Bayesian Variants of Some Classical Semiparametric Regression Techniques

Gary Koop
Department of Economics
University of Glasgow
G.Koop@socsci.gla.ac.uk

and

Dale Poirier
Department of Economics
University of California at Irvine
dpoirier@uci.edu

April 2001

ABSTRACT

This paper develops new Bayesian methods for semiparametric inference in the partial linear Normal regression model: $y = z\beta + f(x) + \varepsilon$ where $f(.)$ is an unknown function. These methods draw solely on the Normal linear regression model with natural conjugate prior. Hence, analytical finite sample results are available which do not suffer from problems of theoretical and computational complexity which plague the existing literature. Constrained and unconstrained estimation are considered as is testing of parametric regression models against semiparametric alternatives and prediction. We discuss how these methods can, at some cost in terms of computational complexity, be extended to other models (e.g. qualitative choice models or those involving censoring or truncation) and provide precise details for semiparametric probit and tobit models. We show how the assumption of Normal errors can easily be relaxed. Our methods are illustrated using artificial and real data sets.

Acknowledgements: We would like to thank Andy Chesher, D. M. Titterington, Adonis Yatchew and seminar participants at the University of Glasgow and the EC^2 2000 conference for helpful comments.

1 Introduction

In a recent survey paper, Yatchew (1998), argues persuasively that economic theory rarely provides a specific functional form for the regression relationship between a dependent variable, y and explanatory variables x and z. Accordingly, it is desirable to to work with the nonparametric regression model $y = f(x, z) + \varepsilon$, where $f(.)$ is an unknown function which is approximated nonparametrically. Unfortunately, if there are many explanatory variables, then nonparametric methods suffer from the curse of dimensionality. That is, nonparametric methods average over "nearby" observations in order to approximate the regression relationship. For a fixed sample size, as the dimension of x and/or z increases, "nearby" observations become farther and farther away and nonparametric methods become more and more unreliable. Accordingly, there has been a recent focus on a semiparametric regression model, the so-called "partial linear model" (see Yatchew, 1998), which is of the form $y = z\beta + f(x) + \varepsilon$ where x is of low dimension. Yatchew (1998) cites a plethora of recent econometric work which derives many key theoretical results for the partial linear model. However, he bemoans the fact that nonparametric regression methods are rarely used in applied work. He highlights three reasons for this:

> "First, nonparametric regression techniques are theoretically more complex than the usual tool kit of linear and nonlinear parametric modelling methods.
>
> Second, nonparametric regression techniques are computationally intensive....
>
> Third, a unified framework for constrained estimation and testing of economic models using nonparametric regression is still in the incipient stage." [Yatchew (1998, pages 672-673)].

The purpose of the present paper is to develop a Bayesian method which surmounts all of these problems. First, the Bayesian semiparametric methods developed here are simple since they are based on the standard Normal linear regression model with natural conjugate prior. This is usually the first model taught to graduate (or advanced undergraduate) students learning Bayesian econometrics and is included in most standard textbooks (e.g. Poirier, 1995, Judge et al., 1985, Greene, 2000). We stress that nothing in the methods we develop for the partial linear model goes beyond the basic results for this simple and commonly-used model. Second, our methods are computationally simple as standard analytical results for the Normal linear regression model with natural conjugate prior are available. Third, Bayesian methods provide for a unified theory of estimation and model comparison. For instance, Bayes factors comparing a semiparametric to a parametric model or even two different semiparametric models can be calculated in a simple way.

Constrained estimation can easily be handled by allocating zero prior weight to regions of the parameter space that violate the constraints (see Geweke, 1986).

It is also worth noting that Bayesian methods provide exact finite sample results for a class of models where classic asymptotic theory can be complicated. Furthermore, Bayesian methods are occasionally criticized on the grounds that they incorporate prior information. In the approach developed here, the only type of prior input required is one prior hyperparameter, η, which controls the degree of smoothness of $f(.)$. We discuss various methods for eliciting η.

The approach developed here differs from the existing Bayesian nonparametric regression literature. This literature includes nonparametric models based on Dirichlet process priors, wavelets, splines and Gaussian processes (see, for instance, Dey, Muller and Sinha, 1998, O'Hagan, 1978 and Silverman, 1985). Below we will compare our work with other Bayesian approaches in greater detail. Although these other Bayesian methods have many advantages, we would argue that our approach has the advantage of theoretical and computational simplicity. The simplicity of our theoretical results allows us to look inside the Bayesian black box and understand what precisely our methods are doing (e.g. we show that the Bayesian posterior means of nonparametric effects exhibit a local averaging property analogous to kernel smoothing approaches). On the computational front, alternative Bayesian approaches require difficult Markov Chain Monte Carlo algorithms while the approach we advocate here for the partial linear Normal model does not.

On a more general level, nonparametric methods involve imposing structure on what can become an infinite dimensional parameter space. Imposing such structure involves making assumptions. Classical nonparametric approaches typically impose structure by making assumptions on the degree of smoothness of $f(.)$. For instance, kernel smoothing methods involve a choice of functional form (i.e. the kernel) and a bandwidth choice. The former of these controls the shape of the weights in the local averaging, the latter controls the magnitudes of the weights. These assumptions are analogous to the assumptions made in our prior. Classical nonparametric econometricians often stress that the implications of economic theory are almost invariably nonparametric and do not specify parametric functional forms (see, e.g., Yatchew, 1998, pages 669-670). Hence, it is more natural to place structure on a model through nonparametric assumptions. In the Bayesian approach developed in this paper, the prior assumptions we make do take such a nonparametric form (in a sense which will be made clear later). That is, our method of imposing structure on an infinite dimensional parameter space involves a prior over the degree of smoothness of the nonparametric regression function. It is not a prior reflecting a strength of belief in a particular parametric form.

Before we begin, a digression on terminology is warranted. Since the partial linear model involves a parametric component, $z\beta$, and a nonparametric one, $f(x)$, we follow a common classical practice and refer to this as a "semiparametric" regression model (see, e.g., Powell, 1994), with $y = f(x) + \varepsilon$ being a "nonparametric" regression model. However, the Bayesian models developed in this paper, in finite samples, all involve a finite dimensional parameter space and, in this sense, are not "semiparametric". Nevertheless, the model we work with is the same as one labelled "semiparametric" in the classical literature. Hence, we adopt the standard classical language of "non-" and "semiparametric" even though in a formal sense it is not precisely correct to do so. This will serve to simplify terminology and highlight the similarities (and differences) between Bayesian and classical methods of inference in this class of models. Useful surveys of classical non- and semiparametric regression techniques are available in many places, including Hardle (1990), Hardle and Linton (1994), Horowitz (1998), Pagan and Ullah (1999) and Yatchew (1998).

The paper is organized as follows. The second section introduces the partial linear model and discusses methods for estimation, testing and prediction, beginning with the case where x is a scalar, before proceeding to the case where x is a vector. The third section discusses how our basic methods for inference in the partial linear can be extended to many more complicated models. We consider semiparametric probit and tobit models to illustrate how such extensions can be implemented in practice. All these sections include empirical illustrations. The fifth section concludes. Most theoretical derivations and proofs are in the Appendix.

2 The Model

Consider the partial linear model

$$y_i = z_i\beta + f(x_i) + \varepsilon_i, \tag{2.1}$$

where y_i is the dependent variable, z_i is a vector of k explanatory variables which are treated parametrically, x_i is a scalar explanatory variable which is treated nonparametrically, $f(.)$ is an unknown function and ε_i is i.i.d. N(0,σ^2) for $i = 1, ..., N$. All explanatory variables are assumed to be fixed (or exogenous). Observations are ordered so that $x_1 \leq x_2 \leq ... \leq x_N$. Define $y = (y_1, ..., y_N)'$, $Z = (z_1, ..., z_N)'$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)'$. Letting $\gamma = (f(x_1), ..., f(x_N))'$, $W = (Z, I_N)$ and $\delta = (\beta', \gamma')'$, we can write (2.1) as:

$$y = W\delta + \varepsilon. \tag{2.2}$$

There are two important features of regression (2.2). Firstly, (2.2) is characterized by *insufficient observations* because there are more unknown elements in δ than there are observations, i.e. $N + k \geq N$ =rank(W). Therefore, δ and σ^{-2} are not identified; rather $\varphi \equiv W\delta$ is identified. This manifests itself in the singularity of $W'W$. It also implies (see Poirier, 1998) that the prior distribution of δ given φ in not affected by the data, i.e. $p(\delta|y, \psi) = p(\delta|\varphi)$, or equivalently, $\delta \perp y|\varphi$. Secondly, a perfect fit is available such that the sum of squared errors is zero. As a result, the likelihood is unbounded along $\delta = d$ as $\sigma^2 \to 0$, where d is any $N \times 1$ solution to the normal equations $(W'W)d = W'y$. This causes problems for maximum likelihood estimation since the MLE of σ^2 is zero regardless of the data. For a Bayesian there are concerns because an unbounded likelihood function may lead to an unbounded posterior, even with a proper prior. However, for the judiciously chosen improper partially informative priors used in the paper, the posterior is proper.

In the sub-sections which follow, we will consider informative priors which are analogous to classical assumptions about the degree of smoothness of the function $f(.)$. Each of these priors will depend on a scalar hyperparameter. We discuss how one might wish to select this hyperparameter optimally and relate this choice to the classical issue of optimal bandwidth selection.

2.1 Priors Based on Smoothness

Following Rice (1984) and in a spirit going back to Whittaker (1923), Yatchew (1998) surveys several different classical approaches to estimating the partial linear model given in (2.1). One approach, referred to as the Differencing Estimator, uses differencing to remove the nonparametric effect. That is, if the first derivative of $f(.)$ is bounded by a constant C, we can write:

$$|f(x_i) - f(x_{i-1})| \leq C|x_i - x_{i-1}|. \tag{2.3}$$

If x_i is drawn from a distribution with finite support, then as sample size increases x_i and x_{i-1} get closer and closer together and the right hand side of (2.3) goes to zero. Hence, if these conditions (i.e. smoothness of $f(.)$ and finite support of x) are satisfied, ordinary least squares estimation of Δy on ΔZ will yield a consistent estimate of β. The priors discussed in this subsection reflect the properties of the classical Differencing Estimator. They embed the idea that $f(x_i) - f(x_{i-1}) = \gamma_i - \gamma_{i-1}$ should be small. One variant incorporates the idea that the posterior mean of β should be equal to the OLS estimator on differenced data. This intuition extends to any degree of differencing.

We assume a natural conjugate Normal-Gamma prior for β, γ and σ^{-2} (see Poirier, 1995, pages 526-527). By adopting such a choice, we are able to obtain simple analytical results which provide a great deal of

intuitive insight. Unlike other Bayesian nonparametric approaches, they do not require computationally-intensive posterior simulation methods (e.g. Markov Chain Monte Carlo methods). In order to focus on the nonparametric part of the partial linear model, we assume the standard noninformative prior for σ^{-2} and β:[1]

$$p(\beta, \sigma^{-2}) \propto \sigma^2. \tag{2.4}$$

For the coefficients in the nonparametric part of the model, we use the partially informative prior (see Poirier, 1995, page 535) on the m^{th} differences of γ:

$$R\delta \sim N(0_{N-m}, \sigma^2 V(\eta)), \tag{2.5}$$

where $V(\eta)$ is a positive definite matrix which depends on a hyperparameter η, 0_{N-m} is the $N-m$ vector of zeros and $R = [0_{(N-m)\times K}, D]$, where D is the $(N-m)\times N$ first-differencing matrix given by:

$$D = \begin{bmatrix} d_0 & d_1 & ... & d_m & 0 & ... & 0 \\ 0 & d_0 & d_1 & ... & d_m & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & 0 & d_0 & ... & d_m \end{bmatrix}, . \tag{2.6}$$

where $d_0, ..., d_m$ are the appropriate differencing weights. For instance, with first differencing $d_0 = -1$, $d_1 = 1$, with second differencing $d_0 = 1$, $d_1 = -2$, $d_2 = 1$, etc. Note that this structure implies that we only have prior information about $\Delta^m \gamma_i$; that is, on the m^{th} difference of $f(.)$. All of our empirical illustrations below set $m = 1$.

As outlined in the Appendix, the resulting posterior is Normal-Gamma. Using the standard notation for this distribution (e.g. Poirier, 1995, page 128), we have $\delta, \sigma^{-2}|y \sim \text{NG}(\widetilde{\delta}, \widetilde{Q}, \widetilde{s}^{-2}, \widetilde{\nu})$, where $\widetilde{\delta}, \widetilde{Q}, \widetilde{s}^{-2}, \widetilde{\nu}$ are defined in equations (A.11), (A.12), (A.10) and (A.9), respectively. As shown in the Appendix, provided $V(\eta)$ is positive definite, all quantities in these equations exist and, hence, the posterior is a valid probability density function, despite the fact that the number of regressors in (2.1) is greater than the number of observations. Of particular importance are the marginal posterior mean and variance of δ, which are given by

[1] Prior information on σ^{-2} and β can be trivially incorporated into the formulae below, provided it is of natural conjugate form. Note also that, if one were to assume prior independence between σ^{-2} and δ, then the results of Fernandez, Osiewalski and Steel (1997) imply that an improper prior for the precision proportional to $\left(\sigma^{-2}\right)^{-\frac{m-1}{2}}$ for $m \geq 0$ leads to a posterior which is improper, even if the prior for δ is proper. Our prior given in (2.4) does not incorporate such prior indepenence and can be rationalized as a limiting form of the conjugate prior below in which σ^{-2} and δ are dependent.

$$E(\delta|y,\eta) \equiv \widetilde{\delta} = \widetilde{Q}W'y, \tag{2.7}$$

where

$$\widetilde{Q} = \left(R'V(\eta)^{-1}R + W'W\right)^{-1}, \tag{2.8}$$

and

$$var(\delta|y,\eta) = \frac{\widetilde{\nu}\widetilde{s}^2}{\widetilde{\nu}-2}\widetilde{Q}. \tag{2.9}$$

The posterior mean exists if $\widetilde{\nu} > 1$ and posterior variance if $\widetilde{\nu} > 2$. Since $\widetilde{\nu} = N$ existence of these moments is assured in reasonable data sets. The precise formula of the posterior mean of σ^{-2} (i.e. $\widetilde{s}^{-2}$) is given in the Appendix.

The general formula for the posterior mean of γ can be rewritten in an instructive manner (see A.13):

$$E(\gamma|y,\eta) = \left[M_Z + D'V(\eta)^{-1}D\right]^{-1} M_Z y, \tag{2.10}$$

where $M_Z = I_N - Z(Z'Z)^{-1}Z'$ is the familiar OLS idempotent matrix and, thus, $M_Z y$ are the OLS residuals from the regression of y on Z. Note also that, if the prior in (2.5) becomes noninformative (i.e. $V(\eta)^{-1} \rightarrow 0_{N-m,N-m}$), then $E(\gamma|y,\eta) = y$ and the nonparametric part of the model merely fits the observed data points (i.e. there is no smoothing). In the subsequent material, we will refer to (2.10) as the fitted nonparametric regression line.

The posterior mean of β has the form (see A.14):

$$E(\beta|y,\eta) = \left\{Z'\left(D'\left[DD' + V(\eta)\right]^{-1}D\right)Z\right\}^{-1} Z'\left(D'\left[DD' + V(\eta)\right]^{-1}D\right)y. \tag{2.11}$$

So far, we have said nothing about $V(\eta)$. A simple form for $V(\eta)$, reflecting only smoothness considerations (i.e. $\Delta^m\gamma_i$ is small), would be to take $V(\eta) = \eta I_{N-m}$. This prior depends only on the scalar hyperparameter η, which can be selected by the researcher to control the degree of smoothness. We refer to this prior as our **Base Prior**. This choice is analogous to that used by Shiller (1973) in his analysis of distributed lag models.

Additional structure on $V(\eta)$ can be obtained if we impose the requirement that the posterior mean for β is equal to the OLS estimate from the regression of $\Delta^m y$ on $\Delta^m Z$. From (2.11) it can be seen that this occurs if

$$V(\eta) = \eta I_{N-m} - DD',$$

where $\eta \geq 4^m$ is sufficient to ensure the positive definiteness of $V(\eta)$. This form for $V(\eta)$ completes the specification of what we shall call the **mth Difference Prior**. It requires only prior elicitation of the scalar hyperparameter η, which controls the degree of smoothness imposed on $f(.)$.

The Base and mth Difference priors have many desirable properties and are similar in spirit to classical nonparametric approaches. Note, however, that posterior inference using this prior only uses order information on x (i.e. the only place x enters the model is through the data being ranked so that $x_1 \leq x_2 \leq ... \leq x_N$.). With large samples and finite support for the distribution of x, no information is lost by doing so. However, in finite samples, information on the distance between x_i and x_{i-1} can potentially be quite important. As described above, the classical nonparametric literature often smooths $f(.)$ by bounding its first derivative as in (2.3). This consideration is embedded in the following prior. Since this prior depends on x, we refer to it as the **X-Prior**. A reasonable way to incorporate first difference information is to set

$$v_i = \eta \left(x_i - x_{i-1} \right), \tag{2.12}$$

where v_i is the $(i,i)^{th}$ element of $V(\eta)$. Since the data is ordered, all of the v_is are non-negative. A further advantage of this prior is that the dependency on x_i in (2.12) is of a form such that the marginal priors are not sensitive to the set of observations in the model. For instance, if a new data point, x^*, is added, we would not like this to affect the prior we previously held for γ_i for $i = 1, .., N$. The prior variance structure in (2.12) has the property that $p(\gamma_i)$ is the same regardless of whether x^* is included or not. This property would not hold if $m > 1$ and, hence, we use this prior only for the $m = 1$ case.

To provide more intuition on how the Bayesian posterior involves an averaging of nearby observations, consider (2.10) for the case with no parametric regressors (i.e., $\beta = 0$) and $m = 1$. In this case, $E(\gamma|y, \eta) = \left[I + D'V(\eta)^{-1}D\right]^{-1} y$ and the matrix $\left[I + D'V(\eta)^{-1}D\right]^{-1}$ contains the weights used in local averaging for all of the observations. To illustrate the manner in which the fitted nonparametric regression line involves local averaging of observations, it would be instructive to write out $E(\gamma_i|y, \eta)$. Unfortunately, a closed form expression for this does not exist except for specific choices of $V(\eta)$ due to the difficulty of inverting $I + D'V(\eta)^{-1}D$. For the First Difference Prior, this matrix can be inverted and (2.10) can be written as:

$$E(\gamma_i|y, \eta) = \frac{y_{i-1}}{\eta} + \left(1 - \frac{2}{\eta}\right) y_i + \frac{y_{i+1}}{\eta},$$

for $i = 2, .., N-1$.[2] Hence, the smoothing implied by the First Difference Prior involves only y_i and its two nearest neighbors: y_{i-1} and y_{i+1}. Furthermore, $V(\eta)$ is only positive definite for all N if $\eta \geq 4$. These two facts imply that this prior can never smooth the fitted nonparametric regression line very much. That is, the smoothest it can ever be is $E(\gamma_i|y,\eta) = .25y_{i-1} + .5y_i + .25y_{i+1}$. In reasonable data sets, this will yield a very irregular fitted nonparametric regression line (see the empirical illustrations below). Thus, the prior which is most closely related to the classical Differencing Estimator does not seem to allow for an adequate degree of smoothing of the nonparametric regression line.

For the Base Prior it is instructive to look at $E(\gamma_i|y,\gamma^{(i)},\eta)$ where $\gamma^{(i)} = (\gamma_1, ..., \gamma_{i-1}, \gamma_{i+1}, ..., \gamma_N)$. For this prior, with $\beta = 0$, it can be shown that:

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

for $i = 2, .., N-1$. $E(\gamma_i|y,\gamma^{(i)},\eta)$ 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}). This implies that, in contrast to the First Difference Prior, $E(\gamma_i|y,\eta)$ will be a weighted average of all of $y_1, .., y_N$. Since η controls the degree of smoothness we wish to impose on $f(.)$, it makes sense that as $\eta \to \infty$ we obtain $E(\gamma_i|y,\gamma^{(i)},\eta) = y_i$ (i.e. no smoothing whatsoever). As $\eta \to 0$ we obtain $E(\gamma_i|y,\gamma^{(i)},\eta) = \frac{1}{2}\left(\gamma_{i-1} + \gamma_{i+1}\right)$. Furthermore, it can be shown that $var(\gamma_i|y,\gamma^{(i)},\eta) = \frac{\sigma^2\eta}{2+\eta}$ which goes to zero as $\eta \to 0$. Thus the limiting case of perfect smoothness yields $\gamma_i = \frac{1}{2}\left(\gamma_{i-1} + \gamma_{i+1}\right)$ and the nonparametric regression component is merely a straight line.

For the Base Prior, Table 1 contains the weights for the 6^{th} observation for the case where $N = 11$ and $V(\eta) = \eta I$ for various choices of η.

Table 1: Local Averaging Implied by Base Prior for 6th observation when N=11

Observation	η=.001	η=.01	η=.1	η=1	η=10
1	.090	.087	.058	.005	.000
2	.091	.087	.064	.010	.000
3	.091	.089	.076	.025	.000
4	.091	.092	.095	.065	.006
5	.091	.095	.125	.171	.071
6	.092	.100	.166	.447	.845
7	.091	.095	.125	.171	.071
8	.091	.092	.095	.065	.006
9	.091	.089	.076	.025	.000
10	.091	.087	.064	.010	.000
11	.090	.087	.058	.005	.000

[2] Results for $i = 1$ and N are slightly different due to endpoint effects.

This table shows, for instance, that setting $\eta = 1$ implies that $E(\gamma_6|y,\eta)$ allocates 44.7% of the weight to y_6, 17.1% to each of the nearest neighbors, y_5 and y_7, 6.5% to the second nearest neighbors y_4 and y_8, etc. The weights are symmetrical about $i = 6$. Small values of η imply weights which die off very slowly away from $i = 6$, while large values allocate most of the weight at or near $i = 6$. This table, of course, is only valid for N=11. However, very similar weights hold for different values of N. For instance, for $\eta = .1$ and $N = 11$ we approximately obtain

$$E(\gamma_i|y,\eta) = ... + .125y_{i-1} + .166y_i + .125y_{i+1} + ...$$

for interior values of i (i.e. values not near 1 or N). Using $N = 111$ the comparable result is:

$$E(\gamma_i|y,\eta) = ... + .114y_{i-1} + .156y_i + .114y_{i+1} + ...$$

Hence, the results of Table 1 are suggestive and of potential use for prior elicitation.

The X-Prior implies a posterior which averages in a similar manner to the Base Prior. For this reason, and since posterior properties depend on the specific values of $x_1, .., x_N$, we do not discuss the X-Prior further at this stage.

2.2 Comparison with Related Bayesian Approaches

Many Bayesian approaches (e.g. those involving Dirichlet process priors) require computationally-intensive posterior simulation. A chief advantage of our approach is that analytical results are available for the semiparametric regression model given in (2.1). Many extensions of the semiparametric regression model (e.g. semiparametric tobit or probit), however, will require posterior simulation. It is in these contexts that the presence of analytical results for the regression part of the model will be of enormous use, since posterior simulation would then be required for only for the model extension. For instance, in the semiparametric probit model discussed below, the probit structure implies that a Gibbs sampler is required. If, in addition, posterior simulation methods are required for the semiparametric regression part of the model, computational demands can become quite high. If the researcher is interested in estimating many models, or if cross-validation is done, it is easy to imagine cases where computational costs become prohibitive. Hence, we would argue that our derivation of analytical results for the partial linear model is of great importance.

There are other Bayesian approaches which allow for analytical results in the nonparametric regression model. O'Hagan (1978) considers a similar specification with a prior covariance matrix which implies prior correlations between points on the nonparametric regression line that depend on the distance they are from

one another (i.e. prior correlations depend on $|x_i - x_j|$ for $i \neq j$. The smoothing spline literature (e.g. Wahba, 1983, Silverman, 1985 and Green and Silverman, 1994) is based on minimizing a linear combination of weighted fit (assuming Normality) plus smoothness. This formulation is equivalent to a Bayesian analysis with an implied prior for γ given σ that is a singular multivariate Normal with rank N-2:

$$p(\gamma|\sigma) \propto \exp\left(-\frac{\eta}{\sigma^2}\int f^{''2}\right)$$

where $f^{''}$ is the second derivative of $f(.)$ and η is a smoothing parameter. Our approach can be trivially extended to allow for D to be the second difference matrix. If we had done so, we would have obtain a similar prior except with sums of second differences instead of integrals of second derivatives. Green and Silverman (1994) also show how the natural cubic spline prior can be written as

$$p(\gamma) \propto \exp\left(-\frac{\eta}{\sigma^2}\gamma' Q \gamma\right)$$

for a certain matrix Q which is rank deficient. Our prior is formulated in terms of $D\gamma$, but the implied prior covariance matrix for γ is also rank deficient. For D being the second difference matrix and a particular choice of $V(\eta)$, our approach could thus match the natural cubic spline approach. Hence, our model nests other specifications in the literature. Our approach, however, is more general, and it is motivated using the simple linear regression model with natural conjugate prior and

2.3 Prior Elicitation

The priors recommended in this paper require only the elicitation of the hyperparameter, η. The exact value for η will usually depend on the particular empirical application at hand. Here we offer a few thoughts on issues that might be useful in prior elicitation. A first point to stress is that, as in any Bayesian analysis, it is usually sensible to carry out a prior sensitivity analysis. In the present case, this will be particularly simple since only one hyperparameter is involved. A second point to stress is that applied classical nonparametric researchers often choose kernels and bandwidths experimentally, trying out different choices and examining graphs of the nonparametric regression line to see if it looks suitably smooth. A similar strategy can be employed by the Bayesian. Thirdly, since η is related to the marginal effect of x on y, it has a natural interpretation which is likely to make prior elicitation fairly simple. Alternatively, hierarchical priors for η can be used in a straightforward fashion. For instance, suppose the researcher is not willing to choose a particular value for η, but is willing to specify a prior, p(η). Note that

$$\mathrm{p}(\gamma|y) \propto \int \mathrm{p}(\gamma|y,\eta)\mathrm{p}(y|\eta)\mathrm{p}(\eta)\ \mathrm{d}\eta,$$

and the results of the previous subsection provide $\mathrm{p}(\gamma|y,\eta)$ and $\mathrm{p}(y|\eta)$ for particular values of η. If a suitable grid of values of η is chosen, the integral in the previous expression can be approximated by weighting results for each value of η by $\mathrm{p}(\eta)$.

The extreme bounds analysis literature (see Leamer, 1982, 1984) provides another approach which can be used by those who do not feel comfortable choosing a particular prior, but are comfortable specifying a range of priors. We investigate this approach below.

It is worth noting in passing that empirical Bayesian methods, which choose hyperparameter values which maximize the marginal likelihood are not suitable here since, as $\eta \to \infty$, the marginal likelihood becomes infinite (Iwata, 1996, Lemma 3, page 246). However, the extreme bounds analysis literature provides another approach which can be used by those who do not feel comfortable eliciting a particular value for η, but are comfortable specifying a range of values for this hyperparameter.

If the researcher does not wish to subjectively elicit even a range of values for η, cross-validation techniques, analogous to those used in classical nonparametrics for bandwidth selection (e.g. Yatchew, 1998), are an alternative. In the empirical parts of this paper, we use a simple cross-validation methodology similar to one outlined in Yatchew (1998), page 692. For simplicity, we illustrate the methodology for the nonparametric regression model. Define the cross-validation function as:

$$CV(\eta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - E(\gamma_i|y^{(i)},\eta))^2, \tag{2.14}$$

where $y^{(i)} = (y_1,..,y_{i-1},y_{i+1},...,y_N)'$.To make things even simpler, in our empirical work we use the following approximation $E(\gamma_1|y^{(1)},\eta) = E(\gamma_2|y^{(1)},\eta)$, $E(\gamma_N|y^{(N)},\eta) = E(\gamma_{N-1}|y^{(N)},\eta)$ and, for i=2,..,N-1,

$$E(\gamma_i|y^{(i)},\eta) = \frac{1}{2}\left[E(\gamma_{i-1}|y^{(i)},\eta) + E(\gamma_{i+1}|y^{(i)},\eta)\right].$$

η is chosen so as to minimize the cross-validation function.

2.3.1 Illustration: Estimation of Fitted Nonparametric Regression Line

To illustrate our methods, we use two different data sets, one an artificial data set, the second a house pricing data set from Yatchew (1998, page 699). For the artificial data set, we use the same data generating mechanism as Yatchew (1998, Figure 3). In particular, for $i = 1,..,100$ we generate

$$y_i = x_i \cos(4\pi x_i) + \varepsilon_i, \tag{2.15}$$

where ε_i is i.i.d. N(0,.09) and x_i is i.i.d. Uniform over [0,1]. Figures 1a-1c plot, for the three priors and $m = 1$, the data generated, the true regression line given by (2.15), the fitted nonparametric regression line given by (2.10) and this fitted line plus/minus two posterior standard deviations. The values for η chosen by cross-validation were .18 for the Base Prior, 4 for the First Difference Prior and 19.7 for the X-Prior.

Insert Figure 1 approximately here

For the reasons discussed above, the First Difference Prior is not able to adequately smooth the nonparametric regression line and this is clearly seen in Figure 1b. However, Figures 1a and 1c exhibit very appealing properties. Even for the relatively small sample size, the fitted nonparametric regression line tracks the true regression line quite well. Furthermore, with few exceptions, the true regression line lies within an approximate 95% highest probability density interval.

The house price data set is described in more detail in Yatchew (1998, Figure 11, page 699). It suffices here to note that y_i is the sales price of house i for $i = 1, .., 92$. The data set includes 10 explanatory variables, 8 of which are assumed to enter linearly (i.e. the dimension of z is 8). These are: lot size, square footage of house, number of bedrooms, average neighborhood income, distance to highway and dummies for the presence of a garage, a fireplace and luxury appointments. The other two explanatory variables, x_1 and x_2, are location coordinates of the house and it is assumed that these enter nonparametrically. In this section we use only x_1. Figures 2a-2c plot $y_i - z_i E(\beta|y)$ and $E(\gamma_i|y, \eta)$ for different priors with hyperparameter values chosen by cross-validation. For the Base, First Difference and X-Priors, η is .24, 4 and 2.35, respectively.

Insert Figure 2 approximately here

As with the artificial data, our Bayesian semiparametric methods seem to work well, with our approximate 95% highest probability density interval containing virtually every data point. However, the fitted nonparametric regression lines are not as smooth as they were for the artificial data. This kind of pattern emerges with our methods when, as in the house price data set, a large amount of random error is in the data and there appears to be no clear relationship between x and y.

For the sake of brevity, we do not present posterior results for β and σ^2. Suffice it to note here, that these results are very reasonable and qualitatively similar to those given in Yatchew (1998, Figure 11).

2.4 Extreme Bounds Analysis

For researchers who are reluctant to present results based only a single, or a few, priors, the extreme bounds analysis literature provides useful guidance (see Leamer, 1982, 1984). It derives regions, which depend on the degree of information assumed about the prior, within which the posterior mean of the regression coefficients must lie. Many of the results in this literature can be applied directly to the present context. We illustrate how this can be done by adapting Theorem 7 of Leamer (1982) to the nonparametric regression model.

Thereom: Consider the Normal linear regression model given in (2.2) without the parametric component for simplicity (i.e. $W = I_N$) and natural conjugate prior given by $p(\beta, \sigma^{-2}) \propto \sigma^2$ and $R\delta \sim N(0_{N-m}, \sigma^2 V)$. Suppose $\underline{V} \leq V \leq \overline{V}$ is the sense that $V - \underline{V}$ and $\overline{V} - V$ are positive semidefinite. Let ϕ be an N-vector of constants. Then the posterior mean of $\xi = \phi'\delta$ satisfies

$$\xi_{\min} \leq \xi \leq \xi_{\max},$$

where

$$\xi_{\min} = \xi_1 + \widetilde{\phi}'\xi_2 - \sqrt{\xi_3 \widetilde{\phi}' \Phi^{-1} \widetilde{\phi}}$$

and

$$\xi_{\max} = \xi_1 + \widetilde{\phi}'\xi_2 + \sqrt{\xi_3 \widetilde{\phi}' \Phi^{-1} \widetilde{\phi}}.$$

In the previous formulae, we use the following notation:

$$\widetilde{\phi}' = \phi' R' \left(RR'\right)^{-1},$$

$$\xi_1 = \phi' y - \phi' R' \left(RR'\right)^{-1} Ry,$$

$$\xi_2 = \left(\left(RR'\right)^{-1} + \underline{V}^{-1}\right)^{-1} \left[\left(RR'\right)^{-1} Ry + \frac{1}{2}\left(\underline{V}^{-1} - \overline{V}^{-1}\right)\left(\left(RR'\right)^{-1} + \overline{V}^{-1}\right)^{-1} \left(RR'\right)^{-1} Ry\right],$$

$$\xi_3 = \frac{1}{4} y' R' \left(RR'\right)^{-1} \left(\left(RR'\right)^{-1} + \overline{V}^{-1}\right)^{-1} \left(\underline{V}^{-1} - \overline{V}^{-1}\right)\left(\left(RR'\right)^{-1} + \underline{V}^{-1}\right)^{-1\prime} \left(RR'\right)^{-1} Ry$$

and

$$\Phi = \left((RR')^{-1} + \overline{V}^{-1}\right)\left(\underline{V}^{-1} - \overline{V}^{-1}\right)^{-1}\left((RR')^{-1} + \overline{V}^{-1}\right) + \left((RR')^{-1} + \overline{V}^{-1}\right).$$

In other words, insofar as one feels comfortable bounding the prior covariance matrix, one can find bounds within which the posterior mean of any function of the regression coefficients must lie. We stress that the derivations in this section hold for arbitrary positive definite forms for V, not just the three forms for $V(\eta)$ specified above.

The results of the previous theorem do not immediately extend to the partial linear model since, in that case, $W'W$ is singular and Leamer (1982)'s proof requires nonsingularity of this matrix. However, Chamberlain and Leamer (1976, Theorem 9, pages 82-83), addresses a case involving singular $W'W$ which can easily be adapted to the develop extreme bounds for the partial linear model.

2.4.1 Illustration: Extreme Bounds Analysis

We carry out an extreme bounds analysis for each element of γ (i.e. to obtain bounds for γ_i, we set ϕ to a vector of zeros with the exception that the i^{th} element is 1) using the artificial data set described in (2.15) and focus on the first difference case (i.e. $m = 1$). The results in Figure 3a (labelled "Reasonable Variance Bounds") are for $\underline{V} = .2I_{N-1}$ and $\overline{V} = .5I_{N-1}$. An examination of Table 1 and a consideration of the fact that Vs within these bounds can have a wide variety (including non-diagonal) forms, indicates that we are capturing a broad range of patterns and degrees of nonparametric smoothing. Figure 3a plots the data, the true regression line and the extreme bounds for the fitted nonparametric regression line (i.e. $\xi_{\min}$ and $\xi_{\max}$ for each γ_i). Figure 3b (labelled "Wide Variance Bounds") is calculated using $\underline{V} = .01I_{N-1}$ and $\overline{V} = 10I_{N-1}$.

It can be in Figure 3a that using a reasonably wide set of bounds for the prior covariance matrix can still imply that the fitted nonparametric regression line must lie in a fairly narrow interval. However, Figure 3b sensibly indicates that some knowledge about the degree of smoothing in the nonparametric regression is necessary to obtain a meaningful estimate of the nonparametric regression line. As Table 1 indicates, the prior covariance bounds $\underline{V} = .01I_{N-1}$ and $\overline{V} = 10I_{N-1}$are so broad that virtually any sort of nonparametric regression line could be found which corresponds to some V in this interval.

Insert Figure 3 here

2.5 Imposing Constraints on the Nonparametric Regression Function

The priors used above can be used to control the degree of smoothness on the nonparametric regression function. However, in many cases, economic theory may provide additional inequality restrictions on $f(.)$ (e.g. monotonicity, concavity, etc.). Classical approaches to the imposition of such constraints tends to be quite complicated (see, e.g., Matzkin, 1994). However, such restrictions can be easily incorporated in our Bayesian approach at a small computational cost by imposing them in the prior (see Geweke, 1986).

Consider the semiparametric regression model given in (2.1) or (2.2). In addition to prior information which says $D\gamma$ is small, we now have the additional restriction that $\gamma \in \Omega$, for some $\Omega \subset R^N$. The latter restriction can be imposed by modifying (2.5) to:

$$R\delta \sim N(0_{N-m}, \sigma^2 V(\eta))\mathbf{1}(\gamma \in \Omega), \tag{2.16}$$

where $\mathbf{1}(.)$ is the indicator function. Let the restricted posterior corresponding to (2.16) be $p_R(\delta, \sigma^{-2}|y)$, and the unrestricted posterior based on (2.5) be $p_U(\delta, \sigma^{-2}|y)$. It follows that:

$$p_R(\delta, \sigma^{-2}|y) = p_U(\delta, \sigma^{-2}|y)\mathbf{1}(\gamma \in \Omega).$$

Depending on the form of the restrictions, it may be possible to directly take random draws from $p_R(\delta, \sigma^{-2}|y)$. Alternatively, a posterior simulator for the restricted model can be developed by simply taking random draws from $p_U(\delta, \sigma^{-2}|y)$ and discarding those which violate the inequality restrictions. This is formally equivalent to importance sampling with $p_U(\delta, \sigma^{-2}|y)$ as the importance function.[3]

2.5.1 Illustration: Imposing Constraints on the Nonparametric Regression Function

In this sub-section, we illustrate the case where $f(.)$ is known to be monotonically increasing, $m = 1$, and a variant of the Base Prior is used. The prior in (2.16) becomes:

$$R\delta \sim N(0_{N-1}, \sigma^2 \eta I_{N-1})\mathbf{1}(R\delta > 0).$$

A posterior simulator can be derived for drawing directly from the restricted posterior. The Appendix provides details. The key step in the derivation is to note that standard results for the Normal-Gamma distribution imply that the resulting marginal posterior for $R\delta$ will be a multivariate-t distribution truncated at 0.

[3] Simulation methods will not be necessary if the inequality restrictions are linear in the parameters (e.g. $\gamma_1 > \gamma_2$). In such cases, textbook results for the truncated Normal or Student-t distributions can be used to obtain posterior moments subject to the constraints. Here we stay with simulation methods because of their greater generality.

Figures 4a and b plot results for the fitted nonparametric regression lines for the constrained and unconstrained models for an artificial data set generated from a monotonically increasing DGP. In particular, for $i = 1, .., 100$ we generate

$$y_i = x_i + \varepsilon_i,$$

where ε_i is i.i.d. N(0,.09) and x_i is i.i.d. Uniform over [0,1]. In both models $\eta = .04$.

It can be seen that imposing monotonicity does seem to provide some additional smoothing of the fitted nonparametric regression line. However, the main difference is posterior standard deviations have been reduced due to the increasing amount of prior information in the restricted model.

Insert Figure 4 here

2.6 Model Comparison

The classical nonparametrics literature provides many tests of specification (e.g. Zheng, 1996). Some of these are fairly easy to implement (e.g. involving an auxiliary regression of the nonparametric residuals on x). However, the properties of these tests are asymptotic and difficult to derive. The finite sample performance of these tests is poorly understood. From a Bayesian perspective, nonparametric regression poses no new complications for model comparison. Hence, the well-known Bayesian tools of model comparison can be applied in a straightforward manner if one wishes to compare a parametric to a nonparametric model. Yatchew (1998) argues that one of the problems with classical nonparametric methods is that "....a unified framework for constrained estimation and testing of economic models using nonparametric regression is still in the incipient stage [Yatchew (1998, page 673)]." For the Bayesian method outlined in this paper, no such problem exists and Bayes factors and posterior odds ratios can be calculated analytically and given the standard interpretation.

Consider comparing the partial linear model given in (2.2), which shall be labelled M_1, to the following parametric model, M_2:

$$y = W_* \delta_* + \varepsilon_*, \tag{2.17}$$

where W_* is a N$\times k_*$ matrix with ith row given by[4] $w_{*i} = (1, z_i', g(x_i)')$ where $g(x_i)$ is a vector of known functions of x. $\varepsilon_* = (\varepsilon_{*1}, ..., \varepsilon_{*N})'$ with ε_{*i} i.i.d. N(0,σ_*^2). $\delta_* = (\beta_*', \gamma_*')'$ is a $k_* \times 1$ vector of parameters

[4] Note that the nonparametric effect implicitly puts an intercept in the model, hence in the partial linear model there is no intercept. In the purely parametric model, it typically is sensible to include an intercept and we do so here.

which can be broken down into β_*, a $k+1$-vector of coefficients on $(1, z_i)$, and γ_*, a $k_* - k - 1$ vector of coefficients on $g(x_i)$. Different choices of $g(.)$ can be made to carry out a wide range of tests. For instance, if $g(.) = 0$ then a purely nonparametric test of whether the marginal effect of x on y is zero is obtained. If $g(x_i) = x_i$ then we obtain of test of whether the nonparametric model is preferred to a standard linear regression model. Alternatively, $g(.)$ could be chosen to be a cubic regression spline function as in Poirier (1973) or a flexible Fourier form as in Hong and White (1995).

The formula for the Bayes factor comparing M_1 to M_2 is given in (A.19).

2.6.1 Illustration: Model Comparison

To illustrate Bayesian model comparison methods, we calculate posterior model probabilities for three models:

- M_1: the partial linear model,
- M_2: a parametric model with $g(x) = x$,
- M_3: a parametric model with $g(x) = 0$.

To focus discussion, we use only the Base Prior with $m = 1$ for the semiparametric partial linear model.

For the house price data set, under M_1 we set $\eta = .24$ (i.e. the value chosen by cross-validation) as well as $\eta = .06$ and $\eta = .96$ (i.e. we halve and double the prior standard deviation as a prior sensitivity analysis). Under M_2 we select a wide range of values for $\underline{q}_*$. To aid in interpretation, note that the prior variance of γ_* is $\underline{q}_* \sigma^2$ (see Appendix for details) and that $E(\sigma^2|y,\mathrm{M}_1) = 319.75$. Under M_3, no further prior input is required. Table 2 presents the resulting posterior model probabilities. Of course, $\Pr(\mathrm{M}_3|y) = 1 - \Pr(\mathrm{M}_1|y) - \Pr(\mathrm{M}_2|y)$.

Table 2: Posterior Model Probabilities for Housing Data Set

$\underline{q}_*$	M_1	M_2	M_1	M_2	M_1	M_2
	$\eta = .24$	$\eta = .24$	$\eta = .06$	$\eta = .06$	$\eta = .96$	$\eta = .96$
.000001	.9465	.0268	.9825	.0087	.3896	.3053
.0001	.9460	.0273	.9824	.0089	.3873	.3093
.01	.9264	.0474	.9757	.0157	.3124	.4428
1	.9607	.0121	.9873	.0039	.4690	.1636
100	.9713	.0012	.9908	.0004	.5498	.0195

Our preferred prior is $\eta = .24$ and $\underline{q}_* = .01$. For this preferred prior and all priors with $\eta = .24$ or .06, there is strong evidence in favor of the semiparametric model. Results exhibit little sensitivity to our very wide range of priors for the parametric component. However, for $\eta = .96$, there is much less evidence for the semiparametric model. The reason for this is that $\eta = .96$ allows for the nonparametric regression line

to be extremely non-smooth. In practice, this yields only a small increase in fit. The benefit of this increase in fit is more than counteracted by the penalty incurred in the Bayes factor for having such a flat prior over the high-dimensional coefficient vector, γ. It is worth stressing that Bayesian model comparison methods contain a strong reward for parsimony. Hence, if relatively flat priors (e.g. those with large prior variances) are used for parameters which are not common to all models (i.e. γ and any coefficients in $g(x)$), then the model containing the relatively flat prior will be penalized. This effect is most clearly seen when we use a relatively flat prior for the nonparametric component (i.e. $\eta = .96$). However, it can also be seen when we increase $\underline{q}_*$ to allow for a relatively noninformative prior in M_2. Noting that $\underline{q}_* = 100$ implies an enormous prior standard deviation of around 200 for the coefficient on x (and the posterior mean, not reported here, on this coefficient is approximately -3), it is not surprising that such a flat prior implies that M_2 receives much less support (see the last row of Table 2).

The artificial data set using (2.15) is so nonlinear that it always provides massive evidence for the M_1. More interestingly, Table 3 contains posterior model probabilities using artificial data generated from a model where M_2 should be favored:

$$y_i = x_i + \varepsilon_i,$$

where ε_i is i.i.d. N(0,.09) and x_i is i.i.d. N(0,1). Cross-validation yields $\eta = .04$.As above, we also halve and double the prior standard deviation as a sensitivity analysis. Given the true value of σ^2 used in the data generating process, our choices for $\underline{q}_*$ of .1, 1 and 10 cover a reasonably wide range.

Table 3: Posterior Model Probabilities for Artificial Data (M_2)

$\underline{q}_*$	M_1	M_2	M_1	M_2	M_1	M_2
	$\eta = .04$	$\eta = .04$	$\eta = .01$	$\eta = .01$	$\eta = .16$	$\eta = .16$
.1	.9989	.0011	.0000	1.000	1.000	.0000
1	.0000	1.000	.0000	1.000	.0001	.9999
10	.0000	1.000	.0000	1.000	.0000	1.000

Except for the prior with $\underline{q}_* = .1$, there is overwhelming posterior support for the correct model, M_2. In fact, given a prior mean of zero, setting $\underline{q}_* = .1$ pulls M_2 towards the (incorrect) model M_3 (i.e. the tight prior pulls the coefficient on x towards zero when the value used in the DGP is one) and, hence, M_2 receives little support. This illustrates a reasonable property: if the prior for the parametric model is greatly at odds with the data, the more flexible semiparametric model is supported.

Table 4 presents comparable posterior model probabilities based on the following artificial data set which favors M_3:

$$y_i = \varepsilon_i,$$

where ε_i is i.i.d. N(0,.09) and x_i is i.i.d. N(0,1). The cross-validation procedure yields $\eta = .002$ (an extremely small value). As a prior sensitivity analysis, we also present results for $\eta = .02$ and .2.

Table 4: Posterior Model Probabilities for Artificial Data (M_3)

$\underline{q}_*$	M_1	M_2	M_1	M_2	M_1	M_2
	$\eta = .002$	$\eta = .002$	$\eta = .02$	$\eta = .02$	$\eta = .2$	$\eta = .2$
.1	.4780	.2166	.2465	.3127	.0013	.4145
1	.5555	.0896	.3087	.1393	.0018	.2011
10	.5916	.0305	.3409	.0493	.0021	.0746

As noted above, for small values of $\underline{q}_*$, M_2 becomes very similar to M_1. Hence, the pattern observed in the column for M_2 is unsurprising (i.e. M_2 receives successively less support as $\underline{q}_*$ increases). The more puzzling feature of Table 4 is why the nonparametric model receives so much support for our preferred choice of $\eta = .002$. The answer lies in the fact that this prior imposes a very high degree of smoothness imposed on M_1. Hence, $M_1$is virtually identical to M_3 (i.e. with $\eta = 0$ the nonparametric regression line is merely a horizontal line). Accordingly, M_1 and M_3 receive a similar amount of support from the data. However, as η increases in Table 4 the nonparametric model receives less and less support.

The purpose of this illustration is largely to show how conventional Bayesian model comparison procedures work well in the present context. For instance, Tables 2, 3, 4 show the typical trade-off between fit and parsimony exhibited by Bayes factors. However, it also illustrates the success of our cross-validation procedure in eliciting values for η. For instance, the data generating process used to produce Table 4 can be interpreted as either being M_3, or the nonparametric regression model with $\eta = 0$. Our cross-validation approach selects an extremely small value for η and, hence, the Bayes factor comparing M_1 to M_3 is roughly one.

2.7 Prediction

Consider the partial linear model given in (2.1) and suppose we wish to predict $y = y^*$ at $z = z^*$ and $x = x^*$, where $x_i < x^* < x_{i+1}$ for some i, $1 \leq i \leq N-1$. Let $\gamma^* = [f(x_1), ..., f(x_i), f(x^*), f(x_{i+1}), ..., f(x_N)]'$, $\delta^* = \left[\beta', \gamma^{*\prime}\right]'$, W^* be a $N \times (N+K+1)$ matrix formed by inserting 0_N between columns i and $i+1$ of the

I_N component of W, D^* be an $(N-m+1)\times(N+1)$ version of D defined in (2.6) and $R^*=[0_{N\times k}, D^*]$. Then (2.2) can be written as $y=W^*\delta^*+\varepsilon$. Furthermore,

$$y^*|y,z^*,x^*,\delta^*,\sigma^{-2}\sim N(z^*\beta+\gamma^*_{i+1},\sigma^2). \tag{2.18}$$

We extend (2.5) to incorporate prior information on $f(x^*)$:

$$R^*\delta^*\sim N(0_{N-m+1},\sigma^2V^*(\eta)), \tag{2.19}$$

where $V^*(\eta)$ is $(N-m+1)\times(N-m+1)$. Any of the three prior choices discussed for $V(\eta)$ can be used for $V_*(\eta)$. Note that $f(x^*)$ appears in the prior, but not the likelihood. It follows immediately that the posterior density of δ^* and σ^{-2} is Normal-Gamma. Using this Normal-Gamma posterior, we can integrate δ^* and σ^{-2} out of (2.18) to give:

$$p(y^*|y,z^*,x^*)=t\left(y^*|z^*E(\beta|y)+E(\gamma^*_{i+1}|y),\widetilde{s}^{*2}\left(1+[z^*,e_{i+1}]\,\widetilde{Q}^*\,[z^*,e_{i+1}]'\right),\widetilde{\nu}^*\right), \tag{2.20}$$

where $t(.|a,b,c)$ indicates the t-density with c degrees of freedom, mean a (if $c>1$) and variance $\frac{c}{c-2}b$ (if $c>2$). In (2.20), $e_{i+1}=(0_i',1,0_{N-i}')$ and $\widetilde{\nu}^*$, $\widetilde{s}^{*2}$, $\widetilde{Q}^*$ are the obvious analogues of $\widetilde{\nu}$, $\widetilde{s}^2$, $\widetilde{Q}$ defined in (A.9), (A.10) and (A.12). For instance, $\widetilde{Q}^*$ is merely (A.12) or (2.8) with all variables superscripted with a $*$.

2.7.1 Illustration: Prediction

We illustrate prediction using the house price data set and the Base Prior with $\eta=.24$ (the value chosen by cross-validation). For a house with median values for all elements of z and $x^*=4.08$ on the 8 point scale of the variable reflecting the location of the house (i.e. a hypothetical house which lies between the 44th and 45th houses in terms of or the ordered data), we obtain a predictive mean of 134.916 and standard deviation of 20.261. For a house with maximum values for all elements of z and $x^*=4.08$ we obtain a predictive mean and standard deviation of 252.586 and 26.274. For a house with minimum values for all elements of z and $x^*=4.08$ we obtain a predictive mean and standard deviation of 77.571 and 23.282.

2.8 Extension to Non-scalar x

In the preceding material, x_i was a scalar for $i=1,..,N$. In this case, the classical Differencing Estimator involved simply reordering the data as $x_1\leq ...\leq x_N$. and then differencing to remove the nonparametric effect. When x_i is a vector, then the classical Differencing Estimator proceeds by using a nearest neighbor algorithm to measure the distance between observations. The data are then reordered according to the

distance between observations. For instance, Yatchew (1998) defines the distance between observations i and j as:

$$dist_{i,j} = \sum_{l=1}^{p} (x_{il} - x_{jl})^2 , \tag{2.21}$$

where $x_i = (x_{i1}, .., x_{ip})'$. 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. Yatchew (1998, page 697) argues that the classical Differencing Estimator works well provided p does not exceed 3.

Once the data have been ordered, the Bayesian procedure described previously can be used with minimal alteration. The likelihood based on (2.1) and (2.2) and the Base and m^{th} Difference Priors are completely unaltered. The X-Prior is slightly altered by replacing (2.12) with

$$v_i = \eta \; dist_{i,i-1} \tag{2.22}$$

Note that, if the x_is are drawn from the p-dimensional Uniform distribution, then the average distance between any two points is $T^{-\frac{1}{p}}$. This indicates that the distance between and observation and its nearest neighbor can be very far indeed if p is large. In such cases, we would suspect that the X-Prior, since it incorporates information on this distance, might perform better than the other priors.

2.8.1 Illustration: Nonscalar x

As described in Section 2.2, the house price data set described in Yatchew (1998) has two measures of location which can be included in $f(.)$. We reorder the data using these two nonparametric regressors and the nearest neighbor algorithm described above. For brevity, we consider only the Base Prior and use cross-validation methods to select $\eta = .10$. Figure 5 plots $y_i - z_i E(\beta|y)$ and $E(\gamma_i|y)$ against the cumulative distance from the initial observation. As before, the fitted nonparametric regression line matches the data well and the approximate 95% highest posterior density interval contains virtually every data point.

Insert Figure 5 here

2.8.2 An Additive Model

The curse of dimensionality means that the preceding approach will not work well if the dimension of x_i is at all large relative to N. Hence, interest often centers on additive models which do not suffer from this

curse. This sub-section will discuss the following additive nonparamatric regression model:

$$y_i = f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) + \varepsilon_i, \tag{2.23}$$

where $f_j(.)$ for $j = 1, ..p$ are unknown scalar functions. A parametric component such as $z\beta$, can be added in a straightforward fashion. Extending the notation of section 2, we can write this model as:

$$y = \gamma_1 + \gamma_2 + ... + \gamma_p + \varepsilon, \tag{2.24}$$

where $\gamma_j = \left(\gamma_{1j}, ..., \gamma_{Nj}\right)' = \left[f_j(x_{1j}), ..., f_j(x_{Nj})\right]'$. In the case where x is a scalar, we used the simple intuition that, if we order the data points so that $x_1 \leq x_2 \leq ... \leq x_N$, then it is 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 data, so there is not one simple ordering which can be adopted. Nevertheless, Bayesian inference can be carried out using the priors and methods of the previous section by setting up a Gibbs sampler which involves sequentially drawing from $p\left(\gamma_1|y, \gamma_2, .., \gamma_p, \sigma\right)$, $p\left(\gamma_2|y, \gamma_1, \gamma_3, .., \gamma_p, \sigma\right)$, ..., $p(\gamma_p|y, \gamma_2, ..., \gamma_p, \sigma)$. Each of these conditional posteriors is of the form given in the previous section. That is, they are all multivariate Normal with mean and variance given by equations analogous to (2.9) and (2.10).

Formally, let $y^{(j)}$ indicate the dependent variable ordered according to the j^{th} explanatory variable (i.e. ordered so that $x_{1j} \leq x_{2j} \leq ... \leq x_{Nj}$) and define $\gamma_l^{(j)}$ analogously. Let

$$y^{(j*)} = y^{(j)} - \left(\gamma_1^{(j)} + ... + \gamma_{j-1}^{(j)} + \gamma_{j+1}^{(j)} + ... + \gamma_p^{(j)}\right).$$

We use priors comparable to the one of section 2.1:

$$D\gamma_j^{(j)} \sim N(0_{N-m}, \sigma^2 V(\eta_j)), \tag{2.25}$$

where D is defined in equation (2.6). It follows immediately from our earlier results that

$$p(\gamma_j^{(j)}|y, \gamma_1, .., \gamma_{j-1}, \gamma_{j+1}, .., \gamma_p, \sigma) = p(\gamma_j^{(j)}|y^{(j*)}, \sigma)$$

is Normal with

$$E(\gamma_j^{(j)}|y^{(j*)}, \sigma) = \left[I_N + D'V\left(\eta_j\right)^{-1} D\right]^{-1} y^{(j*)} \tag{2.26}$$

and

$$var(\gamma_j^{(j)}|y^{(j*)},\sigma) = \sigma^2\left[I_N + D'V\left(\eta_j\right)^{-1}D\right]^{-1}. \tag{2.27}$$

A Gibbs sampler can be set up where sequential draws are taken from $p(\gamma_j^{(j)}|y^{(j*)},\sigma)$ for $j = 1,..,p$ and then $p(\sigma|y,\gamma_1,..,\gamma_p)$. Standard Bayesian results for the linear regression imply that the conditional posterior for σ is inverted-Gamma.

There is an identification problem with the model in (2.23) in that constants may be 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 an arbitrary constant. Insofar as interest centers on the shapes of the $f_j(x_{ij})$ for $j = 1,..,p$ or the overall fit of the nonparametric regression line, the lack of identification is irrelevant. If desired, identification can be imposed in many ways (e.g. by setting $f(x_{12}) = ... = f(x_{1p}) = 0$).

2.8.3 Illustration: An Additive Model

For $i = 1,..,100$ we generate artificial data from

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

where ε_i is i.i.d. N(0,.09) and x_{i1} and x_{12} are i.i.d. Uniform over [0,1]. We take

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

and

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

We use the base prior with $\eta_1 = 0.20$ and $\eta_2 = 0.05$. These values are chosen to illustrate that different degrees of smoothing in the two additive components are allowed for in our approach. Of course, cross-validation or some other approach could be used to select these hyperparameters in a substantive application.

Figures 6a and 6b plot the fitted and true nonparametric regression lines for each of the two additive functions in our nonparametric regression model (i.e. $E\left(\gamma_j|y\right)$ and $f_j(x_{ij})$). The horizontal axis in these figures measures x_{ij} and the vertical axis $y_i - f(x_{il})$ for $j \neq l$. These figures indicate that we are successfully estimating $f_j(.)$ up to an additive constant. The fact that $\eta_1 > \eta_2$ is revealed in the greater smoothness of the curve in Figure 6b.

Insert Figure 6 here

3 Extensions

A wide range of extensions to the basic semiparametric regression model can be handled, at the cost of an increase in computational complexity (i.e. Markov Chain Monte Carlo, MCMC, techniques are required to simulate the posterior). The class of models which can be written in terms of a parameter vector (possibly including latent data) θ, and a nonparametric component, γ with an additive i.i.d. N(0,σ^2) error is a rich one which can be handled using the present set of techniques. In particular, if $p(\gamma, \sigma^{-2}|y, \theta)$ is Normal-Gamma and either $p(\theta|y)$ or $p(\theta|y, \gamma, \sigma^{-2})$ can conveniently be sampled from, then MCMC methods can be used to carry out Bayesian nonparametric inference. The list of models which can be put in this form is huge and includes many qualitative choice models (e.g. probit, logit, etc.), truncated or censored models (e.g. tobit), stochastic frontier models, Normal mixture models, switching regression models, etc. Furthermore, the assumption that ε is multivariate Normal can be relaxed by using mixtures of Normals.[5] That is, if $p(\varepsilon_i|\theta)$ is Normal then the present paper provides complete posterior results *conditional on* θ. For most models, θ can be integrated out using MCMC methods to yield results which are not conditional on θ.

In this section, we illustrate how such extensions can be handled by developing Bayesian semiparametric methods for probit and tobit.

3.1 Semiparametric Probit

Bayesian methods for a semiparametric probit model can be derived by combining the ideas of the preceding part of this paper with standard Bayesian results for probit models (e.g. McCulloch and Rossi, 1994). The semiparametric probit model can be written as:

$$y_i^* = z_i\beta + f(x_i) + \varepsilon_i, \tag{3.1}$$

or

$$y^* = W\delta + \varepsilon. \tag{3.2}$$

where all model assumptions are as in Section 2, except that $y^* = (y_1^*, .., y_N^*)'$ is unobserved. Instead we observe

$$\begin{array}{l} y_i = 1 \text{ if } y_i^* > 0, \\ y_i = 0 \text{ if } y_i^* \leq 0. \end{array} \tag{3.3}$$

[5] Mixtures of Normals are extremely flexible. The Student-t distribution is a scale mixture of Normals. Location mixtures of Normals allow for multi-modal distributions. Dirichlet mixtures of Normals can approximate any continuous distribution and, hence, work with Dirichlet mixtures is often referred to as nonparametric.

Bayesian inference for this model proceeds by noting that[6] $p(\delta, \sigma^{-2}|y^*)$ is precisely what was analyzed in Section 2. Furthermore, $p(y^*|y, \delta, \sigma^{-2}) = \prod_{i=1}^{N} p(y_i^*|y_i, \delta, \sigma^{-2})$ and $p(y_i^*|y_i, \delta, \sigma^{-2})$ is truncated Normal. Hence, a simple Gibbs sampler with data augmentation can be used to carry out Bayesian inference.

Making the usual identifying assumption that $\sigma^2 = 1$, the MCMC algorithm involves sequentially drawing from (see A.11 and A.12 for notation):

$$\delta|y^* \sim N(\widetilde{\delta}, \widetilde{Q}) \tag{3.4}$$

and, for i=1,..,N,

$$\begin{array}{l} y_i^*|y_i, \delta \sim N(z_i\beta + \gamma_i, 1)I(y_i^* > 0) \text{ if } y_i = 1 \\ y_i^*|y_i, \delta \sim N(z_i\beta + \gamma_i, 1)I(y_i^* < 0) \text{ otherwise.} \end{array} \tag{3.5}$$

Extensions to allow for non-Normality of the error term through mixtures of Normals can be done by drawing on the ideas of Geweke and Keane (2000). In other words, a semiparametric mixtures of Normals probit model is a simple extension of the present model.

3.1.1 Illustration: Semiparametric Probit

We simulate artificial data for x_i and y_i^* from specification (2.15) and create y_i according to (3.3). Figure 7 plots the fitted nonparametric regression line using the Base Prior. This figure indicates that our approach is good at picking up the main features of the date generating process. For the sake of brevity, we do not present results for other priors. Nor do we implement constrained estimation or model comparison for the probit model. All these are straightforward extensions of our earlier work with the partial linear model.

Insert Figure 7 here

3.2 Semiparametric Tobit

Bayesian methods for a semiparametric tobit model can be derived along similar lines to semiparametric probit by combining the techniques outlined in Section 2 with standard Bayesian methods for parametric tobit models (Chib, 1992). The semiparametric tobit model can be written as:

$$y_i^* = z_i\beta + f(x_i) + \varepsilon_i, \tag{3.6}$$

or

[6] For completeness, it is worth noting that it can be easily shown that $p(\delta, \sigma^{-2}|y)$ is a valid probability density function.

$$y^* = W\delta + \varepsilon. \tag{3.7}$$

where all model assumptions are as in Section 2, except that $y^* = (y_1^*, .., y_N^*)'$ is unobserved. Instead we observe

$$\begin{array}{l} y_i = y_i^* \text{ if } y_i^* > 0, \\ y_i = 0 \text{ if.} y_i^* \leq 0 \end{array} \tag{3.8}$$

Bayesian inference for this model proceeds by noting that $\mathrm{p}(\delta, \sigma^{-2}|y^*)$ is precisely what was analyzed in Section 2. Furthermore,

$$p(y^*|y, \delta, \sigma^{-2}) = \prod_{i=1}^{N} p(y_i^*|y_i, \delta, \sigma^{-2})$$

and $p(y_i^*|y_i, \delta, \sigma^{-2})$ is either simply a point mass at y_i or truncated Normal. Hence, a simple Gibbs sampler with data augmentation can be used to carry out Bayesian inference.

The MCMC algorithm involves sequentially drawing from (see A.9-A11 for notation):

$$\delta, \sigma^{-2}|y^* \sim NG(\widetilde{\delta}, \widetilde{Q}, \widetilde{s}^{-2}, \widetilde{\nu}) \tag{3.9}$$

and, for i=1,..,N,

$$\begin{array}{l} y_i^* = y_i \text{ if } y_i > 0 \\ y_i^*|y_i, \delta \sim N(z_i\beta + \gamma_i, 1)I(y_i^* < 0) \text{ if } y_i = 0. \end{array} \tag{3.10}$$

3.2.1 Illustration: Semiparametric Tobit

We simulate artificial data for x_i and y_i^* from specification (2.15) and create y_i according to (3.8). Figure 8 plots the fitted nonparametric regression line using the Base Prior. This figure indicates that our approach is good at picking up the main features of the data generating process. As we did for probit, we do not present results for other priors. Nor do we implement constrained estimation or model comparison for the tobit model. All these are straightforward extensions of our work with the partial linear model.

Insert Figure 8 here

4 Conclusions

In this paper, we have developed Bayesian semiparametric methods for estimation (constrained and unconstrained) and testing in the partial linear regression model. Our methods are based on the Normal linear

regression model with natural conjugate prior and, hence, allow for the existence of finite sample analytical results. The theoretical and computational simplicity of our approach contrasts with previous Bayesian and non-Bayesian work in the field of nonparametric regression. We argue that the Bayesian's need for a prior is comparable to the non-Bayesian's need to select a kernel and bandwidth. Furthermore, the prior information used in this paper is on the degree of smoothness in $f(.)$, a natural quantity to elicit in practice.

The paper also either provides precise details or general discussion of many extensions that can easily be done, albeit at a cost in terms of computational complexity. These extensions include semiparametric tobit and probit and models with non-Normal errors.

5 Appendix: Posterior Results for Partially Informative Prior

In this appendix we derive posterior results for the model given in (2.1) with a partially informative prior on $R\delta$, where $R = [0_{N-m,K}, D]$, where D is the $(N-m) \times N$ differencing matrix given in (2.6).

It is useful to transform (2.2) as:

$$y = U_1\alpha_1 + U_2\alpha_2 + \varepsilon = U\alpha + \varepsilon, \tag{A.1}$$

where $U_1 = W_1 - W_2R_2^{-1}R_1, U_2 = W_2R_2^{-1}, \alpha = (\alpha_1', \alpha_2')', \alpha_1 = (\delta_1, ..., \delta_{k+1})'$ and $\alpha_2 = R\delta$. W and R have been partitioned as $W = (W_1, W_2)$ where W_1 is $N \times (k+m)$ and W_2 is $N \times (N-m)$ and $R = (R_1, R_2)$ where R_1 is $(N-m) \times (k+m)$ and R_2 is $(N-m) \times (N-m)$. Note that R_2 is simply equal to D with the first m columns deleted. It can be verified that R_2 is nonsingular.

Assume a natural conjugate, Normal-Gamma prior for α and σ^{-2}. Using the standard notation for the Normal-Gamma (see Poirier, 1995, page 526), we write this prior as NG($\underline{\alpha}$,$\underline{A}$,$\underline{s}^{-2}$,$\underline{\nu}$). Using standard results (e.g. Poirier, 1995, page 527), it follows that the posterior for α and σ^{-2} is NG($\overline{\alpha}, \overline{A}, \overline{s}^{-2}, \overline{\nu}$) where

$$\overline{\alpha} = \overline{A}\left(\underline{A}^{-1}\underline{\alpha} + U'y\right), \tag{A.2}$$

$$\overline{A} = \left(\underline{A}^{-1} + U'U\right)^{-1}, \tag{A.3}$$

$$\overline{\nu} = \underline{\nu} + N \tag{A.4}$$

and

$$\overline{\nu s}^2 = \underline{\nu s}^2 + (y - U\overline{\alpha})'(y - U\overline{\alpha}) + (\overline{\alpha} - \underline{\alpha})'\underline{A}^{-1}(\overline{\alpha} - \underline{\alpha}). \tag{A.5}$$

Results based on the partially informative priors used in this paper can be derived as special limiting cases of (A.2)-(A.5). Note first that the traditional noninformative prior for σ^{-2} given in (2.4) is obtained by setting $\underline{\nu} = 0$. Next, we set $\underline{\alpha} = 0_{N+k}$ and note that the Normal-Gamma prior implies:

$$\alpha|\sigma \sim N(0_{N+k}, \sigma^2\underline{A}). \tag{A.6}$$

Since $\alpha_2 = R\delta$, if we choose

$$\underline{A} = \begin{bmatrix} aI_{k+1} & 0 \\ 0 & V(\eta) \end{bmatrix},$$

and let $a \to \infty$, we obtain the partially informative prior given in (2.5). Under these conditions, it can easily be verified that $\overline{A}$ exists and is non-singular and, hence, that all quantities in (A.2)-(A.5) exist and a valid posterior is obtained despite the fact that the number of regressors in (2.2) is greater than N. In particular, the posterior for α and σ^{-2} under the partially informative prior is NG$(\widetilde{\alpha}, \widetilde{A}, \widetilde{s}^{-2}, \widetilde{\nu})$, where

$$\widetilde{\alpha} = \widetilde{A}U'y, \tag{A.7}$$

$$\widetilde{A} = \begin{bmatrix} U_1'U_1 & U_1'U_2 \\ U_2'U_1 & V(\eta)^{-1} + U_2'U_2 \end{bmatrix}^{-1}, \tag{A.8}$$

$$\widetilde{\nu} = N \tag{A.9}$$

and

$$\widetilde{\nu}\widetilde{s}^2 = (y - U\widetilde{\alpha})'(y - U\widetilde{\alpha}) + \widetilde{\alpha}_2'V(\eta)^{-1}\widetilde{\alpha}_2. \tag{A.10}$$

where $\widetilde{\alpha}$ has been partitioned conformably with α as $(\widetilde{\alpha}_1, \widetilde{\alpha}_2)'$.

Messy, but straightforward, matrix manipulations allows us to transform back to (2.2) and work in terms of δ and σ^{-2}. The posterior for these parameters is NG$(\widetilde{\delta}, \widetilde{Q}, \widetilde{s}^{-2}, \widetilde{\nu})$, where

$$\widetilde{\delta} = \widetilde{Q}W'y, \tag{A.11}$$

and

$$\widetilde{Q} = \left(R'V(\eta)^{-1}R + W'W\right)^{-1}. \tag{A.12}$$

Using results for obtaining the inverse of a partitioned matrix, we can find the posterior means for the parametric and nonparametric components of $\delta = (\beta', \gamma')$:

$$E(\gamma|y) = \left[M_Z + D'V(\eta)^{-1}D\right]^{-1} M_z y \tag{A.13}$$

and

$$E(\beta|y) = \left\{Z'\left(D'\left[DD' + V(\eta)\right]^{-1}D\right)Z\right\}^{-1} Z'\left(D'\left[DD' + V(\eta)\right]^{-1}D\right)y, \tag{A.14}$$

where $M_Z = I_N - Z\left(Z'Z\right)^{-1}Z'$. It can be verified that (A.14) is equivalent to the OLS estimate from a regression of Δy on ΔZ if $V(\eta) = \eta I - DD'$.

To carry out inference in the model with monotonicity imposed (section 2.3), it proves easiest to work in the parameterization given in (A.1). The resulting posterior is $\text{NG}(\widetilde{\alpha}, \widetilde{A}, \widetilde{s}^{-2}, \widetilde{\nu})\mathbf{1}(\alpha_2 > 0)$. Integrating out σ^{-2} we obtain:

$$\alpha|y \sim t(\widetilde{\alpha}, \widetilde{A}, \widetilde{\nu})\mathbf{1}(\alpha_2 > 0),$$

where $t(.,.,.)$ is the multivariate-t distribution (Poirier, 1995, page 126). Simulation from the truncated multivariate-t can be carried out as described in Geweke (1991) yielding draws from the posterior for α, which can be transformed to draws of δ.

The marginal likelihood for this partially informative prior is undefined due to the impropriety of the prior on β and σ^2. However, our purpose is to use the marginal likelihood to compare models which are identical in all respects except for the treatment of the nonparametric component. Thus, β and σ^{-2} are common to all models. Accordingly, it makes sense to use the same prior hyperparameter choices for these parameters (i.e. choose the same values for a, $\underline{s}^{-2}, \underline{\nu}$ for all models). Bayes factors comparing two models will then be defined, even for the noninformative case considered here: $a \rightarrow \infty$ and $\underline{\nu} \rightarrow 0$. Justification of this common Bayesian practice of using improper priors on parameters common to all models is given in many places (e.g. Kass and Raftery, 1995, page 783).

For the parametric model, M_2, given in (2.14), we assumed a Natural-conjugate prior consistent with these principles:

$$(\delta_*, \sigma_*^{-2}) \sim NG(\underline{\delta}_*, \underline{Q}_*, \underline{s}_*^{-2}, \underline{\nu}_*).$$

We set $\underline{\nu}_* = 0$, $\underline{\delta}_* = 0$[7] and for $\underline{Q}_*$ we assume the following structure which is noninformative for β_*, but informative for γ_*:

$$\underline{Q}_* = \begin{bmatrix} qI_{k+1} & 0 \\ 0 & \underline{q}_* \end{bmatrix},$$

and let $q \to \infty$.

The resulting posterior is:

$$(\delta_*, \sigma_*^{-2})|y \sim NG(\overline{\delta}_*, \overline{Q}_*, \overline{s}_*^{-2}, \overline{\nu}_*),$$

where

$$\overline{\delta}_* = (\overline{\beta}_*', \overline{\gamma}_*')' = \overline{Q}_* {W_*}' y, \tag{A.15}$$

$$\overline{Q}_* = \begin{bmatrix} Z'Z & Z'g(x) \\ g(x)'Z & \underline{q}_*^{-1} + g(x)'g(x) \end{bmatrix}^{-1}, \tag{A.16}$$

where $g(x) = (g(x_1), ..., g(x_N))'$

$$\overline{\nu}_* = N \tag{A.17}$$

and

$$\overline{\nu}_* \overline{s}_*^2 = \left(y - W_* \overline{\delta}_*\right)' \left(y - W_* \overline{\delta}_*\right) + \overline{\gamma}_*' \underline{q}_*^{-1} \overline{\gamma}_*. \tag{A.18}$$

Under these assumptions, the Bayes factor in favor of the partial linear model, M_1, relative to the purely parametric model, M_2 is given by:

$$B_{12} = \left[\frac{|\widetilde{A}^{-1}|/|V(\eta)^{-1}|}{|\overline{Q}_*^{-1}|/|\underline{q}_*^{-1}|} \right]^{-\frac{1}{2}} \left[\frac{\left(\widetilde{\nu}\widetilde{s}^2\right)^{-\frac{\widetilde{\nu}}{2}}}{\left(\overline{\nu}_* \overline{s}_*^2\right)^{-\frac{\overline{\nu}_*}{2}}} \right]. \tag{A.19}$$

[7] These assumptions can be trivially be relaxed with minor changes in the following formulae.

6 References

Chamberlain, G. and Leamer, E. (1976). "Matrix weighted averages and posterior bounds," *Journal of the Royal Statistical Society B*, 38, 73-84.

Chib, S. (1992). "Bayes inference in the Tobit censored regression model," *Journal of Econometrics*, 51, 79-99.

Dey, D., Muller, P. and Sinha, D. (eds.), (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer-Verlag.

Fernandez, C., Osiewalski, J. and Steel, M. (1997). "On the use of panel data in stochastic frontier models with improper priors," *Journal of Econometrics*, 79, 169-193.

Geweke, J. (1986). "Exact inference in the inequality constrained Normal linear regression model," *Journal of Applied Econometrics*, 1, 127-141.

Geweke, J. (1991). "Efficient simulation from the multivariate Normal and Student-t distributions subject to linear constraints," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, E.M. Keramidas and S.M. Kaufman (eds.), Interface Foundation of North America.

Geweke, J. and Keane, M. (2000). "Mixtures of Normals probit models," in C. Hsiao, K. Lahiri, L-F Lee and M. H. Pesaran (eds.), *Analysis of Panels and Limited Dependent Variables: A Volume in Honor of G. S. Maddala*. Cambridge: Cambridge University Press, forthcoming.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Greene, W. (2000). *Econometric Analysis*, fourth edition. New Jersey: Prentice Hall.

Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hardle, W. and Linton, O. (1994). "Applied nonparametric methods," chapter 38 in R. Engle and D, McFadden (eds.), *Handbook of Econometrics*, vol. 4. Amsterdam: North Holland.

Hong, Y. and White, H. (1995). "Consistent specification testing via nonparametric series regression," *Econometrica*, 63, 1133-1160.

Horowitz, J. (1998). *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.

Iwata, S. (1996). "Bounding posterior means by model criticism," *Journal of Econometrics*, 75, 239-261.

Judge, G., Griffiths, W., Hill, R.C., Lutkepohl, H. and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*, second edition. New York: John Wiley and Sons.

Kass, R. and Raftery, A. (1995). "Bayes factors," *Journal of the American Statistical Association*, 90,

773-795.

Leamer, E. (1982). "Sets of posterior means with bounded variance priors," *Econometrica*, 50, 725-736.

Leamer, E. (1984). "Global sensitivity results for generalized least squares estimates," *Journal of the American Statistical Association*, 79, 867-870.

McCulloch, R. and Rossi, P. (1994). "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64, 207-240.

Matzkin, R. (1994). "Restrictions of economic theory in nonparametric methods," chapter 42 in R. Engle and D, McFadden (eds.), *Handbook of Econometrics*, vol. 4. Amsterdam: North Holland.

O'Hagan, A. (1978). "Curve fitting and optimal design for prediction," *Journal of the Royal Statistical Society, Series B*, 40, 1-42.

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Poirier, D. (1973). "Piecewise regression using cubic spline," *Journal of the American Statistical Association*, 68, 515-524.

Poirier, D. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge: MIT Press.

Poirier, D. (1998). "Revising beliefs in nonidentified models," *Econometric Theory*, 14, 483-509.

Powell, J. (1994). "Estimation of semiparametric models," chapter 41 in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, vol. 4. Amsterdam: North Holland.

Rice, J. (1984). "Bandwidth choice for nonparametric regression," *Annals of Statistics*, 12, 1215-1230.

Shiller, R. (1973). "A distributed lag estimator derived from smoothness priors," *Econometrica*, 41, 775-788.

Silverman, B. (1985). "Some aspects of the spine smoothing approach to nonparametric regression curve fitting (with discussion)," *Journal of the Royal Statistical Society, Series B*, 47, 1-52.

Wahba, G. (1983). "Bayesian confidence intervals for the cross-validated smoothing spline," *Journal of the Royal Statistical Society, Series B*, 45, 133-150.

Whitaker, E. (1923). "On a new method of graduation," *Proceedings of the Edinburgh Mathematical. Society*, 41, 63-75.

Yatchew, A. (1998). "Nonparametric regression techniques in economics," *Journal of Economic Literature*, 36, 669-721.

Zheng, J. (1996). "A consistent test of functional form via nonparametric estimation techniques," *Journal of Econometrics*, 75, 263-289.

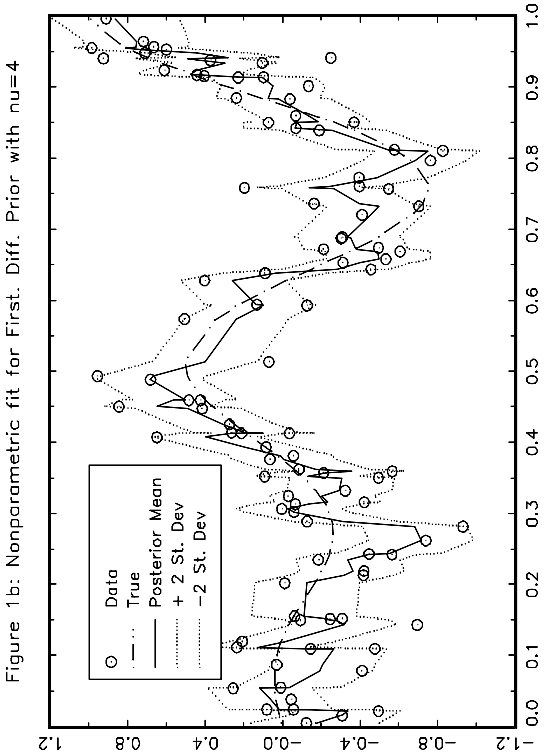

Figure 1b: Nonparametric fit for First. Diff. Prior with nu=4

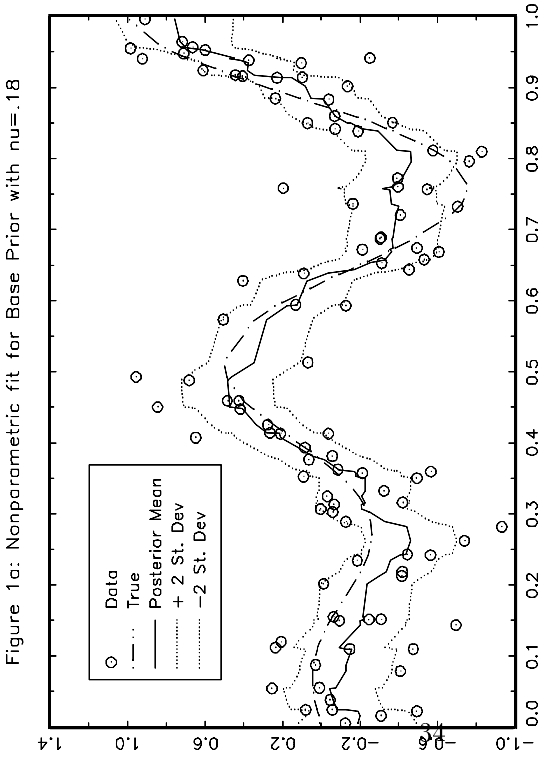

Figure 1a: Nonparametric fit for Base Prior with nu=.18

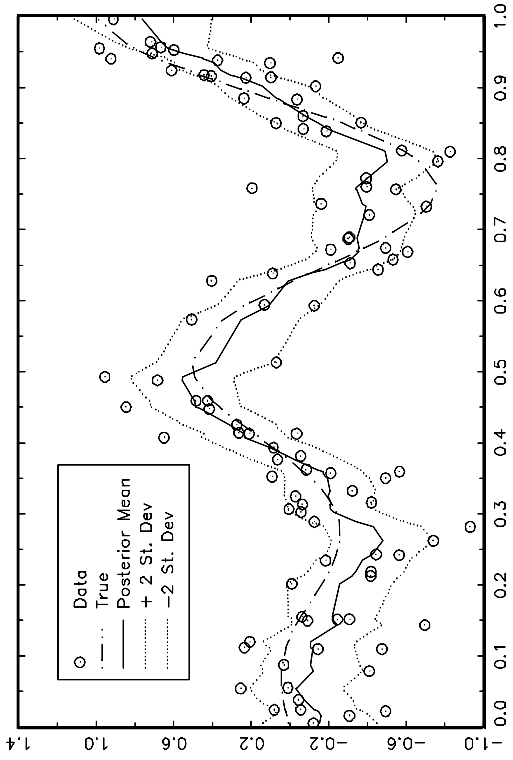

Figure 1c: Nonparametric fit for X−Prior with nu=19.7

Figure 1:

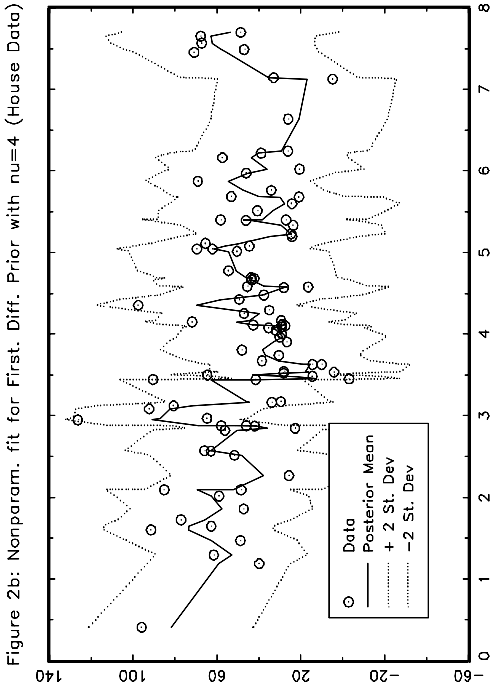
Figure 2b: Nonparam. fit for First. Diff. Prior with nu=4 (House Data)
Data
Posterior Mean
+ 2 St. Dev
−2 St. Dev

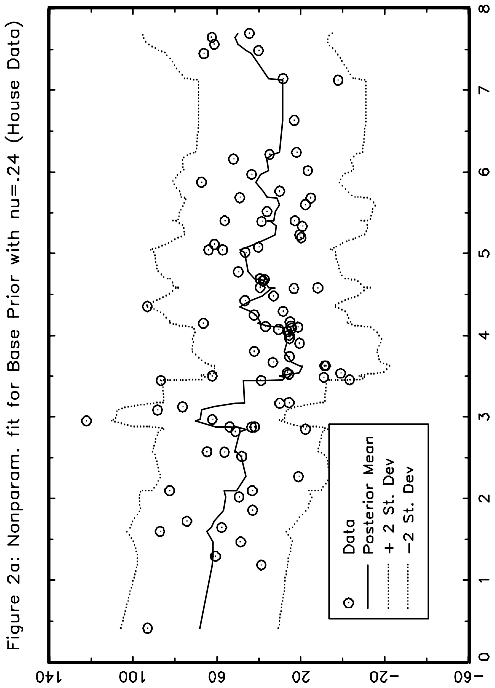
Figure 2a: Nonparam. fit for Base Prior with nu=.24 (House Data)
Data
Posterior Mean
+ 2 St. Dev
−2 St. Dev

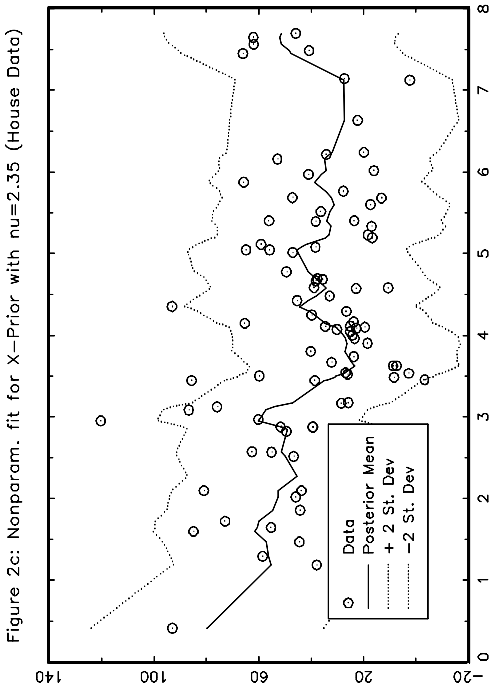
Figure 2c: Nonparam. fit for X−Prior with nu=2.35 (House Data)
Data
Posterior Mean
+ 2 St. Dev
−2 St. Dev

Figure 2:

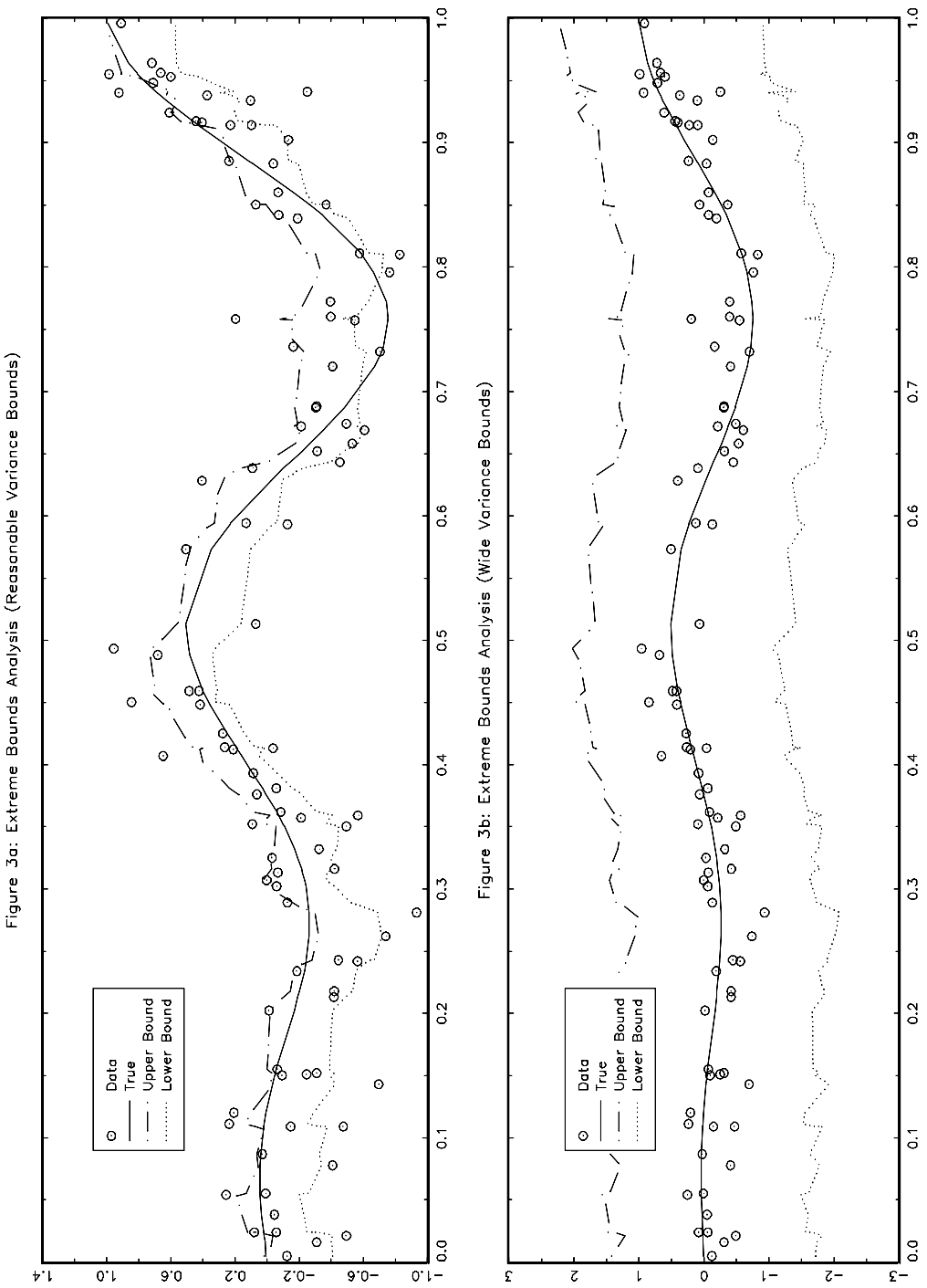
Figure 3a: Extreme Bounds Analysis (Reasonable Variance Bounds)
Data
True
Upper Bound
Lower Bound
1.4
1.0
0.6
0.2
−0.2
−0.6
−1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Figure 3b: Extreme Bounds Analysis (Wide Variance Bounds)
Data
True
Upper Bound
Lower Bound
3
2
1
0
−1
−2
−3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 3:

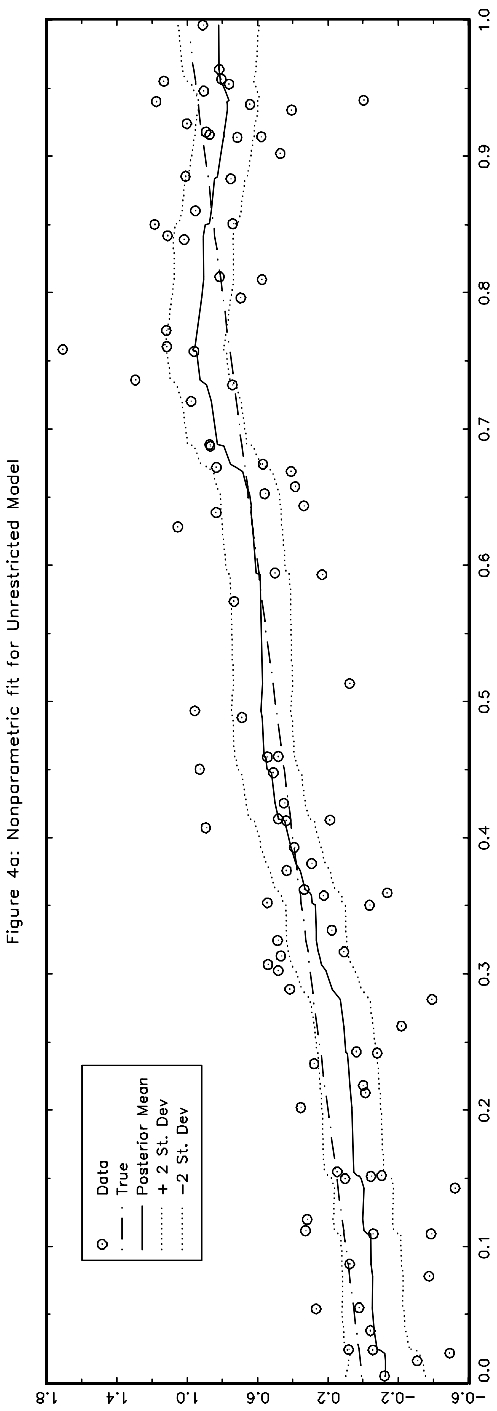
Figure 4a: Nonparametric fit for Unrestricted Model
Data
True
Posterior Mean
+ 2 St. Dev
−2 St. Dev
1.8
1.4
1.0
0.6
0.2
−0.2
−0.6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

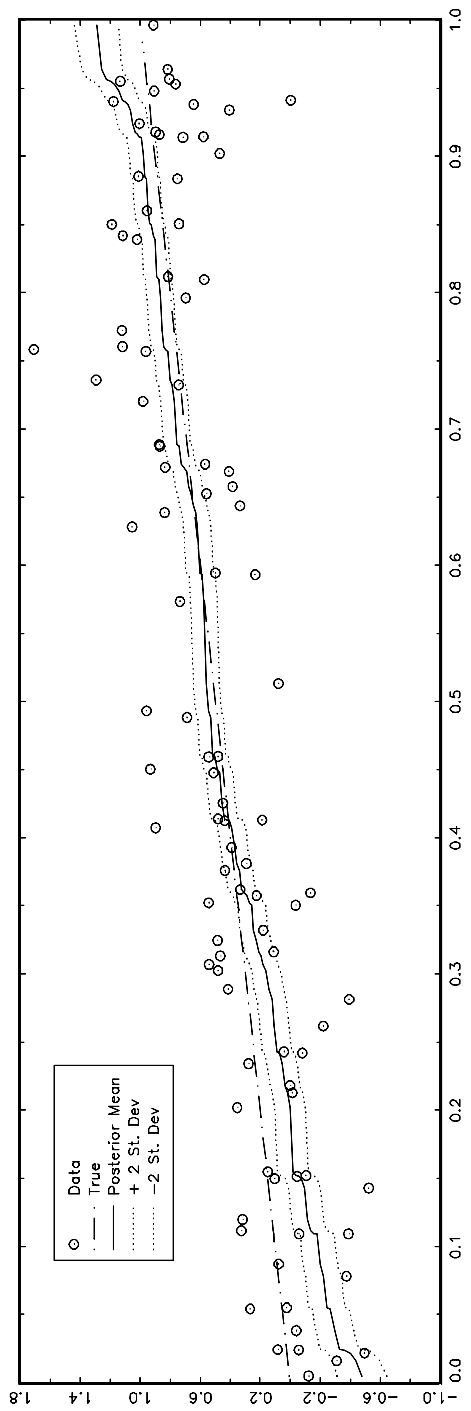
Figure 4b: Nonparametric fit for Restricted Model
Data
True
Posterior Mean
+ 2 St. Dev
−2 St. Dev
1.8
1.4
1.0
0.6
0.2
−0.2
−0.6
−1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4:

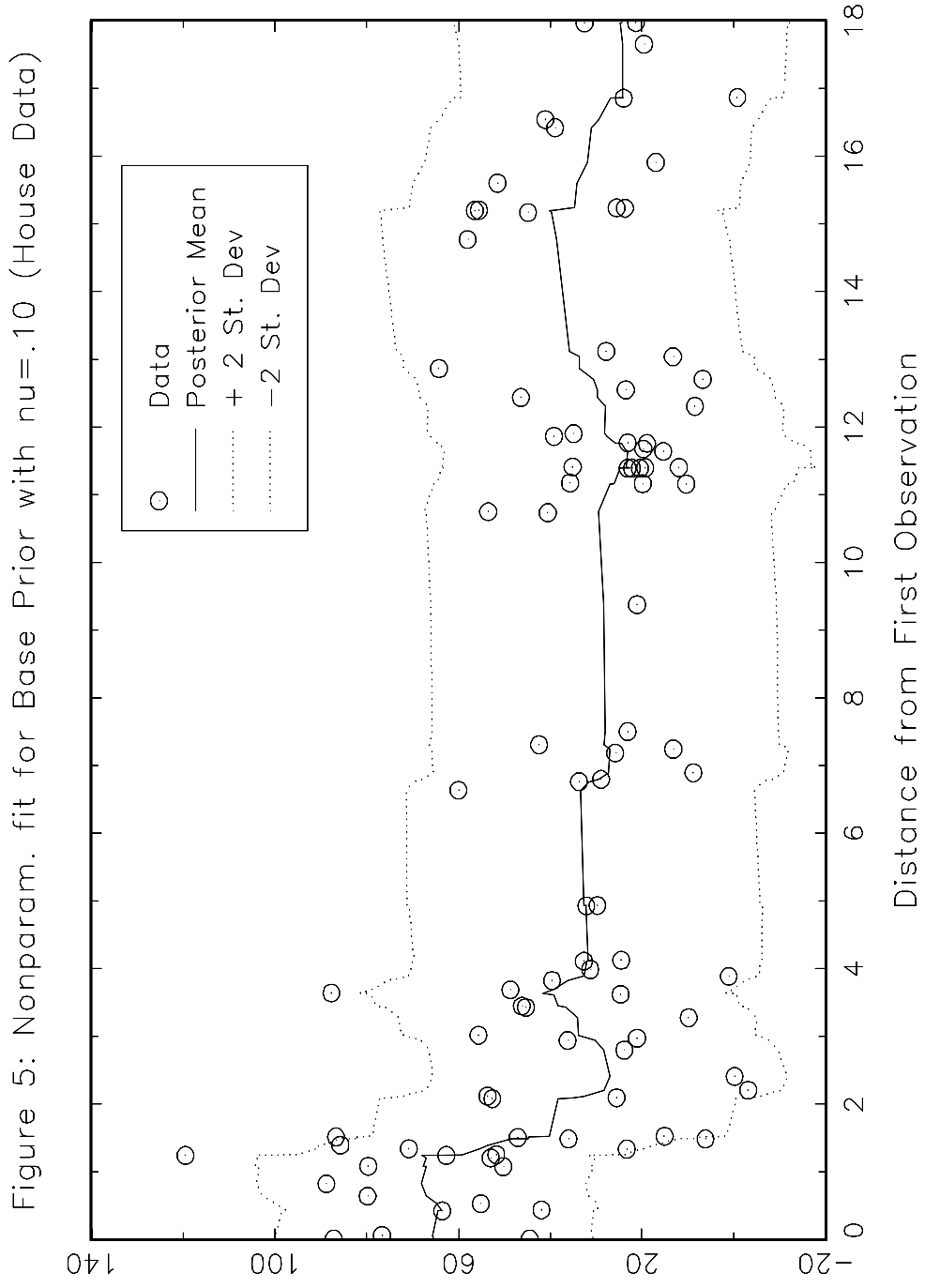
Figure 5: Nonparam. fit for Base Prior with nu=.10 (House Data)
Data
Posterior Mean
+ 2 St. Dev
−2 St. Dev
140
100
60
20
−20
0
2
4
6
8
10
12
14
16
18
Distance from First Observation

Figure 5:

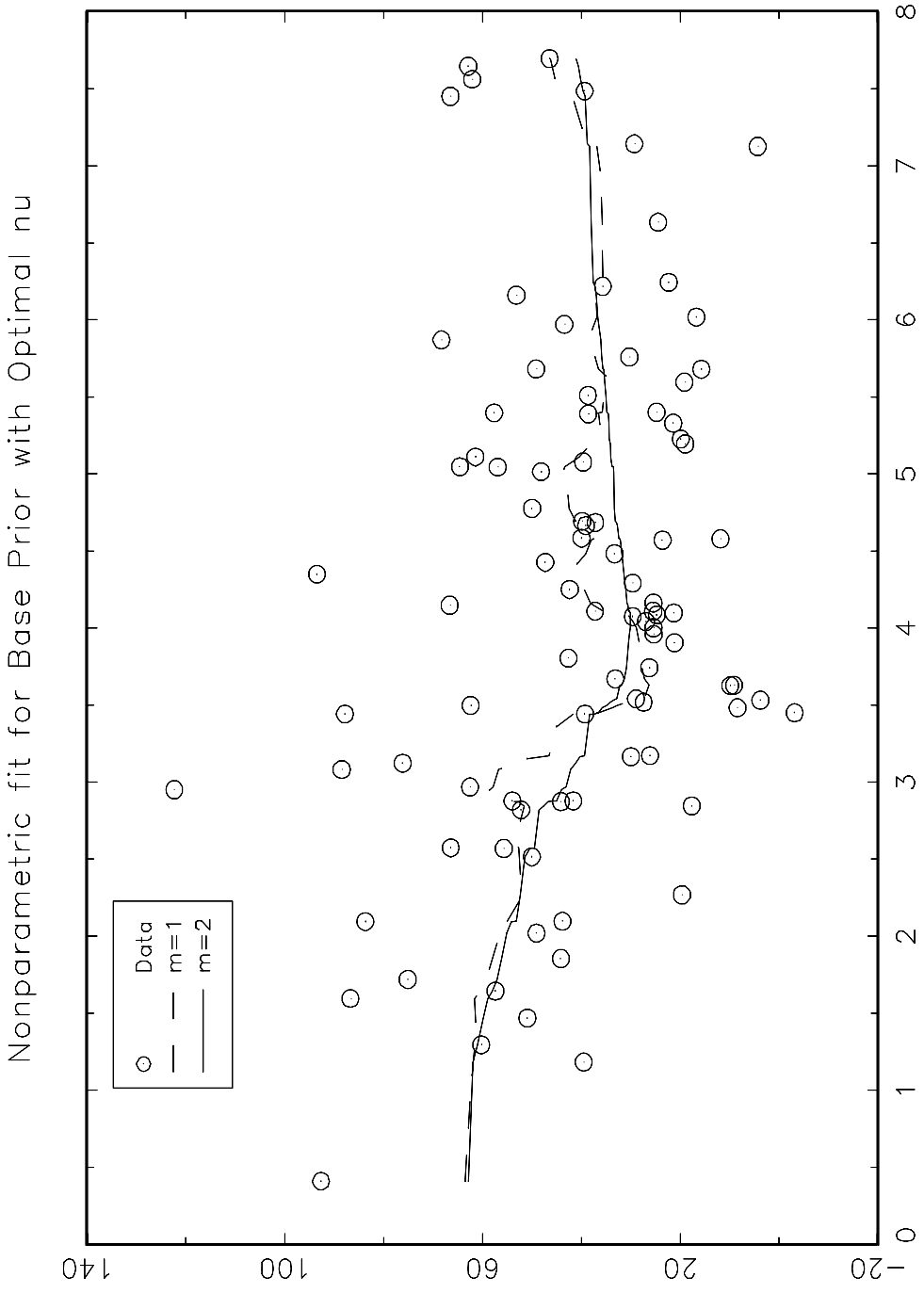
Nonparametric fit for Base Prior with Optimal nu
Data
m=1
m=2
140
100
60
20
−20
0
1
2
3
4
5
6
7
8

Figure 6:

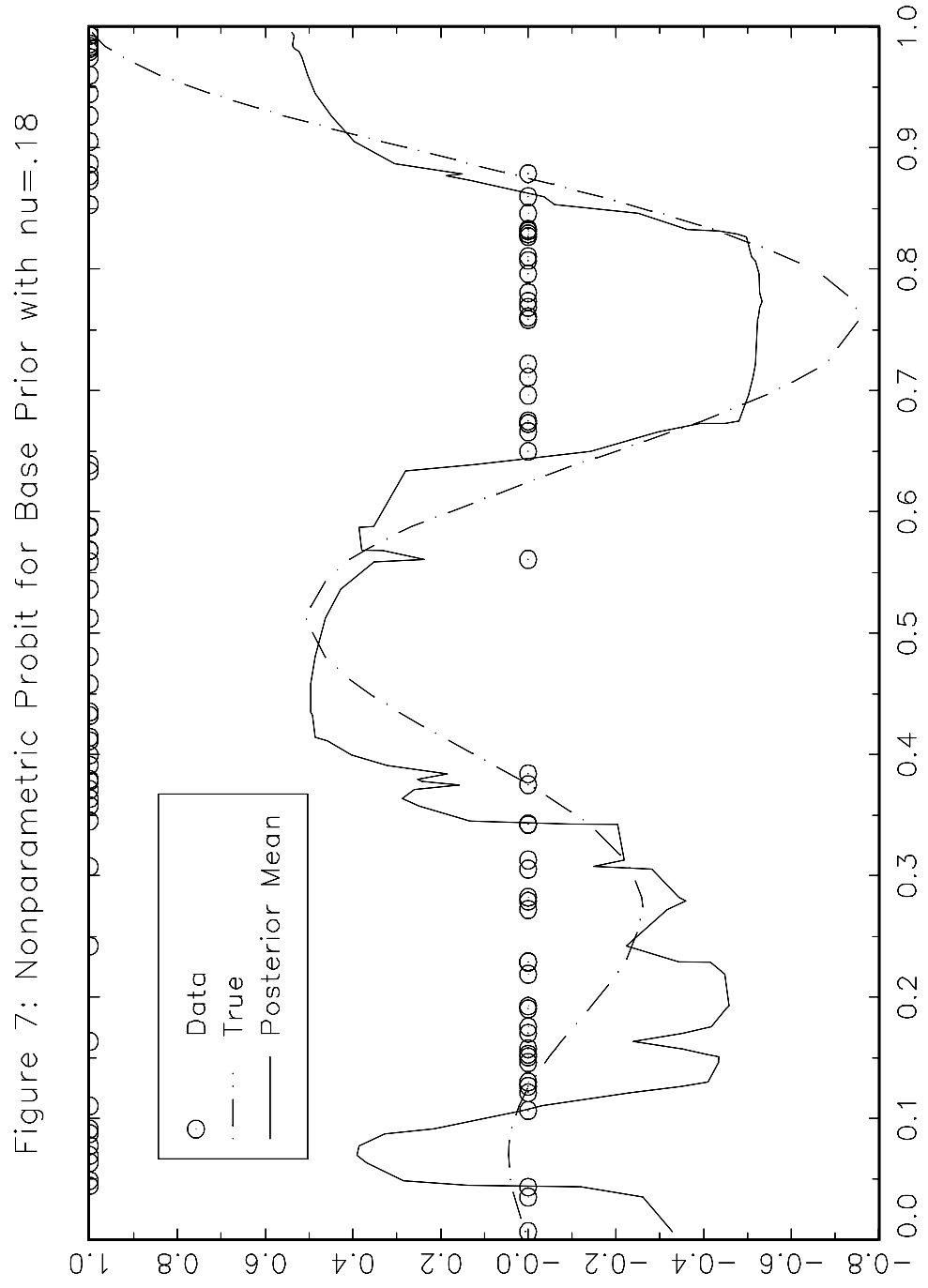
Figure 7: Nonparametric Probit for Base Prior with nu=.18
Data
True
Posterior Mean
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 7:

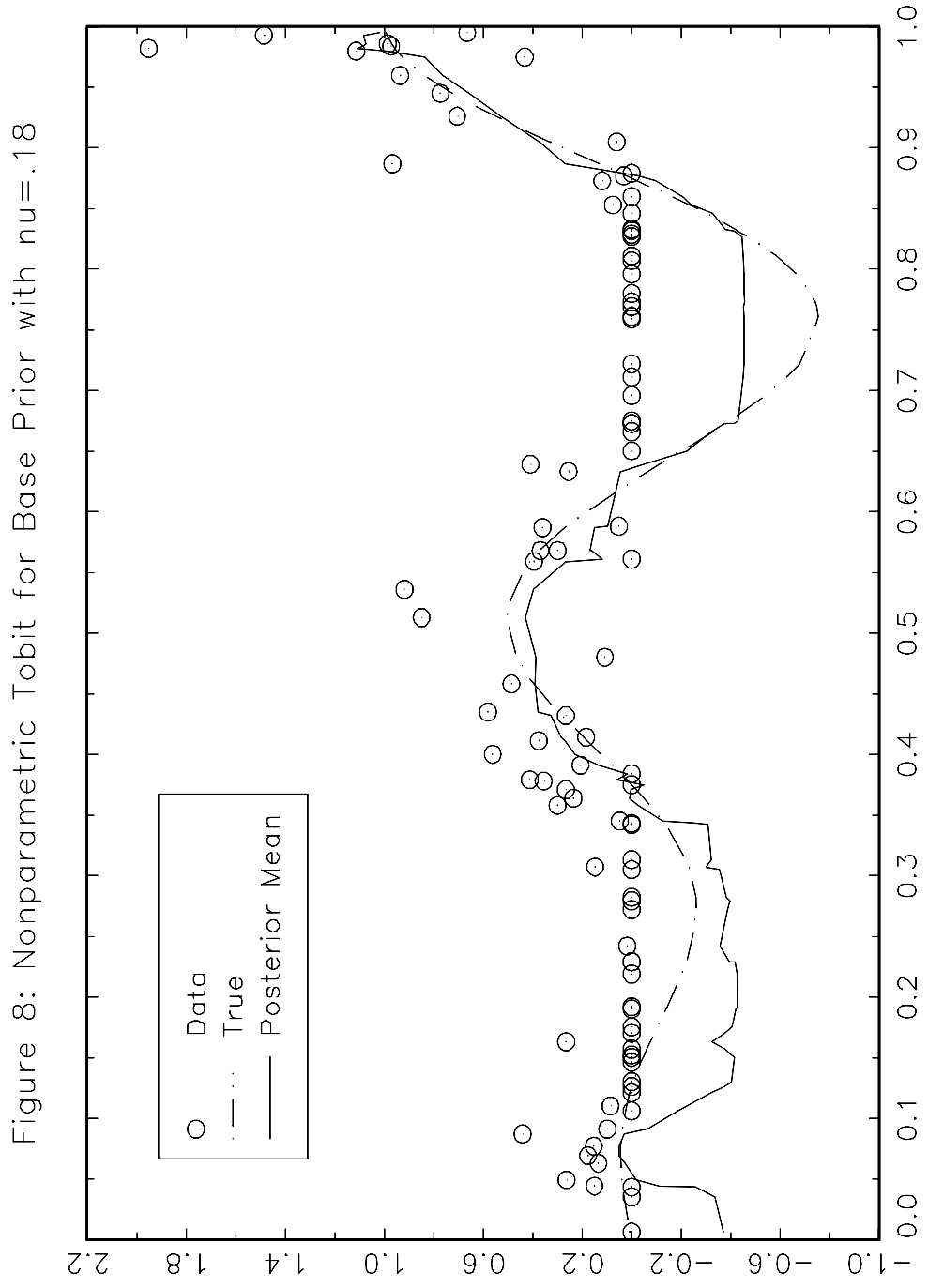
Figure 8: Nonparametric Tobit for Base Prior with nu=.18
Data
True
Posterior Mean
2.2
1.8
1.4
1.0
0.6
0.2
−0.2
−0.6
−1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 8: