

*The following excerpt was written by Mike Keane and John Geweke. It is Section 5 of a chapter titled "Computationally Intensive Methods for Integration in Econometrics" that will appear in the Handbook of Econometrics, volume 5; edited by James J. Heckman and Edward E. Leamer and published by North Holland.*

## 5. Univariate Latent Linear Models

Economic models are often used to study a single decision or outcome. The outcome variable may be fully observed, continuous, and unrestricted (for example, log consumption); fully observed and continuous but restricted to an interval (fraction of expenditure devoted to a certain category of goods); continuous but censored (earnings subject to known withholding limits for social insurance); a mixture of discrete and continuous outcomes (earnings of full-time high school students); categorical (income from survey data known only to be in a designated interval); or discrete (dichotomous choice, such as labor force participation). Depending on the model and data, other kinds of outcomes may be observed as well.

In all of these models, it is useful to conceive first of a latent outcome (denoted  $\tilde{y}_t$ , for observation  $t$ ), and then a corresponding set-valued observed outcome, denoted  $y_t$ . For example, in the case of a continuous outcome censored from above at  $c$ ,  $y_t = \tilde{y}_t$  if  $\tilde{y}_t \leq c$ , and  $y_t = (c, \infty)$  if  $\tilde{y}_t > c$ . In the case of a dichotomous outcome, one observes  $y_t = (-\infty, 0]$  if  $\tilde{y}_t \leq 0$  and  $y_t = (0, \infty)$  if  $\tilde{y}_t > 0$ . This construction is sometimes used explicitly in introducing the tobit model (Amemiya (1985) Section 10.2, or Greene (1997) Section 20.3.2) and probit model (Goldberger (1991) Section 29.1, or Maddala (1992) Section 8.9) respectively.

This section treats the linear model  $\tilde{y}_t = \mathbf{b}'\mathbf{x}_t + u_t$ , with observed outcomes of the form  $y_t = [c_t, d_t]$ ,  $y_t = [c_t, d_t)$ , or  $y_t = (c_t, d_t]$ , it being understood that  $c_t \leq d_t$  and that  $c_t$  and  $d_t$  are extended real numbers. The disturbances  $u_t$  ( $t = 1, \mathbf{K}, T$ ) are independent and identically distributed conditional on  $\mathbf{x}_t$  ( $t = 1, \mathbf{K}, T$ ). The disturbance  $u_t$  has a normal mixture distribution. We make this assumption because the normal mixture density can approximate any density arbitrarily well (Ferguson, 1983), and because it leads to practical methods for inference. It avoids the well-known problems that arise if the

distribution of  $u_t$  is assumed to be Gaussian when in fact this assumption is poor. In the specific case of dichotomous choice models the strategy here has objectives similar to those of nonparametric single-index models.<sup>1</sup> The treatment here differs in that it covers a much wider class of latent variable models, is fully Bayesian, and is computationally less demanding than methods for single-index models.

Section 5.1 presents an overview of the univariate latent linear model (ULLM), leaving technical detail to Appendix A. Section 5.2 provides some results with artificial data, to establish the practicality of the methods. The ULLM is incorporated in the Bayesian Analysis, Computation and Communication (BACC) software system. This system provides extensions to Gauss, Matlab, and S-plus by means of dynamically linked libraries, making it easy for one familiar with one of these commercial software packages to apply the model. Detailed information is available at <http://www.econ.umn.edu/~bacc>.

## 5.1 An overview of the ULLM

*Distribution of disturbances.* In the univariate latent linear model

$$(5.1.1) \quad \tilde{y}_t = \mathbf{b}'\mathbf{x}_t + u_t,$$

the disturbances  $u_t$  ( $t=1, \mathbf{K}, T$ ) are i.i.d. conditional on  $\mathbf{x}_t$  ( $t=1, \mathbf{K}, T$ ). Several alternative assumptions about the distribution of  $u_t$  can be made, and here we shall take up three in detail. The first is the conventional specification  $u_t \sim N(0, \mathbf{s}^2)$ , in which  $\mathbf{s}^2$  may be a free parameter (for example, in the censored linear model) or fixed as a condition of identification (for example,  $\mathbf{s}^2 = 1$  in the probit model).

The second alternative assumption about the distribution is  $u_t \sim t(0, \mathbf{s}^2; I)$ , a Student- $t$  distribution with location parameter 0, scale parameter  $\mathbf{s}$ , and degrees-of-freedom parameter  $I$ . The scale parameter may be fixed as a condition of identification. The disturbances may be represented  $u_t = \mathbf{s}_{(t)}\mathbf{h}_t$ , with  $(\mathbf{s}_{(t)}, \mathbf{K}, \mathbf{s}_{(t)})$  and  $(\mathbf{h}_1, \mathbf{K}, \mathbf{h}_T)$  i.i.d. and mutually independent conditional on  $(\mathbf{x}_1, \mathbf{K}, \mathbf{x}_T)$ . The latent variables  $\mathbf{s}_{(t)}^2$  have

---

<sup>1</sup> See for example Cosslett (1983), Manski (1985), Gallant and Nychka (1987), Powell, Stock and Stoker (1989), Horowitz (1992), Ichimura (1993), Klein and Spady (1993), and Lewbel (1997). For a detailed

independent inverted gamma distributions,  $\mathbf{l}/\mathbf{s}_{(t)}^2 \sim \mathbf{c}^2(\mathbf{l})$ , and  $\mathbf{h}_t \sim \mathbf{N}(0, \mathbf{s}^2)$ .<sup>2</sup> They subsequently play an important part for inference in this model.

The third alternative assumption about the distribution is  $u_t \sim \mathbf{N}(\mathbf{a}_j, \mathbf{s}^2 \mathbf{s}_j^2)$  with probability  $p_j$  ( $j=1, \mathbf{K}, m$ );  $\sum_{j=1}^m p_j = 1$ . This is a normal mixture model, with  $u_t$  drawn at random from one of  $m$  ‘‘urns’’, each urn containing a collection of  $u_t$  with a different normal distribution. By increasing the value of  $m$  and choosing the  $\mathbf{N}(\mathbf{a}_j, \mathbf{s}^2 \mathbf{s}_j^2)$  distributions appropriately, any univariate p.d.f. can be approximated arbitrarily well in the  $L_1$  topology (Ferguson, 1983). In this case, the disturbance may be represented  $u_t = \mathbf{a}' \tilde{\mathbf{z}}_t + \mathbf{s}_{(t)} \mathbf{h}_t$ . In this representation  $\mathbf{a}' = (\mathbf{a}_1, \mathbf{K}, \mathbf{a}_m)$ . The random variables  $(\mathbf{h}_t, \mathbf{K}, \mathbf{h}_t)$  are i.i.d. conditional on  $(\mathbf{x}_t, \mathbf{K}, \mathbf{x}_t)$ :  $\mathbf{h}_t \sim \mathbf{N}(0, \mathbf{s}^2)$ . The latent random vectors  $(\tilde{\mathbf{z}}_t, \mathbf{s}_{(t)})$  are i.i.d. conditional on  $(\mathbf{x}_t, \mathbf{K}, \mathbf{x}_t)$  and  $(\mathbf{h}_t, \mathbf{K}, \mathbf{h}_t)$ . Their values are governed by a latent state variable  $s(t)$  taking on the alternative values  $s(t) = j$  ( $j = 1, \mathbf{K}, m$ ). The  $s(t)$  are i.i.d. conditional on  $(\mathbf{x}_t, \mathbf{K}, \mathbf{x}_t)$  and  $(\mathbf{h}_t, \mathbf{K}, \mathbf{h}_t)$ , with  $\mathbf{P}[s(t) = j] = p_j$ . Conditional on  $s(t) = j$ , we have  $\mathbf{s}_{(t)} = \mathbf{s}_j$ ,  $\tilde{z}_{ij} = 1$ , and  $\tilde{z}_{ii} = 1 \forall i \neq j$ . To identify the model with respect to permutation of the state index, it is assumed that  $\mathbf{s}_1 > \mathbf{K} > \mathbf{s}_m$ . Identification of  $\mathbf{s}$  separately from  $\mathbf{s}_j$  ( $j=1, \mathbf{K}, m$ ) is taken up subsequently as part of the prior distribution.

All three specifications of the distribution of  $u_t$  in (5.1.1) are embedded in

$$\tilde{y}_t = \mathbf{a}' \underset{m \times 1}{\tilde{\mathbf{z}}_t} + \mathbf{b}' \underset{k \times 1}{\mathbf{x}_t} + \mathbf{e}_t,$$

$$\mathbf{e}_t = \mathbf{s}_{(t)} \mathbf{h}_t,$$

in which, conditional on  $(\mathbf{x}_t, \mathbf{K}, \mathbf{x}_t)$ ,  $\mathbf{h}_t \sim \mathbf{N}(0, \mathbf{s}^2)$  is i.i.d. and independent of  $(\tilde{\mathbf{z}}_t, \mathbf{s}_{(t)})$ .

The three specifications of the distribution of  $u_t$  in (5.1.1) are distinguished by the distribution of  $\mathbf{s}_{(t)}$ . Only when  $u_t$  has a normal mixture distribution is  $m > 0$ .

---

discussion of the use of mixture of normal models as an alternative to the probit model, see Geweke and Keane (1999).

*Observable outcomes.* Models are further distinguished by the observable outcome  $y_t$ , which in general is a set-valued function of the latent outcome  $\tilde{y}_t$ . If  $y_t = \tilde{y}_t$  ( $t=1, \mathbf{K}, T$ ) the ULLM reverts to the linear model. For the dichotomous choice model  $y_t = (-\infty, 0]$  if  $\tilde{y}_t \leq 0$  and  $y_t = (0, \infty)$  if  $\tilde{y}_t > 0$ . For an outcome censored from above at  $c$ ,  $y_t = \tilde{y}_t$  if  $\tilde{y}_t \leq c$ , and  $y_t = (c, \infty)$  if  $\tilde{y}_t > c$ . In all cases,

$$P(y_t, \tilde{y}_t | \mathbf{x}_t) = p(\tilde{y}_t | \mathbf{x}_t) p(y_t | \tilde{y}_t) = p(\tilde{y}_t | \mathbf{x}_t) \mathbf{c}_{y_t}(\tilde{y}_t),$$

in which  $\mathbf{c}_s(z)$  is the set indicator function:  $\mathbf{c}_s(z) = 1$  if  $z \in S$  and  $\mathbf{c}_s(z) = 0$  if  $z \notin S$ .

*Prior distributions.* Every model is endowed with a proper prior distribution. This makes it possible to compare different models for the same data using Bayes factors, as discussed below. For each model, we specify a benchmark prior distribution with hyperparameters that can be adjusted to reflect beliefs. These prior distributions are chosen for their combination of flexibility and analytical simplicity. Beyond the choice of hyperparameters, these prior distributions may be adjusted further to include prior distributions not in the benchmark families, by reweighting the output of the posterior simulator constructed subsequently.<sup>3</sup>

The benchmark prior distribution for  $\mathbf{b}$  is Gaussian,  $\mathbf{b} \sim N(\mathbf{b}, \mathbf{H}_b^{-1})$ , and that for  $\mathbf{s}^2$  is inverted gamma,  $\xi^2 / \mathbf{s}^2 \sim \mathbf{c}^2(\boldsymbol{\eta})$ . If  $\xi^2 / \boldsymbol{\eta} = \mathbf{s}^{*2}$  and  $\xi^2 \rightarrow \infty$ , then  $\mathbf{s}^2$  is degenerate at  $\mathbf{s}^{*2}$ . Thus  $\mathbf{s}^2 = 1$  can be enforced by a very large value of  $\xi^2 = \boldsymbol{\eta}$ .

For the Student- $t$  model the benchmark prior for the degrees of freedom parameter is exponential with mean  $\mathbf{l}$ ,  $\mathbf{l} \sim \exp(\mathbf{l})$ . Smaller values of  $\mathbf{l}$  reflect beliefs that the distribution is more leptokurtic.

The normal mixture model for the disturbances has three components:  $\mathbf{p}' = (p_1, \mathbf{K}, p_m)$ ,  $(\mathbf{s}_1^2, \mathbf{K}, \mathbf{s}_m^2)$ , and  $\mathbf{a}' = (\mathbf{a}_1, \mathbf{K}, \mathbf{a}_m)$ . The multinomial distribution of the state index  $s(t) = j$  ( $j = 1, \mathbf{K}, m$ ) involves the probabilities  $p_1, \mathbf{K}, p_m$ ,  $\sum_{j=1}^m p_j = 1$ . The benchmark prior distribution is Dirichlet (multivariate beta) with hyperparameters

---

<sup>2</sup> For a derivation of this construction see Johnson, Kotz and Balakrishnan (1995, Section 28.1) or Geweke (1993).

<sup>3</sup> Such reweighting is discussed in Geweke (1999) Section 6 and is easy to carry out in the BACC software system.

$r_1, \mathbf{K}, r_m$ :  $p(\mathbf{p}) \propto \prod_{j=1}^m p_j^{r_j-1}$ . In this distribution  $p_j$  has mean  $r_j/R$  and standard deviation  $[r_j(R-r_j)]^{1/2}/R(R+1)^{1/2}$ , where  $R = \sum_{i=1}^m r_i$ .

The benchmark prior distribution for the second component of the normal mixture model, the variance scaling parameters  $\mathbf{s}_j^2$ , consists of the  $m$  inverted gamma components  $\mathfrak{s}_j^2/\mathbf{s}_j^2 \sim \mathbf{c}^2(\mathbf{n}_j)$ . These are subject to the restrictions  $\mathbf{s}_1^2 > \mathbf{K} > \mathbf{s}_m^2$  but otherwise independent. The ordering of the  $\mathbf{s}_j^2$  removes the possibility of permuting the states, but imposes the restriction  $\mathbf{s}_i^2 \neq \mathbf{s}_j^2 \forall (i, j)$ . We choose to identify states by variance, because in our applications this is more convenient than identifying states by orderings of state probabilities,  $p_j$ , or state means,  $\mathbf{a}_j$ . The lack of identification in the likelihood imposed by the fact that the variance in component  $j$  is  $\mathbf{s}^2 \mathbf{s}_j^2$  is resolved by the proper prior distributions for  $\mathbf{s}^2$  and  $(\mathbf{s}_1^2, \mathbf{K}, \mathbf{s}_m^2)$ . Identification can also be achieved in the traditional way by taking  $\mathfrak{s}_i^2 \rightarrow \infty$  and  $\mathbf{n}_i \rightarrow \infty$  while  $\mathfrak{s}_i^2/\mathbf{n}_i = 1$ , for a selected state  $i$ , thereby making  $\mathbf{s}^2$  the variance in that state. In either event, in the prior distribution of variance ratios across states,

$$\frac{\mathbf{s}^2 \mathbf{s}_k^2}{\mathbf{s}^2 \mathbf{s}_j^2} \sim \frac{\mathfrak{s}_k^2/\mathbf{n}_k}{\mathfrak{s}_j^2/\mathbf{n}_j} \cdot \mathbf{F}(\mathbf{n}_j, \mathbf{n}_k)$$

subject to  $\mathbf{s}^2 \mathbf{s}_j^2 / \mathbf{s}^2 \mathbf{s}_k^2 < 1$  if  $j > k$ . Thus the prior distribution for the  $\mathbf{s}_j^2$  incorporates only beliefs about relative variances. This is convenient in thinking about shapes (as opposed to scales) of distributions – for example, outliers or other forms of leptokurtosis.

The third component of the normal mixture distribution,  $\mathbf{a}' = (\mathbf{a}_1, \mathbf{K}, \mathbf{a}_m)$ , is multivariate normal,  $\mathbf{a}|\mathbf{s} \sim \mathbf{N}(\mathbf{0}, \mathbf{s}^2 \mathbf{H}_a^{-1})$ . This prior distribution is taken conditional on  $\mathbf{s}$ , and prior variance is proportional to  $\mathbf{s}^2$ , in order to represent beliefs about the shape of the disturbance p.d.f. independent of its scale. In all of the illustrations in the next subsection,  $\mathbf{H}_a = \underline{h}_a \mathbf{I}_m$ . This restriction requires these prior beliefs about means to be interchangeable across the mixture components. For example, given the number of states,  $m$ , the greater the precision  $\underline{h}_a$ , the more likely is the p.d.f. to be unimodal.

*Existence of the posterior.* For any particular ULLM the product of the relevant prior densities and data density is the kernel of the posterior distribution, so long as that product is finitely integrable over the space of all parameters and latent variables. If the data density is bounded above, then the integrability condition is met when the prior distribution is proper (as it is here). For all variants of the ULLM with normal and Student- $t$  densities the data density is bounded, as it is for all variants in which all outcomes are discrete (i.e.,  $c_t < d_t \forall t$ ). If the disturbances are mixed normal and at least some of the  $\tilde{y}_t$  are not latent (i.e.,  $c_t = d_t$  for at least some  $t$ ) then the data density is unbounded. In this case the integrability of the posterior kernel can be demonstrated, but the argument is more technical and is relegated to Appendix A.

*MCMC algorithm for inference.* The explicit development of the data density and prior density, whose product is the posterior density kernel, is given in Appendix A. There is no corresponding closed form for the posterior distribution of all of the parameters jointly. However, the Gibbs sampling algorithm described in Section 2.2 can be applied to eight groups of parameters or latent variables that appear in the posterior kernel:  $(\mathbf{a}, \mathbf{b})$ ;  $\mathbf{s}^2$ ;  $\mathbf{s}_{(t)}^2$  ( $t = 1, \mathbf{K}, T$ );  $\mathbf{l}$ ;  $s(t)$  ( $t = 1, \mathbf{K}, T$ ) and  $\tilde{\mathbf{Z}}$ ;  $\mathbf{p}$ ;  $\mathbf{s}_j^2$  ( $j = 1, \mathbf{K}, m$ ); and  $\tilde{y}_t$  ( $t = 1, \mathbf{K}, T$ ). (Not all parameters appear under each assumption about the distribution of  $u_t$ .)

The Gibbs sampling algorithm is practical because the distribution of each parameter group, conditional on all the others, is simple enough that draws from the conditional distribution can be made. In particular, the conditional distribution of  $(\mathbf{a}, \mathbf{b})$  is multivariate normal; those of  $\mathbf{s}^2$  and  $\mathbf{s}_{(t)}^2$  ( $t = 1, \mathbf{K}, T$ ) are inverted gamma; and each  $\tilde{y}_t$  is truncated normal. In the Student- $t$  model, the conditional distribution of  $\mathbf{l}$  is not standard, but a Metropolis within Gibbs step (Section 2.5) employing a Gaussian approximation to the conditional distribution as the proposal distribution works well. (Details are presented in Appendix A.) In the normal mixture model, the conditional distribution of  $\mathbf{p}$  is Dirichlet and the state assignments are multinomial. The conditional distribution of  $\mathbf{s}_j^2$  is inverted gamma, subject to truncation restrictions imposed by the ordering  $\mathbf{s}_1 > \mathbf{K} > \mathbf{s}_m$ .

For seven of the eight groups of parameters in this algorithm, the support of the conditional posterior distribution is the same as the support of the marginal posterior distribution. The exception is the group of variance parameters  $\mathbf{s}_j^2$  ( $j=1, \mathbf{K}, m$ ) in the normal mixture model. Thus for the normal and Student- $t$  variants of the ULLM, Corollary 2.4.2 assures convergence of the Gibbs Markov chain to the posterior distribution.

For the normal mixture model, consider any point in the support of the posterior density, and any subset of the posterior density support with positive posterior probability. There exists a finite number of iterations of the algorithm such that the transition probability from the point to the subset is positive. (The minimum number of steps will depend on the values of the  $\mathbf{s}_j^2$  ( $j=1, \mathbf{K}, m$ ) at the point in question, and their values in the subset. In the normal and Student- $t$  models, this minimum number of steps is one for any point and any subset combination.) This condition assures that the transition density of the chain is aperiodic and absolutely continuous with respect to the posterior density (Tierney, 1994). From Corollary 1 of Tierney (1994), the sequence of parameters and latent variables  $\{\mathbf{q}^{(m)}\}$  produced by the Markov chain is ergodic: that is, if a posterior moment  $\bar{g} = E[\mathbf{g}(\mathbf{q})|\mathbf{X}, \mathbf{y}]$  exists, then  $\bar{g}_M = M^{-1} \sum_{m=1}^M \mathbf{g}(\mathbf{q}^{(m)}) \xrightarrow{a.s.} \bar{g}$ .

*Marginal likelihoods.* It is useful to be able to compare two alternative specifications of the ULLM – for example, models with different specifications of the disturbance distribution, with different covariates  $\mathbf{X}$ , or with different prior distributions. A formal comparison can be made by means of a posterior odds ratio. Let  $A_1$  and  $A_2$  denote the alternative specifications of the ULLM and  $P(A_1)$  and  $P(A_2)$  the prior probabilities of the alternative model specifications themselves. Then the posterior odds ratio in favor of the specification  $A_1$  is

$$\frac{p(A_1|\mathbf{X}, \mathbf{y})}{p(A_2|\mathbf{X}, \mathbf{y})} = \frac{P(A_1)p(\mathbf{y}|A_1)}{P(A_2)p(\mathbf{y}|A_2)},$$

in which the marginal likelihoods  $p(\mathbf{y}|A_j)$  are given by (2.0.2). The key technical task is to evaluate the integrals in (2.0.2).

For the ULLM, a convenient way to approximate the marginal likelihood is to use the modified harmonic mean method of Gelfand and Dey (1994), as further developed in Geweke (1999) Section 4.3. This method requires that the prior densities  $p(\mathbf{q}_j^{(m)}|A_j)$  and data densities  $p(\mathbf{y}|\mathbf{X},\mathbf{q}_j^{(m)},A_j)$  be evaluated for each iteration  $m$  of the MCMC algorithm. Once this is done, the Monte Carlo approximation of the marginal likelihood may be carried out using generic software described in Geweke (1999) Section 4.5.<sup>4</sup> The evaluation of the data density  $p(\mathbf{y}|\mathbf{X},\mathbf{q}_j^{(m)},A_j)$  is relatively straightforward in the ULLM, but some care is required in the handling of the latent variables. Details are given in Appendix A.

## 5.2 Some evidence from artificial data

Before proceeding to apply the ULLM, a number of practical issues arise. For some variants of the ULLM, it is simply of interest to see how well the posterior distribution recovers the underlying population parameters. This is especially true of models with latent  $\tilde{y}_i$  – for example, what can be learned about the degrees of freedom parameter  $\mathbf{I}$  in the Student- $t$  dichotomous choice model, or the conditional means  $\mathbf{a}_j$  in the censored linear model or dichotomous choice model with mixed normal disturbances?

In all cases, it is important to ascertain some information about computational efficiency. This is not simply a matter of the computation time required for each iteration of the MCMC algorithm. It is also driven by the degree of serial correlation of the parameters drawn from one iteration to the next. The variance of the numerical approximation of the posterior mean is computed using conventional time series methods. The method is, essentially, to apply a set of linearly declining weights to the first 8% of the autocovariances of the sequence  $\{g(\mathbf{q}^{(m)})\}$ , for a given function of interest  $g(\cdot)$ .<sup>5</sup>

---

<sup>4</sup> Other methods for Monte Carlo approximation include Chib (1995) and importance sampling as discussed in Geweke (1996). Neither applies directly to a Gibbs sampling algorithm with Metropolis steps.

<sup>5</sup> For long simulations, care must be taken to achieve computational efficiency in the computation of autocovariances. Geweke (1999, Section 3.8) provides details.

*Parameter posterior moments.* Tables 5.1-5.3 provide model specifications, prior distributions, and some posterior moments for instances of the univariate linear model, censored linear model, and dichotomous choice linear model, respectively. In each case sample size is  $T = 2,000$ ,  $x_{t1} = 1.0$  is an intercept, and  $x_{2t}$  and  $x_{3t}$  are independent, i.i.d. standard normal variates. The nine data sets, including covariates, were drawn independently. The MCMC algorithm was executed for 12,000 iterations in each case, and the last  $M = 10,000$  iterations were used for the computations. Each table shows the parameter values used to generate the data.

Panel C of Table 5.1 provides posterior moments for the textbook normal linear model. There are no surprises: the posterior standard deviations of the  $\mathbf{b}_j$  are all about  $(200)^{-1/2}$  and that of  $\mathbf{s}$  is about  $(1000)^{-1/2}$ , the values suggested by the design of the experiment. Given the approximate orthogonality of  $\mathbf{b}$  and  $\mathbf{s}^2$  in the posterior distribution the MCMC draws should be nearly i.i.d., and this is reflected in RNEs close to 1.0.<sup>6</sup>

The Student- $t$  linear model (Table 5.1, panel D) entails draws of  $\mathbf{s}_{(t)}^2$  ( $t = 1, \mathbf{K}, T$ ) each iteration. This accounts for the doubling of computation time per iteration, compared with the normal linear model. There is a modest increase in the posterior standard deviation of  $\mathbf{b}$  and  $\mathbf{s}$ .<sup>7</sup> The posterior mean of the degrees of freedom parameter  $\mathbf{I}$  is close to the population value and well within one posterior standard deviation. Additional serial correlation (relative to the normal linear model) is introduced to the MCMC algorithm by the addition of  $\mathbf{s}_{(t)}^2$  ( $t = 1, \mathbf{K}, T$ ) and  $\mathbf{I}$  to the parameter list. This has at most a modest impact on the RNE of the approximation of  $E(\mathbf{b}_j | \mathbf{X}, \mathbf{y})$ . The main impact is on the numerical approximation error for  $\mathbf{s}$ . This arises because of the positive posterior correlation of  $\mathbf{s}$  and  $\mathbf{I}$ , and the fact that they are drawn in separate blocks of the MCMC algorithm. The numerical standard error of  $\mathbf{I}$  is reduced to 10% of its

---

<sup>6</sup> The same algorithm is applied to this simple normal linear model as is applied in the ULLM generally. Thus, moment matrices like  $\mathbf{X}'\mathbf{X}$  are recomputed each iteration. Code designed specifically for the normal linear model, such as that in the BACC software (<http://www.econ.umn.edu/~bacc>) is substantially more efficient.

posterior standard deviation in about 3,000 iterations, requiring about two minutes of computing time.

The normal mixture model (Table 5.1, panel E) contains two normal components of the disturbance, each with the same probability but different means and variances. (The same mixture is used in the mixed normal censored linear and dichotomous choice models.) The variance of the disturbance is .52, smaller than in the normal models. If the normal linear model is applied to this data set, posterior standard deviations of  $\mathbf{b}_2$  and  $\mathbf{b}_3$  are about  $.016 = (.52/2000)^{1/2}$  as one would expect. . If the states were known, the posterior standard deviations of  $\mathbf{b}_2$  and  $\mathbf{b}_3$  would be about  $.0062 = (1000/.2^2 + 1000/1^2)^{-1/2}$ . The actual posterior standard deviations are much closer to the latter value than the former. That they exceed .0062 can be attributed to the imperfect sorting of observations by state. The posterior means of the intercept values are well within two posterior standard deviations of population values. Since  $\mathbf{s}_1 = 5\mathbf{s}_2$ ,  $\text{var}(\mathbf{b}_1 + \mathbf{a}_1|\mathbf{X}, \mathbf{y}) > \text{var}(\mathbf{b}_1 + \mathbf{a}_2|\mathbf{X}, \mathbf{y})$ , but  $\text{var}(\mathbf{b}_1 + \mathbf{a}_1|\mathbf{X}, \mathbf{y})/\text{var}(\mathbf{b}_1 + \mathbf{a}_2|\mathbf{X}, \mathbf{y}) < 5$  again due to imperfect sorting by states.

The low value of relative numerical efficiency indicates strong serial correlation in  $\mathbf{s}_2$  in the MCMC algorithm. This arises because of high correlation between  $\mathbf{s}_2$  and the state classifications  $s(t)$ . Because of the contrast in standard deviations ( $\mathbf{s}_1 = 5\mathbf{s}_2$ ), there are only a few observations for which  $p[\mathbf{e}_i|s(t) = 1] \approx p[\mathbf{e}_i|s(t) = 2]$ , and therefore there is substantial persistence in state classification. Simple arithmetic shows that the reclassification of an observation has a much larger effect on the smaller variance. Since the effects of changes in other parameters on  $\mathbf{s}_1$  and  $\mathbf{s}_2$  is about the same,  $\mathbf{s}_2$  is more strongly driven by the slowing moving state assignments.

Table 5.2 presents similar information for the censored linear model. In this model  $\tilde{y}_i$  is observed if and only if  $\tilde{y}_i > 0$ , so about half of the  $T = 2,000$  observations are censored. Compared with the linear model, posterior standard deviations are in every case higher, reflecting the loss of information in censoring. Since about half the  $\tilde{y}_i$  must

---

<sup>7</sup> Note that the population variance of the disturbance in this model is  $\mathbf{s}^2 \mathbf{I}/(\mathbf{I} - 2) = 5/3$ . When the normal linear model is applied to this data set, the posterior standard deviation of the  $\mathbf{b}_j$  is quite close to the value

be drawn each iteration, one would expect an increase in serial correlation. This is reflected in a reduction of RNE for all parameters except  $\mathbf{s} \cdot \mathbf{s}_2$  in the normal mixture model. Comparisons among panels C, D and E in Table 5.2 are similar to the comparisons already discussed for their counterparts in Table 5.1. The increase in RNE for  $\mathbf{s} \cdot \mathbf{s}_2$  is due to the fact that uncertainty about  $\tilde{y}_i$  for those  $\tilde{y}_i < 0$  now contributes in a major way to all parameters, and serial correlation in the draws of  $\tilde{y}_i$  from one simulation to the next contributes to the serial correlation in all parameters in the MCMC algorithm. Thus, the contrast in the impact of state classification on  $\mathbf{s}_1$  and  $\mathbf{s}_2$  is less important, relative to all other factors contributing to serial correlation, than was the case in the linear model with mixed normal disturbances.

Table 5.3 presents the same information for the dichotomous choice linear model. Given the parameter values and the distribution of  $x_i$ , the probabilities of the choices are .5 for the normal and Student-t disturbances, and the probability of choice one ( $\tilde{y}_i < 0$ ) is .47 for the mixed normal disturbances. With the obvious exception of the parameters  $\mathbf{s}$  and  $\mathbf{s}_1$ , which are normalized at 1.0, posterior standard deviations for all parameters are higher than was the case in the censored linear model. The largest increases are in the posterior standard deviations of the parameters of the disturbance distribution (other than  $\mathbf{s}$  and  $\mathbf{s}_1$ ). The increase in the posterior standard deviation of the degrees of freedom parameter  $I$  is especially large. Given the increased latency of  $\tilde{y}_i$  in this model, these developments are unsurprising. There is evidence of increased serial correlation (relative to the censored linear model) in the MCMC algorithm, in the form of reduced RNEs, when panels C, D, and E in Table 5.3 are compared with their counterparts in Table 5.2. This decreased efficiency is most pronounced in the case of mixed normal disturbances. (Note that now the RNE of  $\mathbf{s}_2$  is comparable with that of other parameters.)

Inference in the dichotomous choice linear model with mixed normal disturbances is reliable, in the sense that for all the parameters the posterior standard deviation is substantially less than the prior standard deviation, and all posterior means are within about one posterior standard deviation of the population values. There is very substantial serial persistence in the MCMC algorithm, but each iteration requires only about 0.1

---

of  $[(5/3)/2000]^{1/2}$  that one would expect.

seconds. Based on an RNE of .008, numerical standard errors are driven to one-fourth of posterior standard deviation after about 2000 iterations (about three minutes), to one-tenth after 12,500 iterations (about 20 minutes), and to 1% after  $1.25 \times 10^6$  iterations (about 1.5 days). The practicality of the procedure thus depends on one's standards for accuracy. As we shall now see, a great deal can be learned with just a few thousand iterations.

*Marginal likelihood approximations.* In the case of the ULLM, the additional computations required to produce the marginal likelihood are trivial. The likelihood function and prior distribution must be evaluated for those iterations that are used to approximate the marginal likelihood. Since these evaluations are not used subsequently in the MCMC algorithm they need not be made every iteration, but doing so in each iteration increases the computation time only by about 2%. Given the sequence of likelihood and prior evaluations from the MCMC algorithm, the log marginal likelihood is approximated using the variant of the Gelfand and Dey (1994) procedure described above. Computing time for this approximation, using the implementation detailed in Geweke (1999) Section 4.3, is essentially proportional to the number of iterations – about 2 seconds for 10,000 iterations.

Table 5.4 shows several patterns of results. First, for each data set (column headings) the model (row headings) that generates the observations receives the highest marginal likelihood, and therefore the highest posterior probability, of the three models compared. The lowest odds ratio in favor of any true model is that in favor of the Student- $t$  over the mixed normal censored linear model, about 4.5:1. Second, the highest odds ratios occur when the competitor to a true model does not nest or approximate the true model – that is, Student- $t$  versus normal when disturbances are Student- $t$ , and normal mixture versus either normal or Student- $t$  when the disturbances are normal mixture. Third, odds ratios are usually higher when the outcome ( $\tilde{y}_i$ ) is fully observed than when it is not: they tend to be highest in the linear model, lowest in the dichotomous choice model. Fourth, odds ratios in favor of true models against nesting models (e.g., in favor of normal versus Student- $t$  when disturbances are normal) or in favor of true models against approximating models (e.g., Student- $t$  versus normal mixture when disturbances

are Student-t) are lower, and the degree of latency has little effect on the magnitude of the odds ratio.

We conclude that in these examples, discrimination between the three disturbance distributions is effective. Moreover while the results shown in Table 5.3 are based on 10,000 MCMC iterations after discarding the first 2,000, nearly identical results are obtained with 900 iterations after discarding the first 100. For these examples, an investigator could learn quickly which models account for most of the posterior probability and then concentrate computing resources on those models.

### **5.3 Some evidence from real data**

Especially in large data sets, it is relatively easy to detect departures from normality and establish the form of the non-Gaussian distribution with a high degree of precision. Space constraints do not permit development of these applications in detail, so we confine this discussion to three examples in our recent work.

Geweke and Keane (2000) estimates a reduced form life cycle earnings model of the kind introduced by Lillard and Willis (1978) and used in a succession of studies since. A recurring puzzle in this literature has been the inability of these models to capture the transition of individual earnings in and out of the lowest quintile of the earnings distribution. Geweke and Keane (2000) adopts the standard model, but departs from it in two specific ways. First and most important, it specifies shocks to current earnings to be a mixture of three normals. Second, it sets up the regression of earnings on age and education as a high-order polynomial. (There are other elaborations as well, but these are the most important in the context of the ULLM.) Among the paper's many findings, three are important with respect to the ULLM. First, the evidence against normality is overwhelming: the distribution of the shock to current earnings is strongly skewed and leptokurtic. When the same model is fit using a normal distribution, the .40 quantile of the fitted normal corresponds to the .20 quantile of the mixture of three normal distributions. Second, the normal mixture model implies dynamics for the movement of individual earnings in and out of the lowest quintile that are quite similar to those in the data, and much closer than has been captured previously by reduced form life cycle earnings models in the literature. Third, maximum likelihood estimation in this model is

not possible, because the likelihood function has a multitude of isolated singularities. (This point is discussed briefly in Appendix A of this chapter, and in greater detail in on-line Appendix F of Geweke and Keane (2000).)

Two simpler applications appear in Geweke, McCausland and Stevens (2000) and Geweke and Keane (1999). The former example is a simple hedonic regression model for residential real estate prices. The sample size is modest ( $n = 546$ ) and the departure from normality is small (least squares residual kurtosis 4.02). The Bayes factor in favor of a mixture of two normals is about 20. The latter example is a dichotomous choice model of women's labor force participation ( $n = 1555$ ) with conventional covariates. Of a dozen models including the conventional probit model and mixtures of up to five normals, the model with the highest marginal likelihood is a mixture of four normals. All the mixture models are highly favored relative to the conventional probit model, the Bayes factors ranging from  $2 \times 10^5$  to  $9 \times 10^7$ .

**Table 5.1:** Univariate linear model ( $T = 2,000$ )

A. Model specification

$$y_t = \mathbf{b}_1 + \sum_{j=2}^3 x_{jt} + u_t$$

Population for all variants:  $x_{2t} \stackrel{iid}{\sim} N(0, 1)$ ,  $x_{3t} \stackrel{iid}{\sim} N(0, 1)$ ,  $\mathbf{b}_1 = 0$ ,  $\mathbf{b}_2 = 1$ ,  $\mathbf{b}_3 = -1$

Normal disturbances:  $u_t \stackrel{iid}{\sim} N(0, 1)$ ; Student-t disturbances:  $u_t \stackrel{iid}{\sim} t(0, 1; 5)$

Mixed normal disturbances:  $u_t \stackrel{iid}{\sim} N(-.3, 1)$ ,  $p_1 = .5$ ;  $u_t \stackrel{iid}{\sim} N(.3, .2^2)$ ,  $p_1 = .5$

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
$\mathbf{b}_j$ ( $j=1,2,3$ )	$\mathbf{b}_j \sim N(0, 1)$	0.0	1.0
$\mathbf{s}$	$4/\mathbf{s}^2 \sim \mathbf{c}^2(4)$	1.253	.655
$\mathbf{l}$	$\mathbf{l} \sim \exp(5)$	5.0	3.162
$\mathbf{a}_j$ ( $j=1,2$ )	$\mathbf{a}_j \sim N(0, 5\mathbf{s}^2)$	0.0	$2.236\mathbf{s}$
$\mathbf{s}_j$	$4/\mathbf{s}_1^2 \sim \mathbf{c}^2(4)$	1.253	1.414
	$0.4/\mathbf{s}_2^2 \sim \mathbf{c}^2(4)$	.396	.447
$p_1$	Beta(2, 2)	.500	.224

C. Some posterior moments: Normal disturbances; .018 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_1 = 0$	-.031	.023	.808	$\mathbf{b}_3 = -1$	-.991	.022	1.380
$\mathbf{b}_2 = 1$	1.014	.022	1.112	$\mathbf{s} = 1$	1.000	.016	1.380

D. Some posterior moments: Student- $t$  disturbances; .037 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_1 = 0$	-.026	.025	.505	$\mathbf{b}_3 = -1$	-1.016	.024	1.052
$\mathbf{b}_2 = 1$	.968	.025	.538	$\mathbf{s} = 1$	.971	.029	.063
				$\mathbf{l} = 5$	5.111	.702	.032

E. Some posterior moments: Normal mixture disturbances; .052 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_2 = 1$	1.004	.007	1.054	$\mathbf{s} \cdot \mathbf{s}_1 = 1$	.965	.022	.745
$\mathbf{b}_3 = -1$	-1.019	.007	.412	$\mathbf{s} \cdot \mathbf{s}_2 = .2$	.193	.008	.007
$\mathbf{b}_1 + \mathbf{a}_1 = -.3$	-.303	.008	.520	$p_1 = .5$	.511	.018	.107
$\mathbf{b}_1 + \mathbf{a}_2 = .3$	.260	.033	.477				

**Table 5.2:** Univariate censored linear model ( $T = 2,000$ )

A. Model specification

$$\tilde{y}_t = \mathbf{b}_1 + \sum_{j=2}^3 x_{jt} + u_t; \quad y_t = \mathbf{c}_{(0,\infty)}(\tilde{y}_t)\tilde{y}_t + \mathbf{c}_{(-\infty,0)}(\tilde{y}_t)(-\infty,0)$$

Data generating process otherwise as in Table 5.1

B. Prior distributions and moments: As in Table 5.1

C. Some posterior moments: Normal disturbances; .044 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_1 = 0$	-.048	.034	.101	$\mathbf{b}_3 = -1$	-1.013	.031	.154
$\mathbf{b}_2 = 1$	1.020	.032	.234	$\mathbf{s} = 1$	1.009	.024	.188

D. Some posterior moments: Student- $t$  disturbances; .070 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_1 = 0$	-.063	.037	.181	$\mathbf{b}_3 = -1$	-1.036	.034	.167
$\mathbf{b}_2 = 1$	1.030	.036	.262	$\mathbf{s} = 1$	.943	.040	.017
				$\mathbf{l} = 5$	4.478	.708	.011

E. Some posterior moments: Normal mixture disturbances; .095 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_2 = 1$	1.002	.015	.060	$\mathbf{s} \cdot \mathbf{s}_1 = 1$	.978	.031	.672
$\mathbf{b}_3 = -1$	-1.019	.008	.839	$\mathbf{s} \cdot \mathbf{s}_2 = .2$	.197	.021	.107
$\mathbf{b}_1 + \mathbf{a}_1 = -.3$	-.293	.020	.038	$p_1 = .5$	.477	.025	.070
$\mathbf{b}_1 + \mathbf{a}_2 = .3$	.216	.049	.139				

**Table 5.3:** Dichotomous choice linear model ( $T = 2,000$ )

A. Model specification:

$$\tilde{y}_t = \mathbf{b}_1 + \sum_{j=2}^3 x_{jt} + u_t; \quad y_t = \mathbf{c}_{(0,\infty)}(\tilde{y}_t)(0,\infty) + \mathbf{c}_{(-\infty,0)}(\tilde{y}_t)(-\infty,0)$$

Data generating process otherwise as in Table 5.1

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
$\mathbf{s}$	$\mathbf{s} = 1$	1.0	0.0
$\mathbf{s}_j$	$\mathbf{s}_1 = 1$	1.0	0.0
	$0.4/\mathbf{s}_2^2 \sim \mathbf{c}^2(4)$	.396	.447

$\mathbf{b}_j$  ( $j=1,2,3$ ),  $\mathbf{I}$ ,  $\mathbf{a}_j$  ( $j=1,2$ ),  $p_1$ : As in Table 5.1

C. Some posterior moments: Normal disturbances; .067 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_1 = 0$	-.035	.036	.285	$\mathbf{b}_3 = -1$	-1.086	.050	.096
$\mathbf{b}_2 = 1$	1.040	.049	.555				

D. Some posterior moments: Student- $t$  disturbances; .096 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_1 = 0$	-.011	.048	.018	$\mathbf{b}_3 = -1$	-1.283	.147	.004
$\mathbf{b}_2 = 1$	1.265	.141	.004	$\mathbf{I} = 5$	3.639	1.808	.003

E. Some posterior moments: Normal mixture disturbances; .139 sec./iter.

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
$\mathbf{b}_2 = 1$	1.032	.080	.003	$\mathbf{s}_2 = 0.2$	.302	.055	.007
$\mathbf{b}_3 = -1$	-1.051	.080	.004	$p_1 = .5$	.462	.080	.004
$\mathbf{b}_1 + \mathbf{a}_1 = -.3$	-.361	.066	.004				
$\mathbf{b}_1 + \mathbf{a}_2 = .3$	.395	.129	.011				

**Table 5.4**

Log marginal likelihoods  
in some univariate latent linear models with artificial data

A. Linear model

Data disturbances:	Normal	Student-t	Mixed normal
Model:			
Normal	-2851.4	-3268.2	-2297.7
Student-t	-2855.5	-3207.2	-2155.3
Mixed normal	-2857.8	-3213.5	-1900.4

B. Censored linear model

Data disturbances:	Normal	Student-t	Mixed normal
Model:			
Normal	-1821.5	-2037.4	-1575.1
Student-t	-1826.3	-2001.8	-1528.6
Mixed normal	-1826.5	-2003.3	-1403.3

C. Dichotomous choice model

Data disturbances:	Normal	Student-t	Mixed normal
Model:			
Normal	-802.1	-851.8	-677.4
Student-t	-804.5	-847.0	-673.7
Mixed normal	-807.8	-849.3	-662.4

## References

- Amemiya, T. (1985), *Advanced Econometrics* (Cambridge, Harvard University Press).
- Chib, S. (1995), "Marginal Likelihood from the Gibbs Output." *Journal of the American Statistical Association* 90: 1313–1321.
- Cosslett, S.R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica* 51: 765–782.
- Ferguson, T.S. (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," in H. Rivizi and J. Rustagi, eds., *Recent Advances in Statistics* (New York, Academic Press) 287-302.
- Gallant, A.R., and D.W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica* 55: 363–390.
- Gelfand, A.E., and D.K. Dey (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society Series B* 56: 501-514.
- Geweke, J. (1993), "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics* 8: S19-S40.
- Geweke, J. (1996), "Monte Carlo Simulation and Numerical Integration," in H.M. Amman, D.A. Kendrick and J. Rust, eds., *Handbook of Computational Economics*. (Amsterdam, North–Holland) 731-800.
- Geweke, J. (1999), "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication" (with discussion and reply). *Econometric Reviews* 18: 1-127.
- Geweke, J. and M. Keane (1999), "Mixture of Normals Probit Models," in C. Hsiao, K. Lahiri, L-F. Lee and H. Pesaran, eds., *Analysis of Panels and Limited Dependent Variable Models: An Edited Volume in Honor of G.S. Maddala*, Cambridge University Press, 49-78.
- Geweke, J. and M. Keane (2000), "An Empirical Analysis of Earnings Dynamics Among Men in the PSID: 1968-1989," *Journal of Econometrics* 92: 293-356.
- Geweke, J., W. McCausland and J. Stevens (2000), "Using Simulation Methods for Bayesian Econometric Models," in D. Giles, ed., *Computer Aided Econometrics* (New York, Marcel Dekker) forthcoming.
- Goldberger, A.S. (1991), *A Course in Econometrics* (Cambridge, Harvard University Press).

- Greene, W.H. (1997), *Econometric Analysis* (Upper Saddle River, Prentice Hall). (Third Edition)
- Horowitz, J.L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica* 60: 505–531.
- Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics* 58: 71–120.
- Johnson, N.L., S. Kotz, and N. Balakrishnan (1995), *Continuous Univariate Distributions* (Volume 2) (New York, Wiley). (Second Edition)
- Klein, R.W. and R.H. Spady (1993), "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61: 387–421.
- Lewbel, A. (1997), "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory* 13: 32–51.
- Lillard, L. and R. Willis (1978), "Dynamic Aspects of Earnings Mobility," *Econometrica* 46: 985-1012.
- Maddala, G.S. (1992), *Introduction to Econometrics* (New York, Macmillan). (Second Edition)
- Manski, C.F. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics* 27: 313–333.
- Powell, J.L., J.H. Stock, and T.M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57: 1403–1430.
- Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion and rejoinder), *Annals of Statistics* 22: 1701–1762.