}$$

Because MCMC methods deliver draws from the posterior density and the marginal likelihood is the integral with respect to the prior$\pi(\theta|\mathscr{M},\gamma)$ $\pi(\gamma|\mathscr{M},\lambda)$, MCMC output cannot be used directly to average $p(y|\mathscr{M},\theta)$. Nonetheless, computation is feasible by the method of Chib (1995), a widely used method that we now briefly describe.Chib (1995) begins by noting that $m(y|\mathscr{M},\lambda)$ can be expressed as

$$m(y|\mathscr{M}) = \frac{p(y|\mathscr{M},\theta^*)\pi(\theta^*|\mathscr{M},\gamma^*)\pi(\gamma^*|\mathscr{M},\lambda)}{\pi(\theta^*,\gamma^*|\mathscr{M},y)}, \tag{21}$$

for a given (θ^*, γ^*), usually taken to be a high density point such as the posterior mean. Thus, if we have an estimate $\widehat{\pi}(\theta^*,y^*|\mathscr{M},y)$ of the posterior ordinate, the marginal likelihood on the log scale can be estimated as

$$\begin{aligned}\log m(y|\mathscr{M}) &= \log p(y|\mathscr{M},\theta^*) + \log\pi(\theta^*|\mathscr{M},\gamma^*)\\ &+ \log\pi(\gamma^*|\mathscr{M},\lambda) - \log\widehat{\pi}(\theta^*,\ \gamma^*|\mathscr{M},y).\end{aligned} \tag{22}$$

It turns out that it is possible to get an efficient estimate of the posterior ordinate. The basic idea is to write the posterior ordinate as

$$\begin{aligned}\pi(\theta^*,\gamma^*|\mathscr{M},y) &= \pi\left(\theta_1^*|\mathscr{M},y\right)\times\cdots\\ \times\pi\left(\theta_p^*|\mathscr{M},y,\theta_1^*,\ldots,\theta_{p-1}^*\right)&\times\pi\left(\gamma_1^*|\mathscr{M},y,\theta^*\right)\times\cdots\\ \times\pi\left(\gamma_q^*|\mathscr{M},y,\theta^*,\gamma_1^*,\ldots,\gamma_{p-1}^*\right)&\end{aligned} \tag{23}$$

and then to estimate each of these ordinates from the output of appropriate MCMC runs. To see what is involved, consider the ordinate $\pi\left(\theta_j^*|\mathscr{M},y,\theta_1^*,\ldots,\theta_{j-1}^*\right)$ that appears in this decomposition. By definition,

$$\begin{aligned}&\pi\left(\theta_j^*|\mathscr{M},y,\theta_1^*,\ldots,\theta_{j-1}^*\right) =\\ &\int\pi\left(\theta_j^*|y,\theta_1^*,\ldots,\theta_{j-1}^*,\theta_{j+1},\ldots,\theta_p,\gamma\right)\ \mathrm{d}\pi\\ &\left(\theta_{j+1},\ldots,\theta_p,\gamma|y,\theta_1^*,\ldots,\theta_{p-1}^*\right)\end{aligned}$$

is the full conditional density integrated with respect to the distribution $\pi\left(\theta_{j+1},\ldots,\theta_p,\gamma|y,\theta_1^*,\ldots,\theta_{p-1}^*\right)$. To calculate this integral by Monte Carlo one can run an MCMC algorithm in which the blocks $\left(\theta_1,\ldots,\theta_{p-1}\right)$ are fixed at their starred values and sampling is over the remaining free blocks, namely $\left(\theta_j,\theta_{j+1},\ldots,\theta_p,\gamma\right)$. This is called a *reduced* MCMC run. Let the sampled draws from this reduced run be denoted by $\left(\theta_{j+1}^{(r)},\ldots,\theta_p^{(r)},\gamma^{(r)}\right), r=1,\ldots,M$. Then, provided the full conditional of θ_j is in closed form, we have the estimate

$$\begin{aligned}&\widehat{\pi}\left(\theta_j^*|\mathscr{M},y,\theta_{j-1}^*\right)\\ &= M^{-1}\sum_{r=1}^{M}\pi\left(\theta_j^*|y,\theta_1^*,\ldots,\theta_{j-1}^*,\theta_{j+1}^{(r)},\ldots,\theta_p^{(r)},\gamma^{(r)}\right).\end{aligned}$$

Each ordinate is estimated in this way from the output of the appropriate reduced runs. Notice that as more blocks are fixed, fewer distributions appear in the reduced runs.

Example 6 *Consider again the hierarchical Bayesian model for clustered data. In this case, we can decompose* $\pi(\theta^*,\gamma^*|\mathscr{M},y)$ *as*

$$\pi\left(D^{-1*}, \sigma^{2*}, \beta^{*}|y\right) = \pi\left(D^{-1*}|y\right)\pi\left(\sigma^{2*}|y, D^{*}\right)\pi\left(\beta^{*}|y, D^{*}, \sigma^{2*}\right),$$

so that all computations are marginalized over $\{b_i\}$. *The first term can be estimated by averaging the Wishart density given in Example 5 over draws on* $\{b_i\}$ *from the full MCMC run. To estimate the second ordinate, which is conditioned on* D^{*}, *we run a reduced MCMC simulation with the full conditional densities*

$$\pi\left(\beta|y, D^{*}, \sigma^{2}\right), \pi\left(\sigma^{2}|y, \beta, D^{*}, \{b_i\}\right), \pi\left(\{b_i\}|y, \beta, D^{*}, \sigma^{2}\right),$$

where each conditional utilizes the fixed value of D. The second ordinate is now estimated by averaging the inverse gamma full conditional density of σ^2 *at* $\sigma^{2}*$ *over the draws on* $(\beta\ \{b_i\})$ *from this reduced run. The third ordinate is multivariate normal as given in Example 5 and available directly.*

If the full conditional densities are not in closed form, the marginal likelihood can be computed by the modified Chib method as discussed in Chib and Jeliazkov (2001).

See Also

- Bayesian Econometrics
- Bayesian Statistics
- Econometrics
- Fixed Effects and Random Effects
- Longitudinal Data Analysis
- Markov Chain Monte Carlo Methods
- Model Selection
- Simulation-Based Estimation
- Statistical Inference

Bibliography

Albert, J.H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88: 669–679.

Berger, J. 1985. *Statistical decision theory and Bayesian analysis*. New York: Springer.

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

Chib, S., and B.P. Carlin. 1999. On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing* 9: 17–26.

Chib, S., and E. Greenberg. 1995. Understanding the metropolis-Hastings algorithm. *American Statistician* 49: 327–335.

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

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.

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.

Goel, P.K., and M.H. Degroot. 1981. Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association* 76: 140–147.

Lehmann, E., and G. Casella. 1998. *Theory of point estimation*. New York: Springer.

Lindley, D.V., and A.F.M. Smith. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B* 34: 1–41.

Tanner, M.A., and W.H. Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82: 528–550.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 21: 1701–1762.

Wakefield, J.C., A.F.M. Smith, A. Racine-Poon, and A.E. Gelfand. 1994. Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Applied Statistics* 43: 201–221.