10

Flexible Models: Nonparametric and Semiparametric Methods

10.1 INTRODUCTION

All the models considered in previous chapters involved making assumptions about functional forms and distributions. For instance, the Normal linear regression model involved the assumptions that the errors were Normally distributed, and the relationship between the dependent and explanatory variables was a linear one. Such assumptions are necessary to provide the likelihood function, which is a crucial component of Bayesian analysis. However, economic theory rarely tells us precisely what functional forms and distributional assumptions we should use. For instance, in a production example economic theory often tells us that a firm's output in increasing in its inputs and eventually diminishing returns to each input will exist. Economic theory will not say "a constant elasticity of substitution production function should be used". In practice, a careful use of the model comparison and fit techniques described in previous chapters (e.g. posterior predictive p-values and posterior odds ratios) can often be used to check whether the assumptions of a particular likelihood function are reasonable. However, in light of worries that likelihood assumptions may be inappropriate and have an effect on empirical results, there is a large and growing non-Bayesian literature on *nonparametric* and *semiparametric methods*.[1] To motivate this terminology, note that likelihood functions depend on parameters and, hence, making particular distributional or functional form assumptions yields a *parametric likelihood function*. The idea underlying the nonparametric literature is to try and get rid of the such parametric assumptions either completely (in the case of nonparametric methods) or partially (in the case of semiparametric methods).[2]

---

[1] Horowitz (1998) and Pagan and Ullah (1999) provide good introductions to this literature.

[2] This noble goal of "letting the data speak" is often hard to achieve in practice since it is necessary to place some structure on a problem in order to get meaningful empirical results. Nonparametric methods do involve making assumptions, so it is unfair to argue that likelihood-based inference 'makes assumptions' while nonparametric inference 'lets the data speak'. The issue at heart of the

Bayesian inference is always based on a parametric likelihood function and, hence, in a literal sense we should not refer to Bayesian 'nonparametric' or 'semiparametric' methods. This is the reason why the main title to this chapter is 'Flexible Models'. Nevertheless, there are many Bayesian models which are similar in spirit to non-Bayesian nonparametric methods and, thus, there is a large and growing literature which uses the name Bayesian nonparametrics. This field is too large to attempt to survey in a single chapter and, hence, we focus on two sorts of Bayesian nonparametric approaches which are particularly simple and can be done using the methods of previous chapters. The interested reader is referred to Dey, Muller and Sinha (1998) for a broader overview of Bayesian nonparametrics.

To motivate the Bayesian nonparametric approaches discussed here, it is useful to consider the assumptions underlying the Normal linear regression model. The researcher may wish to relax the assumption of a linear relationship (i.e. relax a functional form assumption) or relax the assumption of Normal errors (i.e. relax a distributional assumption). The two approaches described here relate to these two aspects. The section called *Bayesian non- and semiparametric regression* relaxes functional form assumptions, and the section on modeling with mixtures of Normals relaxes distributional assumptions. As we shall see, we can do Bayesian semiparametric regression using only techniques from Chapter 3 on the Normal linear regression model with natural conjugate prior. Modeling with mixtures of Normals can be done using a Gibbs sampler which is an extension of the one introduced in Chapter 6 (Section 6.4) for the regression model with Student-t errors.

10.2 BAYESIAN NON- AND SEMIPARAMETRIC REGRESSION

10.2.1 Overview

In Chapter 5, we discussed the nonlinear regression model

$$y_i = f(X_i, \gamma) + \varepsilon_i$$

where X_i is the ith row of X, $f(\cdot)$ is a known function which depends upon X_i and a vector of parameters, γ. In this section, we begin with a very similar starting point in that we write the nonparametric regression model as

$$y_i = f(X_i) + \varepsilon_i \tag{10.1}$$

but $f(\cdot)$ is an *unknown function*. Throughout this section, we make the standard assumptions that

---

distinction between nonparametric and likelihood-based methods is what kind of assumptions are made. For instance, a nonlinear regression model makes the assumption "the relationship between y and x takes a specific nonlinear form", whereas a nonparametric regression model makes assumptions relating to the smoothness of the regression line. The question of which sort of assumptions are more sensible can only be answered in the context of a particular empirical application.

1. ε is $N(0_N, h^{-1}I_N)$.
2. All elements of X are either fixed (i.e. not random variables) or, if they are random variables, they are independent of all elements of ε with a probability density function $p(X|\lambda)$, where λ is a vector of parameters that does not include any of the other parameters in the model.

Before discussing nonparametric regression, it is worth mentioning that nonlinear regression methods using extremely flexible choices for $f(X_i, \gamma)$ allow the researcher to achieve a goal similar to the nonparametric econometrician without the need for any new methods. For instance, by using one of the common series expansions (e.g. a Taylor series, Fourier or Muntz–Szatz expansion) one can obtain a parametric form for $f(X_i, \gamma)$ which is sufficiently flexible to approximate any unknown function. The choice of a truncation point in the series expansion allows the researcher to control the accuracy of the approximation.[3]

Nonparametric regression methods hinge on the idea that $f()$ is a smooth function. That is, if X_i and X_j are close to one another, then $f(X_i)$ and $f(X_j)$ should also be close to one another. Nonparametric regression methods, thus, estimate the nonparametric regression line by taking local averages of nearby observations. Many nonparametric regression estimators of $f(X_i)$ have the form

$$\widehat{f}(X_i) = \sum_{j \in N_i} w_j y_j \tag{10.2}$$

where w_j is the weight associated with the jth observation and N_i denotes the neighborhood around X_i. Different approaches vary in how the weights and neighborhood are defined. Unfortunately, if there are many explanatory variables, then nonparametric methods suffer from the so-called *curse of dimensionality*. That is, nonparametric methods average over 'nearby' observations to approximate the regression relationship. For a fixed sample size, as the dimension of X_i increases 'nearby' observations become further and further apart, and nonparametric methods become more and more unreliable. Thus, it is rare to see the nonparametric regression model in (10.1) directly used in applications involving many explanatory variables. Instead, various models which avoid the curse of dimensionality are used. In this section, we discuss two such models, beginning with the *partial linear model*.

10.2.2 The Partial Linear Model

The partial linear model divides the explanatory variables into some which are treated parametrically (z) and some which are treated nonparametrically (x). If x is of low dimension, then the curse of dimensionality can be overcome. The choice of which variables receive a nonparametric treatment is an application-specific one. Usually, x contains the most important variable(s) in the analysis for which it is crucial to correctly measure their marginal effect(s). Here we

---

[3] Koop, Osiewalski and Steel (1994) is a paper which implements such an approach.

assume x is a scalar, and briefly discuss below how extensions to nonscalar x can be handled.

Formally, the partial linear model is given by

$$y_i = z_i\beta + f(x_i) + \varepsilon_i \tag{10.3}$$

where y_i is the dependent variable, z_i is a vector of k explanatory variables, x_i is a scalar explanatory variable and $f(\cdot)$ is an unknown function. Note that z_i does not contain an intercept, since $f(x_1)$ plays the role of an intercept. We refer to $f()$ as the *nonparametric regression line*.

The basic idea underlying Bayesian estimation of this model is that $f(x_i)$ for $i = 1, \ldots, N$ can be treated as unknown parameters. If this is done, (10.3) is a Normal linear regression model (albeit one with more explanatory variables than observations). A Bayesian analysis of this model using a natural conjugate prior can be done exactly as described in Chapter 3. Thus, it is simple and straightforward to carry out Bayesian inference in the partial linear model.

We begin by ordering observations so that $x_1 \leq x_2 \leq \cdots \leq x_N$. Since the data points are independent of one another, their precise ordering is irrelevant and choosing to order observations in ascending order makes the definition of what a 'nearby' observation is clear. Stack all variables into matrices in the usual way as $y = (y_1, \ldots, y_N)'$, $Z = (z_1,' \ldots, z_N')'$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)'$. If we let $\gamma = (f(x_1), \ldots, f(x_N))'$,

$$W = [Z : I_N]$$

and $\delta = (\beta', \gamma')'$, then we can write (10.3) as

$$y = W\delta + \varepsilon \tag{10.4}$$

Note first that γ is an N-vector containing each point on the nonparametric regression line. At this stage, we have not placed any restrictions on the elements of γ. Hence, we are being nonparametric in the sense that $f(x_i)$ can be anything and $f()$ is a completely unrestricted unknown function. Secondly, (10.4) is simply a regression model with explanatory variables in the $N \times (N + k)$ matrix W. However, (10.4) is an unusual regression model, since there are more unknown elements in δ than there are observations, i.e. $N + k \geq N$. An implication of this is that a perfect fit is available such that the sum of squared errors is zero. For instance, if we had an estimate of δ of the form

$$\widehat{\delta} = \begin{pmatrix} 0_k \\ y \end{pmatrix}$$

then the resulting errors would all be zero. Note that $\widehat{\delta}$ implies the points on the nonparametric regression line are estimated as $\widehat{f}(x_i) = y_i$. Hence, this estimate implies no smoothing at all of the nonparametric regression line. In terms of (10.2), the implied weights are $w_i = 1$ and $w_j = 0$ for $j \neq i$. Such an estimator is unsatisfactory. Prior information can be used to surmount this pathology.

In the nonparametric regression literature, estimators are based on the idea that $f()$ is a smooth function. That is, x_i and x_{i-1} are close to one another, then $f(x_i)$ should also be close to $f(x_{i-1})$. In a Bayesian analysis, such information can be incorporated in a prior. There are many ways of doing this, but here we implement one simple approach discussed in Koop and Poirier (2002). We assume a natural conjugate Normal-Gamma prior for β, γ and h. By adopting such a choice, we are able to obtain simple analytical results which do not require posterior simulation methods. To focus on the nonparametric part of the partial linear model, we assume the standard noninformative prior for h and β:

$$p(\beta, h) \propto h \tag{10.5}$$

For the coefficients in the nonparametric part of the model, we use the partially informative prior (see Chapter 3, Exercise 4) on the first differences of γ:

$$R\delta \sim N(0_{N-1}, h^{-1}V(\eta)) \tag{10.6}$$

where $V(\eta)$ is a positive definite matrix which depends upon a hyperparameter η (which will be explained later), and $R = [0_{(N-1)\times k} : D]$, where D is the $(N-1) \times N$ first-differencing matrix:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \dots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \\ 0 & \cdots & \cdots & 0 & 0 & -1 & 1 \end{bmatrix} \tag{10.7}$$

Note that this structure implies that we only have prior information on $f(x_i) - f(x_{i-1})$. The fact that we expect nearby points on the nonparametric regression line to be similar is embedded in (10.6) through the assumption that $E[f(x_i) - f(x_{i-1})] = 0$. $V(\eta)$ can be used to control the expected magnitude of $f(x_i) - f(x_{i-1})$ and, thus, the degree of smoothness in the nonparametric regression line.

In this discussion of prior information, it is worth mentioning that the researcher sometimes wants to impose inequality restrictions on the unknown function describing the nonparametric regression line. For instance, the researcher may know that $f()$ is a monotonically increasing function. This is simple to do using the techniques described in Chapter 4 (Section 4.3).

Before presenting the posterior for this model, a brief digression on two points is called for. First, the perceptive reader may have noticed that the structure of the partial linear model is almost identical to the local level model of Chapter 8. In fact, if we omit the parametric term (i.e. drop Z) and change the i subscripts in this chapter to t subscripts, then this nonparametric regression model is identical to the state space model. This is not surprising once one recognizes that both models have ordered data and the structure in the state equation of (8.5) is identical to that of the prior given in (10.6). The fact that state space methods can be used to carry out nonparametric regression has been noted in several places (e.g. Durbin and Koopman, 2001). Everything written in Chapter 8 (Section 8.2) is thus relevant here. For instance, empirical Bayes methods can be used as described in Section 8.2.3 if the researcher does not wish to elicit prior hyperparameters such as η. Secondly, the reader with a mathematical training may be bothered by the fact that we have referred to (10.6) as controlling 'the degree of smoothness' in the nonparametric regression line through prior information about first differences. Usually, the degree of smoothness of a function is measured by its second derivative, which would suggest we use prior information about second differences (i.e. $[f(x_{i+1}) - f(x_i)] - [f(x_i) - f(x_{i-1})]$). Prior information about second differences can be incorporated in a trivial fashion by redefining D in (10.7) to be a second-differencing matrix.

It is straightforward to prove (see Chapter 3, Exercise 4), that the posterior for the Normal linear regression model with partially noninformative Normal-Gamma prior is

$$\delta, h|y \sim NG(\widetilde{\delta}, \widetilde{V}, \widetilde{s}^{-2}, \widetilde{\nu}) \tag{10.8}$$

where

$$\widetilde{V} = (R'V(\eta)^{-1}R + W'W)^{-1} \tag{10.9}$$

$$\widetilde{\delta} = \widetilde{V}(W'y) \tag{10.10}$$

$$\widetilde{\nu} = N \tag{10.11}$$

and

$$\widetilde{\nu}\widetilde{s}^2 = (y - W\widetilde{\delta})'(y - W\widetilde{\delta}) + (R\widetilde{\delta})'V(\eta)^{-1}(R\widetilde{\delta}) \tag{10.12}$$

Furthermore, the posterior is a valid p.d.f., despite the fact that the number of explanatory variables in the regression model is greater than the number of observations. Intuitively, prior information about the degree of smoothness in the nonparametric regression function suffices to correct the perfect fit pathology noted above.

In an empirical study, interest usually centers on the nonparametric part of the model. Using (10.8) and the properties of the multivariate Normal distribution (see Appendix B, Theorem B.9), it follows that

$$E(\gamma|y) = [M_Z + D'V(\eta)^{-1}D]^{-1}M_Z y \tag{10.13}$$

where $M_Z = I_N - Z(Z'Z)^{-1}Z'$. Equation (10.13) can be used as an estimate of $f()$, and we refer to it as the 'fitted nonparametric regression line'. To aid in interpretation, note that M_Z is a matrix which arises commonly in frequentist studies of the linear the regression model. $M_Z y$ are the OLS residuals from the regression of y on Z. Hence, (10.13) can be interpreted as removing the effect of y on Z (i.e. since $M_Z y$ are residuals) and then smoothing the result using the matrix $[M_Z + D'V(\eta)^{-1}D]^{-1}$. Note also that in the purely nonparametric case (i.e. Z does not enter the model), if the prior in (10.6) becomes noninformative (i.e. $V(\eta)^{-1} \rightarrow 0_{N-1,N-1}$), then $E(\gamma|y) = y$ and the nonparametric part of the model merely fits the observed data points (i.e. there is no smoothing).

So far, we have said nothing about $V(\eta)$, and many different choices are possible. A simple choice, reflecting only smoothness considerations (i.e. $f(x_i) - f(x_{i-1})$ is small), would be to take $V(\eta) = \eta I_{N-1}$.[4] This prior depends only upon the scalar hyperparameter η, which can be selected by the researcher to control the degree of smoothness. To provide more intuition on how the Bayesian posterior involves an averaging of nearby observations, it is instructive to look at $E(\gamma_i|y, \gamma^{(i)})$, where $\gamma^{(i)} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_N)$. For the pure nonparametric regression case (i.e. where Z does not enter), it can be shown that:

$$E(\gamma_i|y, \gamma^{(i)}) = \frac{1}{2+\eta}(\gamma_{i-1} + \gamma_{i+1}) + \frac{\eta}{2+\eta} y_i$$

for $i = 2, \ldots, N-1$. $E(\gamma_i|y, \gamma^{(i)})$ is a weighted average of y_i and the closest points on the nonparametric regression curve above and below i (i.e. γ_{i-1} and γ_{i+1}). Since η controls the degree of smoothness we wish to impose on $f(\cdot)$, it makes sense that as $\eta \rightarrow \infty$ we obtain $E(\gamma_i|y, \gamma^{(i)}) = y_i$ (i.e. no smoothing whatsoever). As $\eta \rightarrow 0$ we obtain $E(\gamma_i|y, \gamma^{(i)}) = \frac{1}{2}(\gamma_{i-1} + \gamma_{i+1})$. Furthermore, it can be shown that $var(\gamma_i|y, \gamma^{(i)}) = \frac{\sigma^2\eta}{2+\eta}$ which goes to zero as $\eta \rightarrow 0$. Thus, the limiting case of $\eta \rightarrow 0$ yields $\gamma_i = \frac{1}{2}(\gamma_{i-1} + \gamma_{i+1})$, and the nonparametric regression component is merely a straight line.

In summary, Bayesian inference in the partial linear model can be carried out using the familiar Normal linear regression model with natural conjugate prior if we treat the unknown points on the nonparametric regression line as parameters. Despite the fact that the number of explanatory variables in the partial linear model is greater than the number of observations, the posterior is proper. Model comparison and prediction can be done in exactly the same manner as in Chapter 3.

In many cases, the researcher may be willing to choose a particular value for η. Or, as in the following application, empirical Bayes methods as described in Chapter 8 (Section 8.2.3) can be used to estimate η. However, it is worthwhile to briefly mention another method for selecting a value for η in a data-based fashion. This new method, which is commonly used by nonparametric statisticians, is referred to as *cross-validation*. The basic idea of cross-validation is that some of the data is withheld. The model is estimated using the remaining data and used to predict the withheld data. Models are compared on the basis of how well they predict the withheld data.[5] In the present context, we could define a

---

[4] In small data sets, the distance between x_i and x_{i-1} may be large, and it might be desirable to incorporate this into the prior. A simple way of doing this would be to use a prior involving $V(\eta)$ being a diagonal matrix with (i, i)th elements equal to $v_i = \eta(x_i - x_{i-1})$.

[5] It is worth mentioning that cross-validation can be used as a model comparison/evaluation tool for any model, not just nonparametric ones.

cross-validation function as

$$CV(\eta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - E(\gamma_i|y^{(i)}))^2$$

where $y^{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N)'$. That is, we delete one observation at a time and calculate the fitted nonparametric regression line using the remaining data. We then use $(y_i - E(\gamma_i|y^{(i)}, \eta))^2$ as a metric of how well the resulting nonparametric regression line fits the left out data point. η is chosen so as to minimize the cross-validation function.

Empirical Illustration: The Partial Linear Model

To illustrate Bayesian inference in the partial linear model, we use an artificial data set using a very nonlinear data generating mechanism. For $i = 1, \ldots, 100$ we generate

$$y_i = x_i \cos(4\pi x_i) + \varepsilon_i \tag{10.14}$$

where ε_i is i.i.d. $N(0, 0.09)$ and x_i is i.i.d. $U(0, 1)$. The data is then re-ordered so that $x_1 \leq x_2 \leq \cdots \leq x_{100}$.

For simplicity, we assume a purely nonparametric model (i.e. do not include Z). The partially informative prior, given in (10.5) and (10.6), requires us to select a value for η. Once a value for η is selected, posterior inference about the nonparametric regression line can be done based on (10.8)–(10.13). Here we use the empirical Bayes methods described in Chapter 8 (Section 8.2.3) to estimate η. As stressed in Section 8.2.3, (very weak) prior information about η, γ_1 or h is required to do empirical Bayes methods in this model. Here we use prior information about η, and assume

$$\eta \sim G(\underline{\mu}_\eta, \underline{\nu}_\eta)$$

and choose nearly noninformative values of $\underline{\nu}_\eta = 0.0001$ and $\underline{\mu}_\eta = 1.0$.

Remember that empirical Bayes estimation involves finding the maximum of the marginal likelihood times $p(\eta)$ (see Chapter 8 (8.21)). With the partially informative prior, the integrating constant is not defined. However, insofar as we are interested in comparing models with different values for η, such integrating constants are irrelevant, since they cancel out in the Bayes factors. These considerations suggest that we should choose the value of η which maximizes

$$p(\eta|y) \;\propto p(y|\eta)p(\eta) \propto (|\widetilde{V}||R'V(\eta)^{-1}R|)^{\frac{1}{2}}(\widetilde{\nu s^2})^{-\frac{\widetilde{\nu}}{2}} f_G(\eta|\underline{\mu}_\eta, \underline{\nu}_\eta)$$

We do the one-dimensional maximization of $p(\eta|y)$ through a grid search. The reader who finds this brief discussion of implementing empirical Bayes methods confusing is urged to re-read Chapter 8 (Section 8.2.3) for a more thorough explanation.

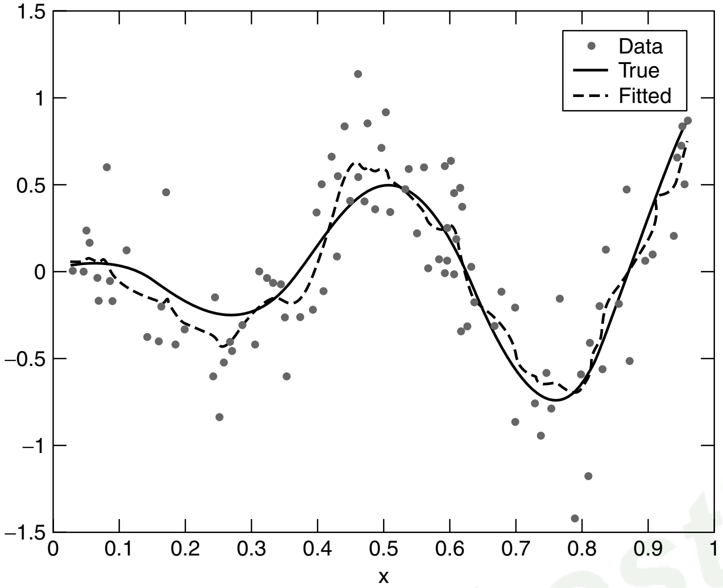

Figure 10.1 True and Fitted Nonparametric Regression Lines

The value of η chosen by the empirical Bayes procedure is 0.1648. Figure 10.1 plots the fitted nonparametric regression line using this choice for η along with the actual data and the true regression line given in (10.14) used to generate the data. It can be seen that the fitted nonparametric regression line tracks the (very nonlinear) shape of the true regression line quite well. If an empirical application requires a smoother curve, then a prior on the second differences, $[f(x_{i+1}) - f(x_i)] - [f(x_i) - f(x_{i-1})]$, can be used.

Since Chapter 3 already contains an empirical illustration using the Normal linear regression model with natural conjugate prior, no further empirical results will be presented here. Of course, all the tools presented there can be used to carry out further posterior inference (e.g. HPDIs can be presented at each point on the nonparametric regression line), model comparison (e.g. Bayes factors comparing this model to a parametric model can be calculated) or prediction. Furthermore, the perceptive reader may have noticed that this figure looks very similar to Figure 8.2, and it is worth stressing yet again that state space and nonparametric regression methods are very similar.

Extensions: Semiparametric Probit and Tobit

In this book we have emphasized the modular nature of Bayesian modeling, especially in terms of posterior simulation. In the present context, the partial linear regression model can serve as one component of a more complicated non- or semiparametric model. Many models can be written in terms of a parameter vector (possibly including latent data) θ, and the partial linear model (with parameters δ and h). Hence, the results derived in this section can be used in an MCMC algorithm to carry out Bayesian inference. That is, many models can be written such that $p(\delta, h|y, \theta)$ is Normal-Gamma, and either $p(\theta|y)$ or $p(\theta|y, \delta, h)$ can conveniently be sampled from. The list of models which can be put in this form is huge. Here we show how Bayesian semiparametric methods for probit and tobit can be developed.

Bayesian methods for a semiparametric probit model can be derived by combining the ideas of this section with results from Section 9.4 of Chapter 9 on the probit model. The semiparametric probit model can be written as

$$y_i^* = z_i\beta + f(x_i) + \varepsilon_i \tag{10.15}$$

or

$$y^* = W\delta + \varepsilon \tag{10.16}$$

where all model assumptions are the same as for the partial linear model, except that $y^* = (y_1^*, \ldots, y_N^*)'$ is unobserved. Instead, we observe

$$\begin{aligned} y_i &= 1 \text{ if } y_i^* \geq 0 \\ y_i &= 0 \text{ if } y_i^* < 0 \end{aligned} \tag{10.17}$$

Bayesian inference for this model proceeds by noting that $p(\delta, h|y^*)$ is precisely that given in (10.8)–(10.13) except that y is replaced by y^* in these formulae. In fact, if we make the usual identifying assumption that $h = 1$, the conditional posterior distribution for δ is simply Normal. Furthermore,

$$p(y^*|y, \delta, h) = \prod_{i=1}^{N} p(y_i^*|y_i, \delta, h)$$

and $p(y_i^*|y_i, \delta, h)$ is truncated Normal (see Chapter 9, Section 9.4). Hence, a simple Gibbs sampler with data augmentation which involves only the Normal and truncated Normal distributions can be used to carry out Bayesian inference.

To be precise, the MCMC algorithm involves sequentially drawing from

$$\delta|y^* \sim N(\widetilde{\delta}, \widetilde{V}) \tag{10.18}$$

and, for $i = 1, \ldots, N$,

$$\begin{aligned} y_i^*|y_i, \delta, \beta &\sim N(z_i\beta + \gamma_i, 1)1(y_i^* \geq 0) \quad \text{if } y_i = 1 \\ y_i^*|y_i, \delta, \beta &\sim N(z_i\beta + \gamma_i, 1)1(y_i^* < 0) \quad \text{otherwise} \end{aligned} \tag{10.19}$$

where $1(A)$ is the indicator function which equals 1 if condition A is true and otherwise equals 0.

Bayesian methods for a semiparametric tobit model can be derived along similar lines to semiparametric probit by combining the techniques for the partial linear model with those for parametric tobit models (see Chapter 9, Section 9.3). Comparable to (10.16) and (10.17), the semiparametric tobit model can be written as

$$y_i^* = z_i\beta + f(x_i) + \varepsilon_i \tag{10.20}$$

or

$$y^* = W\delta + \varepsilon \tag{10.21}$$

where $y^* = (y_1^*, \ldots, y_N^*)'$ is unobserved. In the tobit model, we observe

$$\begin{aligned} y_i &= y_i^* \quad \text{if } y_i^* > 0 \\ y_i &= 0 \qquad \text{if } y_i^* \leq 0 \end{aligned} \tag{10.22}$$

Bayesian inference for this model proceeds by noting that our results for the partial linear model provide us with $p(\delta, h|y^*)$. Furthermore,

$$p(y^*|y, \delta, h) = \prod_{i=1}^{N} p(y_i^*|y_i, \delta, h)$$

and $p(y_i^*|y_i, \delta, h)$ is either simply y_i or truncated Normal. Hence, a simple Gibbs sampler with data augmentation can be used to carry out Bayesian inference. Formally, the MCMC algorithm involves sequentially drawing from

$$\delta, h|y^* \sim NG(\widetilde{\delta}, \widetilde{V}, \widetilde{s}^{-2}, \widetilde{\nu}) \tag{10.23}$$

and, for $i = 1, \ldots, N$,

$$\begin{aligned} &y_i^* = y_i \quad \text{if } y_i > 0 \\ &y_i^*|y_i, \delta, \beta, h \sim N(z_i\beta + \gamma_i, h^{-1})1(y_i^* < 0) \quad \text{if } y_i = 0 \end{aligned} \tag{10.24}$$

Hence, Bayesian semiparametric probit or tobit (as well as many other models) can be carried out in a straightforward fashion using MCMC methods that combine the results for the partial linear model with some other model component.

It is also worth mentioning briefly that there is a myriad of other ways to do Bayesian non- or semiparametric regression. One particular class of model which does much the same thing as nonparametric regression is the class of *spline models*. We do not discuss them here, but refer the interested reader to Green and Silverman (1994), Silverman (1985), Smith and Kohn (1996) or Wahba (1983). There are many other methods for flexible modeling on a regression function which are not discussed in this book. The interested reader is referred to Dey, Muller and Sinha (1998) for a discussion of some of these models and methods.

10.2.3 An Additive Version of the Partial Linear Model

Thus far we have assumed x_i to be a scalar in the partial linear model. In this scalar case, the prior used to impose smoothness on the nonparametric regression line involved simply reordering the observations so that $x_1 \leq \cdots \leq x_N$. As discussed at the beginning of this chapter, when x_i is a vector the curse of dimensionality may preclude sensible nonparametric inference. However, if x_i is of low dimension, then it may be possible to implement Bayesian inference by using a nearest neighbor algorithm to measure the distance between observations. The data can then be reordered according to the distance between observations and the posterior given in (10.8) used to carry out Bayesian inference. For instance, a common definition of the distance between observations i and j is

$$dist_{i,j} = \sum_{l=1}^{p} (x_{il} - x_{jl})^2$$

where $x_i = (x_{i1}, \ldots, x_{ip})'$ is a p-vector. The procedure for ordering the data involves selecting a first observation (e.g. the observation with the minimum value for the first element of x). The second observation is the one which is closest to the first observation. The third observation is the one closest to the second (after deleting the first observation), etc. Once the data have been ordered, the Bayesian procedure described above can be used. However, if p is large (e.g. $p > 3$), then this procedure may work very poorly (and may be sensitive to the choice of first observation and the definition of distance between observations). Accordingly, many variants of the partial linear model have been proposed which place restrictions on $f()$ to break the curse of dimensionality. Here we describe one common model, and develop Bayesian methods for carrying out econometric inference.

The additive version of the partial linear model is given by

$$y_i = z_i\beta + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i \qquad (10.25)$$

where $f_j(\cdot)$ for $j = 1, \ldots, p$ are unknown functions. In other words, we are restricting the nonparametric regression line to be additive in p explanatory variables:

$$f(x_i) = f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip})$$

In many applications, such an additivity assumption may be sensible, and it is definitely much more flexible than the linearity assumption of standard regression methods.

Extending the notation described between (10.3) and (10.4), we can write this model as

$$y = Z\beta + \gamma_1 + \gamma_2 + \cdots + \gamma_p + \varepsilon \qquad (10.26)$$

where $\gamma_j = (\gamma_{1j}, \ldots, \gamma_{Nj})' = [f_j(x_{1j}), \ldots, f_j(x_{Nj})]'$. In other words, the N points on the nonparametric regression line corresponding to the jth explanatory variable are stacked in γ_j for $j = 1, \ldots, p$. The data are ordered according to the first explanatory variable so that $x_{11} \leq x_{21} \leq \cdots \leq x_{N1}$. We refer to this ordering below as the ‘correct’ ordering.

In the case where x was a scalar, we used the simple intuition that, if we ordered the data points so that $x_1 \leq x_2 \leq \cdots \leq x_N$, then it was sensible to put a prior on $f(x_i) - f(x_{i-1})$. Here we have p explanatory variables which can be used to order the observations, so there is not one simple ordering which can be adopted. However, remember that the ordering information was only important as a way of expressing prior information about the degree of smoothness of the nonparametric regression line. If we express prior information for each of $\gamma_1, \ldots, \gamma_p$ with observations ordered according to its own explanatory variable, then transform back to the correct ordering, we can carry out Bayesian inference in a manner virtually identical to that in Section 10.2.2. To emphasize the intuition, let me repeat the econometric strategy in slightly different words. With independent data, it does not matter how the data is ordered, provided all variables are ordered in the same way. Here we have our observations ordered as $x_{11} \leq x_{21} \leq \cdots \leq x_{N1}$. However, prior information on the degree of smoothness for $f_j()$ should be elicited with observations ordered so that $x_{1j} \leq x_{2j} \leq \cdots \leq x_{Nj}$. But this means that, for $j = 2, \ldots, p$, the prior will be elicited with the observations ordered incorrectly (i.e. the correct ordering does not have $x_{1j} \leq x_{2j} \leq \cdots \leq x_{Nj}$, but rather has $x_{11} \leq x_{21} \leq \cdots \leq x_{N1}$). How do we solve this problem? After eliciting each prior, we simply re-order the data back to the correct ordering. Once we have done this, we are back in the familiar world of the Normal linear regression model with natural conjugate prior.

To write out this strategy formally, some new notation is required. Remember that our previous notation (e.g. $\gamma_1, \ldots, \gamma_p$) used an ordering of observations such that $x_{11} \leq x_{21} \leq \cdots \leq x_{N1}$. Define $\gamma_j^{(j)}$ as being equal to γ_j with observations ordered according to the jth explanatory variable (i.e. all data is ordered so that $x_{1j} \leq x_{2j} \leq \cdots \leq x_{Nj}$ for $j = 2, \ldots, p$). For individual elements of $\gamma_j^{(j)}$ we use the notation

$$\gamma_j^{(j)} = \begin{pmatrix} \gamma_{1j}^{(j)} \\ \gamma_{2j}^{(j)} \\ \cdot \\ \cdot \\ \gamma_{Nj}^{(j)} \end{pmatrix} = \begin{pmatrix} \gamma_{1j}^{(j)} \\ \gamma_j^{(j*)} \end{pmatrix}$$

That is, we have isolated out the first point on the jth component of the nonparametric regression line ($\gamma_{1j}^{(j)}$) from all the remaining points which we stack in an $(N-1)$-vector $\gamma_j^{(j*)}$. We define a similar notation when the observations are ordered according to the first explanatory variable with $\gamma_j^{(*)}$ equaling γ_j with one element deleted. This element is the one corresponding to the smallest value of the jth explanatory variable.

Before formally deriving the requisite posterior, it is important to note that there is an identification problem with the additive model, in that constants may be added and subtracted appropriately without changing the likelihood. For instance, the models $y_i = f_1(x_{i1}) + f_2(x_{i2}) + \varepsilon_i$ and $y_i = g_1(x_{i1}) + g_2(x_{i2}) + \varepsilon_i$ are equivalent if $g_1(x_{i1}) = f_1(x_{i1}) + c$ and $g_2(x_{i2}) = f_2(x_{i2}) - c$, where c is any constant. Insofar as interest centers on the marginal effect of each variable on y (i.e. on the shapes of $f_j(x_{ij})$ for $j = 1, \ldots, p$) or the overall fit of the nonparametric regression model, the lack of identification is irrelevant. Here we impose identification in a particular way, but many other choices can be made, and the interpretation of the empirical results will not change in any substantive way. We impose identification by setting $\gamma_{1j}^{(j)} = 0$ for $j = 2, \ldots, p$ (i.e. all except the first additive functions are restricted to have intercepts equalling zero).

For γ_1, β and h we use the same partially informative prior as before. In particular, the noninformative prior for β and h is given in (10.5) and, for γ_1 (i.e. the nonparametric regression line corresponding to the first explanatory variable) we use the prior on the degree of smoothness

$$D\gamma_1 \sim N(0_{N-1}, h^{-1}V(\eta_1)) \tag{10.27}$$

where D is the first-differencing matrix defined in (10.7). For $\gamma_j^{(j)}$ for $j = 2, \ldots, p$ the smoothness prior can be written as

$$D\gamma_j^{(j)} \sim N(0_{N-1}, h^{-1}V(\eta_j)) \tag{10.28}$$

Alternatively, since we impose the identifying assumption $\gamma_{j1}^{(j)} = 0$, we can write (10.28) as

$$D^*\gamma_j^{(j*)} \sim N(0_{N-1}, h^{-1}V(\eta_j)) \tag{10.29}$$

where D^* is an $(N-1)\times(N-1)$ matrix equal to D with the first column removed. Note that, as desired (10.28) and (10.29) imply that if $x_{i-1,j}$ and x_{ij} are close to one another, then $f_j(x_{i-1,j})$ and $f_j(x_{i,j})$ should also be close to one another. As discussed previously, other priors can be used (e.g. D can be replaced with the second-differencing matrix) with minimal changes in the following posteriors.

The prior in (10.28) is for $j = 2, \ldots, p$, and is expressed using the observations ordered in an incorrect manner (i.e. they are ordered as $x_{1j} \le x_{2j} \le \cdots \le x_{Nj}$), so we have to re-order them before proceeding further. Hence, we define D_j, which is equivalent to D except that the rows and columns are re-ordered so that observations are ordered correctly (i.e. as $x_{11} \le x_{21} \le \cdots \le x_{N1}$). We also introduce the notation D_j^* which is comparable to D^*. That is, D_j^* is equal to D_j with the column corresponding to the first point on the nonparametric regression line removed.

A concrete example of how this works might help. Suppose we have $N = 5$ and two explanatory variables which have values:

$$X = \begin{bmatrix} 1 & 3 \\ 2 & 4 \\ 3 & 1 \\ 4 & 2 \\ 5 & 5 \end{bmatrix}$$

The data has been ordered in the correct manner so that the first explanatory variable is in ascending order, $x_{11} \leq \cdots \leq x_{51}$. However, when observations are ordered in this way, the second explanatory variable is not in ascending order. The prior given in (10.28), written for the observations ordered according to $x_{12} \leq \cdots \leq x_{52}$, must be rearranged to account for this. This involves creating a rearranged version of D:

$$D_2 = \begin{bmatrix} 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{bmatrix}$$

It can be verified that $D_2\gamma_2$ defines the distance between neighboring values for the second explanatory variable and, thus, it is sensible to put smoothness prior on it. The identification restriction implies $\gamma_{32} = 0$ and, hence,

$$D_2^* = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

In summary, with the additive model we use the same smoothness prior on each of p unknown functions. Since the observations are ordered so that $x_{11} \leq \cdots \leq x_{N1}$, the smoothness prior for γ_1 can be written using the first difference matrix D. However, for $\gamma_2, \ldots, \gamma_p$ the same smoothness prior must be written in terms of a suitably rearranged version of D. We label these rearranged first difference matrices D_j for $j = 2, \ldots, p$. Imposing the identification restriction involves removing the appropriate column of D_j, and we label the resulting matrix D_j^*.

One more piece of notation relating to the imposition of the identification restriction is required. Let I_j^* equal the $N \times N$ identity matrix with one column deleted. The column deleted is for the observation which has the lowest value for the jth explanatory variable.

With this notation established, we can proceed in a similar manner as for the partial linear model. The model can be written as a Normal linear regression model:

$$y = W\delta + \varepsilon \tag{10.30}$$

where

$$W = [Z : I_N : I_2^* : \ldots : I_p^*]$$

and $\delta = (\beta', \gamma_1', \gamma_2^{(*)'}, \ldots, \gamma_p^{(*)'})'$ contains $K = k + N + (p-1) \times (N-1)$ regression coefficients. The prior for this model can be written in compact notation as

$$R\delta \sim N(0_{p(N-1)}, h^{-1}\underline{V}) \tag{10.31}$$

where

$$R = \begin{bmatrix} 0_{(N-1)\times k} & D & 0 & \cdot & \cdot & 0 \\ 0_{(N-1)\times k} & 0 & D_2^* & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 & \cdot \\ \cdot & \cdot & \cdot & 0 & \cdot & 0 \\ 0_{(N-1)\times k} & \cdot & \cdot & 0 & \cdot & D_p^* \end{bmatrix}$$

and

$$\underline{V} = \begin{bmatrix} V(\eta_1) & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & 0 & V(\eta_p) \end{bmatrix}$$

At this point, it is useful to stress that, although the notation has become complicated due to questions of identification and the ordering of observations, this is still simply a Normal linear regression model with natural conjugate prior. Thus, all of the familiar results and techniques for this model are relevant, and we have

$$\delta, h|y \sim NG(\widetilde{\delta}, \widetilde{V}, \widetilde{s}^{-2}, \widetilde{\nu}) \tag{10.32}$$

where

$$\widetilde{V} = (R'\underline{V}^{-1}R + W'W)^{-1} \tag{10.33}$$

$$\widetilde{\delta} = \widetilde{V}(W'y) \tag{10.34}$$

$$\widetilde{\nu} = N \tag{10.35}$$

and

$$\tilde{\nu}\tilde{s}^2 = (y - W\widetilde{\delta})'(y - W\widetilde{\delta}) + (R\widetilde{\delta})'\underline{V}^{-1}(R\widetilde{\delta}) \tag{10.36}$$

Bayesian inference in this additive model is complicated by the fact that it is potentially difficult to elicit prior hyperparameters in a data-based fashion. Note that the prior allows for a different degree of smoothing in each unknown function (i.e. we have η_j for $j = 1, \ldots, p$). In some cases, the researcher may have prior information that allows her to choose values for each η_j. However, in many cases it will be sensible to smooth each unknown function by the same amount (i.e. setting $\eta_1 = \cdots = \eta_p \equiv \eta$ will be reasonable), and only one prior hyperparameter needs to be chosen. Empirical Bayesian inference can be carried out exactly as for the partial linear model. Model comparison and prediction can be carried out using the familiar methods for the Normal linear regression model.

Illustration: An Additive Model

To illustrate Bayesian inference in the partial linear model with additive nonparametric regression line, we generate artificial data from

$$y_i = f_1(x_{i1}) + f_2(x_{i2}) + \varepsilon_i$$

for $i = 1, \ldots, 100$, where ε_i is i.i.d. $N(0, 0.09)$ and x_{i1} and x_{12} are i.i.d. $U(0, 1)$. We take

$$f_1(x_{i1}) = x_i \cos(4\pi x_{i1})$$

and

$$f_2(x_{i2}) = \sin(2\pi x_{i2})$$

The partially informative prior in (10.31) requires elicitation of prior hyperparameters η_1 and η_2. We set $\eta \equiv \eta_1 = \eta_2$, and use the same empirical Bayesian methods as for the partial linear model to select a value for η. This value for η is then used to make posterior inference about the two components of the nonparametric regression line using (10.32)–(10.36).

As in the previous section, we use (very weak) prior information about η. In particular, we assume

$$\eta \sim G(\underline{\mu}_\eta, \underline{\nu}_\eta)$$

and choose nearly noninformative values of $\underline{\nu}_\eta = 0.0001$ and $\underline{\mu}_\eta = 1.0$. We choose the value of η which maximizes

$$p(\eta|y) \propto p(y|\eta)p(\eta) \propto (|\widetilde{V}||R'\underline{V}^{-1}R|)^{\frac{1}{2}}(\widetilde{\nu s^2})^{-\frac{\widetilde{\nu}}{2}} f_G(\eta|\underline{\mu}_\eta, \underline{\nu}_\eta)$$

The value of η chosen by the empirical Bayes procedure is 0.4210. Figures 10.2a and b plot the fitted and true nonparametric regression lines for each of the two additive functions in our nonparametric regression model (i.e. $E(\gamma_j|y)$ and $f_j(x_{ij})$ for $j = 1, 2$). These figures indicate that we are successfully estimating $f_j(\cdot)$. Remember that the identifying restriction means we can only estimate the functions up to an additive constant. This is reflected in the slight shifting of the two components of the fitted nonparametric regression lines in Figures 10.2a and b. As noted in our illustration of the partial linear model, if the researcher requires a smoother curve, then a prior on the second differences, $[f(x_{i+1}) - f(x_i)] - [f(x_i) - f(x_{i-1})]$, can be used. Furthermore, in a serious empirical application other posterior features (e.g. HPDIs), model comparison tools (e.g. Bayes factors comparing this model to a parametric model) or predictive distributions could be presented.

Extensions

With the partial linear model we noted that many extensions were possible that would allow for Bayesian inference in, for example, semiparametric probit or

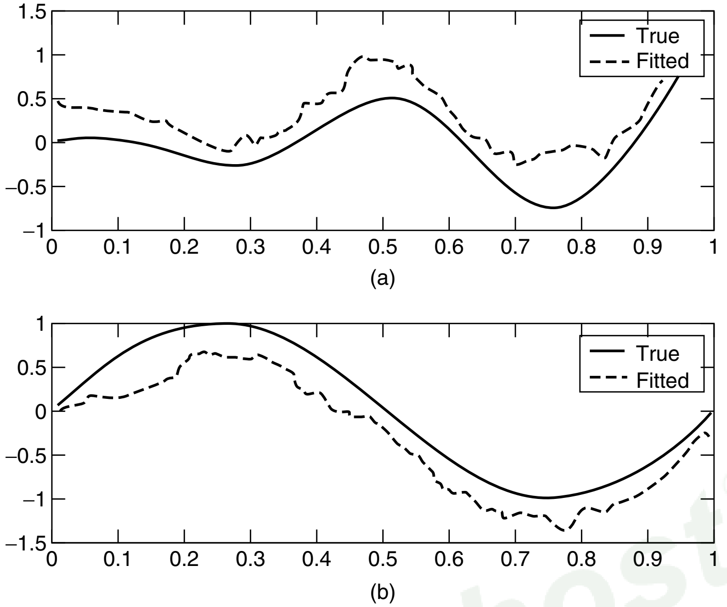

Figure 10.2 True and Fitted Lines for (a) First and (b) Second Additive Term

tobit models. With the additive variant of the partial linear model, the exact same extensions can be done in the same manner.

10.3 MIXTURES OF NORMALS MODELS

10.3.1 Overview

The partial linear model and its additive variant allowed for *the regression line* to have an unknown functional form. There are also many techniques for allowing *whole distributions* to have unknown forms. Here, we describe one such set of techniques. The basic idea underlying the model in this section is that a very flexible distribution can be obtained by mixing together several distributions. The resulting flexible distribution can be used to approximate the unknown distribution of interest. In this section, we discuss mixtures of Normal distributions, since these are commonly used and simple to work with. However, it should be mentioned that any set of distributions can be mixed, resulting in a more flexible distribution than would be obtained by simply choosing a single distribution.

The models considered in this section are not 'nonparametric', in the sense that they cannot become any unknown distribution. This is because they are so-called *finite mixtures of Normals*. For instance, a distribution which mixes five different Normal distributions, although very flexible, cannot accommodate any possible distribution. Thus, finite mixtures of Normals should be considered only as an extremely flexible modeling strategy. However, we note that *infinite mixtures* are, to all intents and purposes, nonparametric. Infinite mixtures of Normals will not be discussed here. Robert (1996), which is Chapter 24 of *Markov Chain Monte Carlo in Practice,* provides an introduction to this area of Bayesian statistics. A particular infinite mixture model involving *Dirichlet process priors* is very popular. Escobar and West (1995) and West, Muller and Escobar (1994) provide thorough discussions of this model.

We have already seen one particular example of a mixture of Normals model. Chapter 6 (Section 6.4) considered the case of the linear regression model with independent Student-t errors, and showed how it could be obtained using a particular mixture of Normals. Since the Student-t distribution is more flexible than the Normal (i.e. the Normal is a special case of the Student-t which arises when the degrees of freedom parameter goes to infinity), Section 6.4 provides a simple example of how mixing Normals can lead to a more flexible distribution. Here we consider more general mixtures of Normals in the context of the linear regression model. However, the basic concepts can be used anywhere the researcher wishes to make a flexible distributional assumption. For instance, the panel data models in Chapter 7 assumed hierarchical priors having particular distributions for the individual effects (e.g. Normal in (7.7) and exponential in (7.46) for the stochastic frontier model). Mixtures of Normals can be used to make these hierarchical priors more flexible. Posterior simulation can be done by combining the relevant components of the Gibbs sampler outlined below with the appropriate Gibbs sampler from Chapter 7. Geweke and Keane (1999) offers another nice use of mixtures of Normals in that it develops a mixtures of Normals probit model.

10.3.2 The Likelihood Function

The linear regression model can be written as

$$y = X\beta + \varepsilon \tag{10.37}$$

where the notation is the same as in previous chapters (e.g. see Chapter 3, Section 3.2). The likelihood function for the Normal linear regression model was based on the assumptions that

1. ε_i is i.i.d. $N(0, h^{-1})$ for $i = 1, \ldots, N$.
2. All elements of X are either fixed (i.e. not random variables) or, if they are random variables, they are independent of all elements of ε with a probability density function, $p(X|\lambda)$ where λ is a vector of parameters that does not include β and h.

Here we replace the first assumption with one where ε_i is a mixture of m different distributions. That is,

$$\varepsilon_i = \sum_{j=1}^{m} e_{ij} \left(\alpha_j + h_j^{\frac{1}{2}} \eta_{ij} \right) \tag{10.38}$$

where η_{ij} is i.i.d. $N(0, 1)$ for $i = 1, \ldots, N$, $j = 1, \ldots, m$ and e_{ij}, α_j and h_j are all parameters. e_{ij} indicates the component in the mixture that the ith error is drawn from. That is $e_{ij} = 0$ or 1 for $j = 1, \ldots, m$ and $\sum_{j=1}^{m} e_{ij} = 1$. Since η_{ij} is Normal, it follows that $(\alpha_j + h_j^{-\frac{1}{2}} \eta_{ij})$ is a Normal random variable with mean α_j and precision h_j. Thus, (10.38) specifies that the regression error is a weighted average of m different distributions. Each of these component distributions is $N(\alpha_j, h_j^{-1})$. This motivates the terminology *mixture of Normals*. The special case where $\alpha_j = 0$ for all j is referred to as a *scale mixture of Normals*. The special case where $h_1 = \cdots = h_m$ is referred to as a *mean (or location) mixture of Normals*. The mixture of Normals used in Chapter 6 (Section 6.4) was a scale mixture of Normals involving a particular hierarchical prior. To simplify notation, we stack these new parameters into vectors in the usual way: $\alpha = (\alpha_1, \ldots, \alpha_m)'$, $h = (h_1, \ldots, h_m)'$, $e_i = (e_{i1}, \ldots, e_{im})'$ and $e = (e_1', \ldots, e_N')'$.

In practice, it is unknown which component the ith error is drawn from and, thus, we let p_j for $j = 1, \ldots, m$ be the probability of the error being drawn from the jth component in the mixture. That is, $p_j = P(e_{ij} = 1)$. Formally, this means that e_i are i.i.d. draws from the Multinomial distribution (see Appendix B, Definition B.23):

$$e_i \sim M(1, p) \tag{10.39}$$

where $p = (p_1, \ldots, p_m)'$. Remember that, since p is a vector of probabilities, we must have $0 \leq p_j \leq 1$ and $\sum_{j=1}^{m} p_j = 1$.

As with many models, there is some arbitrariness as to what gets labeled 'prior' information and what gets labeled 'likelihood' information. Equation (10.39) could be interpreted as a hierarchical prior for e_i. However, following standard practice, here we refer to β, h, α and p as the parameters of the model and $p(y|\beta, h, \alpha, p)$ as the likelihood function. The component indicators, e_i for $i = 1, \ldots, N$, will be treated as latent data (and will prove useful in the Gibbs sampling algorithm outlined below). Since p_j is the probability of the error being drawn from the jth component in the Normal mixture, it can be seen that the likelihood function is

$$p(y|\beta, h, \alpha, p) = \frac{1}{(2\pi)^{\frac{N}{2}}} \prod_{i=1}^{N} \left\{ \sum_{j=1}^{m} p_j \sqrt{h_j} \exp\left[-\frac{h_j}{2} (y_i - \alpha_j - \beta' x_i)^2 \right] \right\} \tag{10.40}$$

where x_i is a k-vector containing the explanatory variables for individual i.

10.3.3 The Prior

As with any Bayesian model, any prior can be used. Here we describe a commonly-used prior which allows for convenient computation and is flexible enough to accommodate a wide range of prior beliefs. However, before describing precise forms for prior densities, there are two underlying issues which must be discussed.

First, the mixtures of Normals model is an example of a model where the likelihood function is unbounded.[6] This means that the standard frequentist theory underlying maximum likelihood estimation breaks down. For the Bayesian, the pathology implies that the researcher should not use a noninformative prior. Bayesian inference with an informative prior, however, can be done in the usual way.[7]

Secondly, there is an identification problem in this model, in that multiple sets of parameter values are consistent with the same likelihood function. For instance, consider a mixture with two components (i.e. $m = 2$). The probabilities associated with each component are $p_1 = 0.25$ and $p_2 = 0.75$. The first distribution in the mixture has $\alpha_1 = 2.0$ and $h_1 = 2.0$, while the second has $\alpha_2 = 1.0$ and $h_2 = 1.0$. This distribution is identical to one where the labeling of the two components is reversed. That is, it is exactly the same as one with parameter values $p_1 = 0.75$, $p_2 = 0.25$, $\alpha_1 = 1.0$, $h_1 = 1.0$, $\alpha_2 = 2.0$ and $h_2 = 2.0$. Because of this, it is necessary for the prior to impose a *labelling restriction*, such as

$$\alpha_{j-1} < \alpha_j \tag{10.41}$$

$$h_{j-1} < h_j \tag{10.42}$$

or

$$p_{j-1} < p_j \tag{10.43}$$

for $j = 2, \ldots, m$. Only one such restriction need be imposed. Here (10.41) will be chosen, although imposing (10.42) or (10.43) will only cause minor modification in the following material.

We begin with a prior for β and h, which is a simple extension of the familiar independent Normal-Gamma prior (see Chapter 4, Section 4.2). In particular,

$$\beta \sim N(\underline{\beta}, \underline{V}) \tag{10.44}$$

and we assume independent Gamma priors for h_j for $j = 2, \ldots, m$,

$$h_j \sim G(\underline{s}_j^{-2}, \underline{\nu}_j) \tag{10.45}$$

---

[6] To see this, set β to $\widehat{\beta}$, the OLS estimate, h_j^{-1} to the OLS estimate of the error variance and $\alpha_j = 0$ for $j = 2, \ldots, m$. For some $c > 0$ set $p_1 = c$ and $p_j = \frac{1-c}{m-1}$ for $j = 2, \ldots, m$. If $\alpha_1 = (y_1 - \widehat{\beta}' x_1)$, then the likelihood function goes to infinity as $h_1 \to \infty$.

[7] A proof of this statement is provided in Geweke and Keane (1999) for the prior used in this section.

The Dirichlet distribution (see Appendix B, Definition B.28) is a flexible and computationally convenient choice for parameters such as p which lie between zero and one and sum to one (remember that $0 \leq p_j \leq 1$ and $\sum_{j=1}^{m} p_j = 1$). Thus, we take

$$p \sim D(\underline{p}) \tag{10.46}$$

where $\underline{p}$ is an m-vector of prior hyperparameters. Appendix B, Theorem B.17 lists some properties which show how $\underline{p}$ can be interpreted.

Here we impose the labeling restriction through α. Hence, we assume the prior for this parameter vector to be Normal with the restrictions in (10.41) imposed:

$$p(\alpha) \propto f_N(\alpha|\underline{\alpha}, \underline{V}_{\alpha})1(\alpha_1 < \alpha_2 < \cdots < \alpha_m) \tag{10.47}$$

Remember that $1(A)$ is the indicator function equalling 1 if condition A holds and otherwise equalling zero.

10.3.4 Bayesian Computation

As with many other models in this book, Bayesian inference can be carried out using a Gibbs sampler with data augmentation. Intuitively, if we knew which component in the mixture each error was drawn from, then the model would reduce to the Normal linear regression model with independent Normal-Gamma prior (see Chapter 4, Section 4.2). Thus, treating e as latent data will greatly simplify things. This intuition motivates a Gibbs sampler which sequentially draws from the full posterior conditional distributions $p(\beta|y, e, h, p, \alpha)$, $p(h|y, e, \beta, p, \alpha)$, $p(p|y, e, \beta, h, \alpha)$, $p(\alpha|y, e, \beta, h, p)$ and $p(e|y, \beta, h, p, \alpha)$. Below we derive the precise form for each of these distributions. These derivations are relatively straightforward, involving multiplying the appropriate prior times $p(y|e, \beta, h, \alpha, p)$ and re-arranging the result. Using methods comparable to those used to derive (10.40), it can be shown that

$$p(y|e, \beta, h, \alpha, p) = \frac{1}{(2\pi)^{\frac{N}{2}}} \prod_{i=1}^{N} \left\{ \sum_{j=1}^{m} e_{ij} \sqrt{h_j} \exp\left[-\frac{h_j}{2}(y_i - \alpha_j - \beta' x_i)^2 \right] \right\} \tag{10.48}$$

Conditional on e, $p(\beta|y, e, h, p, \alpha)$ and $p(h_j|y, e, \beta, p, \alpha)$ for $j = 1, \ldots, m$ simplify, and results from Chapter 4, Section 4.2 can be applied directly. In particular, $p(\beta|y, e, h, p, \alpha)$ does not depend upon p and

$$\beta|y, e, h, \alpha \sim N(\overline{\beta}, \overline{V}) \tag{10.49}$$

where

$$\overline{V} = \left(\underline{V}^{-1} + \sum_{i=1}^{N} \sum_{j=1}^{m} e_{ij} h_j x_i x_i' \right)^{-1}$$

and

$$\overline{\beta} = \overline{V}\left(\underline{V}^{-1}\underline{\beta} + \sum_{i=1}^{n}\sum_{j=1}^{m} e_{ij} h_j x_i [y_i - \alpha_j]\right)$$

Furthermore, for $j = 1, \ldots, m$, the posterior conditionals for the $h_j s$ are independent of one another and simplify to

$$h_j | y, e, \beta, \alpha \sim G(\overline{s}_j^{-2}, \overline{\nu}_j) \tag{10.50}$$

where

$$\overline{\nu}_j = \sum_{i=1}^{N} e_{ij} + \underline{\nu}_j$$

and

$$\overline{s}_j^2 = \frac{\sum_{i=1}^{N} e_{ij}(y_i - \alpha_j - x_i'\beta)'(y_i - \alpha_j - x_i'\beta) + \underline{\nu}_j \underline{s}_j^2}{\overline{\nu}_j}$$

To aid in interpretation, remember that e_{ij} is an indicator variable equalling 1 if the ith error comes from the jth component in the mixture. Hence, $\sum_{i=1}^{N} e_{ij}$ simply counts the number of observations in the jth component, the term $\sum_{i=1}^{N}\sum_{j=1}^{m} e_{ij} h_j x_i x_i'$ is comparable to the term $hX'X$ in Chapter 4 (4.4), but for the ith observation it picks out the appropriate h_j. Other terms have similar intuition.

Noting that α_j enters in the role of an intercept from a Normal linear regression model in (10.48) and (10.47) describes a Normal prior (subject to the labelling restrictions), it can be seen that the conditional posterior of α is Normal (subject to the labeling restrictions). In particular,

$$p(\alpha | y, e, \beta, h) \propto f_N(\alpha | \overline{\alpha}, \overline{V}_\alpha) 1(\alpha_1 < \alpha_2 < \cdots < \alpha_m) \tag{10.51}$$

where

$$\overline{V}_\alpha = \left(\underline{V}_\alpha^{-1} + \sum_{i=1}^{N}\left\{\sum_{j=1}^{m} e_{ij} h_j\right\} e_i e_i'\right)^{-1}$$

and

$$\overline{\alpha} = \overline{V}_\alpha\left[\underline{V}_\alpha^{-1}\underline{\alpha} + \sum_{i=1}^{N}\left\{\sum_{j=1}^{m} e_{ij} h_j\right\} e_i (y_i - \beta' x_i)\right]$$

These formulae may look somewhat complicated, but they are calculated using methods which are minor modifications of those used for the Normal linear regression model. The term $\{\sum_{j=1}^{m} e_{ij} h_j\}$ picks out the relevant error precision for observation i.

Multiplying (10.46) by (10.48) yields the kernel of the conditional posterior, $p(p|y, e, \beta, h, \alpha)$. Straightforward manipulations show that this only depends upon e, and has a Dirichlet distribution

$$p \sim D(\overline{\rho}) \tag{10.52}$$

where

$$\overline{\rho} = \underline{\rho} + \sum_{i=1}^{N} e_i$$

Remember that e_i shows which component in the mixture the ith error is drawn from. It is an m-vector containing all zeros except for a 1 in the appropriate location. Thus $\sum_{i=1}^{N} e_i$ is an m-vector containing the number of observations drawn from each Normal distribution in the mixture.

The last block in the Gibbs sampler is $p(e|y, \beta, h, p, \alpha)$. The rules of conditional probability imply $p(e|y, \beta, h, p, \alpha) \propto p(y|e, \beta, h, p, \alpha)p(e|\beta, h, p, \alpha)$. The prior independence assumptions imply $p(e|\beta, h, p, \alpha) = p(e|p)$ and, thus, $p(e|y, \beta, h, p, \alpha)$ can be obtained by multiplying (10.48) by (10.39) and re-arranging. If this is done, we find that $p(e|y, \beta, h, p, \alpha) = \prod_{i=1}^{N} p(e_i|y, \beta, h, p, \alpha)$, and each of the $p(e_i|y, \beta, h, p, \alpha)$ is a Multinomial density (see Appendix B, Definition B.23). To be precise,

$$e_i|y, \beta, h, p, \alpha \sim$$

$$M\left(1, \left[\frac{p_1 f_N(y_i|\alpha_1 + \beta' x_i, h_1^{-1})}{\sum_{j=1}^{m} p_j f_N(y_i|\alpha_j + \beta' x_i, h_j^{-1})}, \ldots, \frac{p_m f_N(y_i|\alpha_m + \beta' x_i, h_m^{-1})}{\sum_{j=1}^{m} p_j f_N(y_i|\alpha_j + \beta' x_i, h_j^{-1})}\right]'\right) \tag{10.53}$$

Posterior inference in the linear regression model with mixture of Normals errors can be carried out using a Gibbs sampler which sequentially draws from (10.49), (10.50), (10.51), (10.52) and (10.53).

10.3.5 Model Comparison: Information Criteria

All the model comparison methods described in previous chapters can be used with mixtures of Normals. With this class of model, an important issue is the selection of m, the number of components in the mixture. This can be done by calculating the marginal likelihood for a range of values for m and choosing the value which yields the largest marginal likelihood. Either the Gelfand–Dey method (see Chapter 5, Section 5.7) or the Chib method (see Chapter 7, Section 7.5) can be used to calculate the marginal likelihood. A minor complication arises, since both these methods require the evaluation of prior densities, and the labeling restriction means that (10.47) only gives us the prior kernel for α. However, the necessary integrating constant can be calculated using prior simulation. A crude prior simulator would simply take draws from $f_N(\alpha|\underline{\alpha}, \underline{V}_\alpha)$ and calculate the proportion of draws which satisfy $\alpha_1 < \alpha_2 < \cdots < \alpha_m$. One over this proportion is the required integrating constant. More efficient prior simulators can be developed using algorithms for drawing from the truncated Normal.

However, calculating marginal likelihoods can be computationally demanding and care has to be taken with prior elicitation (e.g. marginal likelihoods are usually not defined when using noninformative priors). Accordingly, interest exists in shortcut methods for summarizing the data evidence in favor of a model. Motivated by this consideration, various *information criteria* have been developed. In this section, a few of these will be introduced. Their advantage is that they are easy to calculate, and typically do not depend on prior information. Their disadvantage is that it is hard to provide a rigorous justification for their use. That is, the logic of Bayesian inference says that a model should be evaluated based on the probability that it generated the data. Hence, for the pure Bayesian, the posterior model probability should be the tool for model comparison. Information criteria do not have such a formal justification (at least from a Bayesian perspective). However, as noted below, they can often be interpreted as approximations to quantities which have a formal Bayesian justification.

Information criteria can be used with any model. Accordingly, let us temporarily adopt the general notation of Chapter 1, where θ is a p-vector of parameters and $p(y|\theta)$, $p(\theta)$ and $p(\theta|y)$ are the likelihood, prior and posterior, respectively. Information criteria typically have the form

$$IC(\theta) = 2\ln[p(y|\theta)] - g(p) \tag{10.54}$$

where $g(p)$ is an increasing function of p. The traditional use of information criteria involves evaluating $IC(\theta)$ at a particular point (e.g. the maximum likelihood value for θ) for every model under consideration, and choosing the model with the highest information criteria. Most information criteria differ in the functional form used for $g(p)$. This is a function which rewards parsimony. That it, it penalizes models with excessive parameters.

In Bayesian circles, the most common information criterion is the *Bayesian Information Criterion* (or BIC)

$$BIC(\theta) = 2\ln[p(y|\theta)] - p\ln(N) \tag{10.55}$$

As shown in Schwarz (1978), twice the log of the Bayes factor comparing two models is approximately equal to the difference in BICs for the two models. Two other popular information criteria are the *Akaike Information Criterion* (or AIC), given by

$$AIC(\theta) = 2\ln[p(y|\theta)] - 2p \tag{10.56}$$

and the *Hannan–Quinn Criterion* (or HQ)

$$HQ(\theta) = 2\ln[p(y|\theta)] - pc_{HQ}\ln[\ln(N)] \tag{10.57}$$

In (10.57) c_{HQ} is a constant. HQ is a consistent model selection criterion[8] if $c_{HQ} > 2$.

These are the most popular of the many information criteria which exist. There are many places where the interested reader can find out more. The discussion and citations in Poirier (1995, p. 394) provide a good starting point. Kass and Raftery (1995) is a fine survey paper on Bayes factors which, among many other things, draws out the relationship between Bayes factors and information criteria. Carlin and Louis (2000) include much relevant discussion, including a newly developed information criterion called the *Deviance Information Criterion*, which is designed to work well when models involve latent data and hierarchical priors. In this book, we note only that a quick and dirty method of model selection is to choose the model with the highest value for an information criterion. In the following empirical illustration, we investigate how effective this strategy is in selecting the number of components in a Normal mixture.

10.3.6 Empirical Illustration: A Mixture of Normals Model

We illustrate the mixtures of Normals model using two artificial data sets. To focus on the mixtures of Normals aspect of the model, we do not include any explanatory variables (i.e. β does not enter the model). The two data sets are, thus, all generated from

$$y_i = \varepsilon_i$$

where ε_i takes the mixtures of Normals form of (10.38). All three data sets have $N = 200$. The data sets are given by:

1. Data Set 1 has $m = 2$. The first Normal has $\alpha_1 = -1, h_1 = 16$ and $p_1 = 0.75$. The second Normal has $\alpha_2 = 1, h_2 = 4$ and $p_2 = 0.25$.
2. Data Set 1 has $m = 3$. The first Normal has $\alpha_1 = -1, h_1 = 4$ and $p_1 = 0.25$. The second Normal has $\alpha_2 = 0, h_2 = 16$ and $p_2 = 0.5$. The third Normal has $\alpha_3 = 1, h_3 = 16$ and $p_3 = 0.25$.

Histograms of these data sets are given in Figures 10.3a and b. These figures are included to show just how flexible mixtures of Normals can be. By mixing just two or three Normals together, we can get distributions which are very non-Normal. Mixtures of Normals can be used to model skewed, fat-tailed or multi-modal distributions.

We use a prior which is proper but very near to being noninformative. In particular, using the prior in (10.45), (10.46) and (10.47) we set $\underline{\alpha} = 0_m, \underline{V}_\alpha = (10\,000^2)I_m, \underline{s}_j^{-2} = 1, \underline{\nu}_j = 0.01$ and $\underline{p} = \iota_m$, where ι_m is an m-vector of ones. Bayesian inference is carried out using the Gibbs sampler involving (10.49), (10.50), (10.51), (10.52) and (10.53). For each data set, Bayesian inference is

---

[8] A consistent model selection criterion is one which chooses the correct model with probability one as sample size goes to infinity.

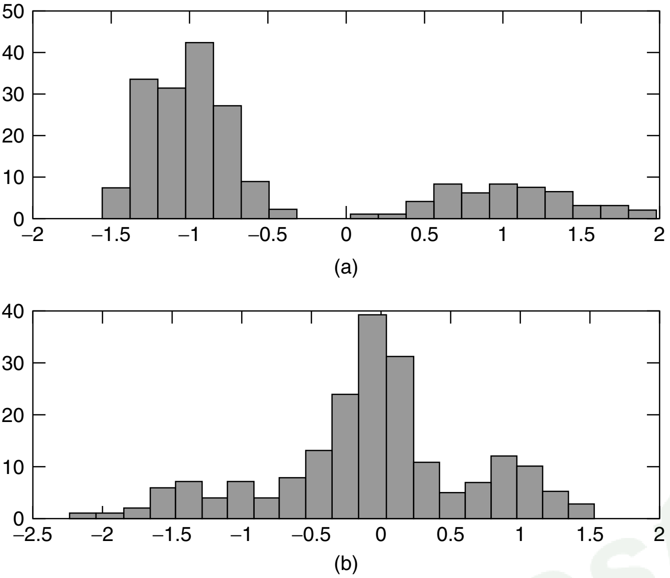

Figure 10.3 Histogram of (a) Data Set 1, (b) Data Set 2

done using $m = 1, 2$ and 3. The information criteria are evaluated at the posterior mean of the parameters in the model. The Gibbs sampler was run for 11 000 replications, with 1000 burn-in replications discarded and 10 000 replications retained. MCMC diagnostics indicate that this is an adequate number of replications to ensure convergence of the Gibbs sampler.

Tables 10.1 and 10.2 contain information criteria for Data Sets 1 and 2, respectively. The information criteria are consistent with one another and conclusive. For Data Set 1 (which was generated with $m = 2$), all of the information criteria select $m = 2$ as the preferred model. For Data Set 2, the information criteria all select the correct value of $m = 3$. Thus, at least for these data sets, information criteria do seem to be useful for selecting the number of components in a Normal mixture.

Table 10.3 presents posterior means and standard deviations of all parameters for the selected model for each data set. A comparison of posterior means with the values used to generate the data sets indicate that we are obtaining very reliable

Table 10.1 Information Criteria for Data Set 1

Model	AIC	BIC	HQ
$m = 1$	−174.08	−183.98	−183.09
$m = 2$	92.41	77.62	74.40
$m = 3$	−52.24	−81.92	−79.25

Table 10.2 Information Criteria for Data Set 2

Model	AIC	BIC	HQ
$m = 1$	-120.99	-130.88	-130.00
$m = 2$	-103.72	-123.51	-121.23
$m = 3$	-76.77	-106.35	-103.69

Table 10.3 Posterior Results for Two Data Sets

	Data Set 1		Data Set 2	
	Mean	St. Dev.	Mean	St. Dev.
α_1	-1.01	0.02	-0.86	0.22
α_2	1.02	0.06	-0.04	0.04
α_3	—	—	1.02	0.04
h_1	18.43	2.14	3.38	1.40
h_2	5.65	1.19	21.97	8.29
h_3	—	—	17.80	5.19
p_1	0.76	0.03	0.33	0.09
p_2	0.24	0.03	0.41	0.08
p_{3-}	—	—	0.25	0.04

estimates of all parameters. An examination of posterior standard deviations indicates that the parameters are reasonably precisely estimated, despite having only a moderately large sample size.

10.4 EXTENSIONS AND ALTERNATIVE APPROACHES

As we have stressed throughout, virtually any model in this book can be used as a component of a larger model. In many cases, posterior simulation for the larger model can be done using a Gibbs sampler, where one or more blocks of the Gibbs sampler can be lifted directly from the simpler model discussed in this book. We have shown how such a strategy can be used to develop posterior simulators for semiparametric probit and tobit models. A myriad of other such extensions are also possible. Similar extensions exist for the mixtures of Normals linear regression model considered above. There are many obvious extensions of models from previous chapters (e.g. mixtures of Normals nonlinear regression or any of the panel data models can be extended to have mixtures of Normals errors). Mixtures of Normals can also be used to make hierarchical priors more flexible. The possibilities are virtually limitless.

Bayesian nonparametrics is, at present, a very active research area and there are many approaches we have not discussed (e.g. Dirichlet process priors, wavelets, splines, etc.). Dey, Muller and Sinha (1998), *Practical Nonparametric and Semiparametric Bayesian Statistics*, provides an introduction to many such approaches in this rapidly developing field.

10.5 SUMMARY

In this chapter, Bayesian inference in several flexible models has been discussed. These models are designed to achieve similar goals as the non- or semiparametric models so popular in the non-Bayesian literature. There are numerous Bayesian nonparametric methods, but in this chapter we focus on simple models which are straightforward extensions of models discussed in previous chapters. The chapter is divided into sections containing models involving nonparametric regression (i.e. the regression line has an unknown functional form) and models involving a flexible error distribution.

The first model considered was the partial linear model. This is a regression model where some explanatory variables enter in a linear fashion and another one enters in a nonparametric fashion. We showed how this model can be put in the form of a Normal linear regression model with natural conjugate prior and, thus, analytical results from Chapter 3 apply directly. We showed how the partial linear model could be used as a component of a more complicated model (e.g. a semiparametric probit or tobit model), and how a Gibbs sampler could be constructed in a straightforward manner. We next considered the partial linear model where $p > 1$ explanatory variables were treated in a nonparametric fashion. Although such a model can be analyzed by ordering the data using a distance function, such an approach may not work well if p is more than 2 or 3. Hence, an additive version of the partial linear model was discussed. This model can also be put in the form of a Normal linear regression model with natural conjugate prior and, thus, a posterior simulator is not required. For modeling flexible error distributions, mixtures of Normals are powerful tools. We showed how a Gibbs sampler with data augmentation can be used to carry out posterior inference in the linear regression model with mixture of Normals errors.

No new tools for Bayesian computation were developed in this chapter. Bayesian quantities such as posterior and predictive means, Bayes factors, etc., can all be calculated using methods described in previous chapters. The only new tool introduced is a model selection technique involving information criteria. Information criteria do not have a rigorous Bayesian interpretation (other than as an approximation). However, they are typically very easy to calculate (and do not depend upon prior information) and, thus, are popular in practice. In this chapter, we showed how information criteria can be used to select the number of components in a Normal mixture.

10.6 EXERCISES

The exercises in this chapter are closer to being small projects than standard textbook questions. Remember that some data sets and MATLAB programs are available on the website associated with this book. The house price data set is available on this website, or in the *Journal of Applied Econometrics* Data Archive listed under Anglin and Gencay (1996) (http://qed.econ.queensu.ca/jae/1996-v11.6/anglin-gencay/).

1. The empirical illustration in Chapter 3 (Section 3.9) used the Normal linear regression model with natural conjugate prior with the house price data set. Please refer to this illustration for data definitions.
   (a) Use the house price data set and the partial linear model to investigate whether the effect of lot size on house price is nonlinear. Experiment with different priors (including empirical Bayes methods).
   (b) Calculate the Bayes factor comparing the partial linear model to the Normal linear regression model for this data set.
   (c) Carry out a prior sensitivity analysis to investigate how robust your answer in part (b) is to prior choice.
2. The additive version of the partial linear model may be too restrictive in some applications. This motivates interest in specifications which are more general, but less likely to suffer from the curse of dimensionality than the partial linear model. One such specification includes interaction terms between explanatory variables. In the case where $p = 2$, we would have the model

$$y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i1}x_{i2}) + \varepsilon_i$$

   (a) Describe how the methods of Section 10.2.3 can be extended to allow for Bayesian analysis of this model.
   (b) Using artificial data sets of your choice, investigate the empirical performance of the methods developed in part (a).
3. Chapter 9 included empirical illustrations of the tobit and probit models (see Sections 9.3.1 and 9.4.1, respectively).
   (a) Re-do these empirical illustrations using semiparametric tobit and probit as described in the present chapter. Note: the empirical illustrations describe the artificial data sets used (see also the website associated with this book).
   (b) Describe how Bayesian inference in a semiparametric ordered probit model could be developed.
   (c) Write a program which carries out Bayesian inference in the semiparametric ordered probit model and investigate its performance using artificial data.
4. The empirical illustration of the mixtures of Normals model (Section 10.3.6) used information criteria to select the number of elements in the mixture.
   (a) Write a program (or modify the one on the website associated with this book) which uses marginal likelihoods to choose the number of elements in the mixture.
   (b) Using an informative prior, investigate the performance of your program using the artificial data sets described in Section 10.3.6.
   (c) Repeat part (b) using different data sets, and compare results obtained using marginal likelihoods with those obtained using information criteria.