University of Leicester

DEPARTMENT OF ECONOMICS

SEMIPARAMETRIC BAYESIAN INFERENCE IN SMOOTH COEFFICIENT MODELS

Gary Koop
University of Leicester

Justin L. Tobias
University of California-Irvine

Working Paper No. 04/18
October 2003

Semiparametric Bayesian Inference In Smooth Coefficient Models

Gary Koop
University of Leicester
Department of Economics
Gary.Koop@leicester.ac.uk

and

Justin L. Tobias
Department of Economics
University of California-Irvine
jtobias@uci.edu

October 2003

Abstract

We describe procedures for Bayesian estimation and testing in both cross sectional and longitudinal data *smooth coefficient models* (with and without endogeneity problems). The smooth coefficient model is a generalization of the partially linear or additive model wherein *coefficients* on linear explanatory variables are treated as unknown functions of an observable covariate. In the approach we describe, points on the regression lines are regarded as unknown parameters and priors are placed on differences between adjacent points to introduce the potential for smoothing the curves. The algorithms we describe are quite simple to implement - estimation, testing and smoothing parameter selection can be carried out *analytically* in the cross-sectional smooth coefficient model, and estimation in the hierarchical models only involves simulation from standard distributions.

We apply our methods by fitting several hierarchical models using data from the National Longitudinal Survey of Youth (NLSY). We explore the relationship between ability and log wages and flexibly model how returns to schooling vary with measured cognitive ability. In a generalization of this model, we also permit endogeneity of schooling and describe simulation-based methods for inference in the presence of the endogeneity problem. We find returns to schooling are approximately constant throughout the ability support and that simpler (and often used) parametric specifications provide an adequate description of these relationships.

1 Introduction

Perhaps the single most important limitation to the use of fully nonparametric regression techniques in practice is the well known *curse of dimensionality* problem, wherein the rate of convergence of the nonparametric estimator slows with the number of variables treated in a nonparametric fashion. In light of these dimensionality considerations, and given the need to control for many variables in most empirical studies in economics, many researchers have made use of the *partially linear* or *semilinear* regression model (e.g., Robinson (1988), Yatchew (1998) and DiNardo and Tobias (2001)). This model mitigates the dimensionality problem by treating one (or a few) key variables nonparametrically while maintaining parametric assumptions regarding the remaining set of explanatory variables.

An important variant of the partially linear model which has received decidedly less attention in empirical work is the *smooth coefficient model* (e.g. Li, Huang, Li and Fu (2002)). In addition to simply treating one or a few explanatory variables nonparametrically (as a partial linear model would), the smooth coefficient model lets the marginal effect of a given variable be represented as an unknown function of an observable covariate. That is, instead of restricting the marginal effect of y with respect to x_j to be constant and equal to a parameter β_j, the smooth coefficient model writes this marginal effect as an unknown function of some explanatory variable, say z_j. This specification nests the traditional linear model as a special case when the marginal effect is found to be constant over the support of z.

In this paper we continue to motivate the use of the smooth coefficient model in applied work and introduce and employ Bayesian methods for estimating various models which have a smooth coefficient form. We begin by showing how Bayesian methods can be used to fit a cross-sectional model as described in Li *et al* (2002). We then develop a generalized set of tools for estimating smooth coefficient models in a hierarchical (longitudinal) context, and finally, in a longitudinal data context with an endogeneity problem.

The types of models we describe in this paper are of general interest, and the methods we apply to estimate them are intuitive and can easily be applied by practitioners. In our view, the particular approaches described in this paper also offer some advantages over existing methods, and we highlight the following benefits:

1. Estimation of the various models is relatively simple and only requires simulation from standard distributions. In the *cross-sectional* model, posterior distributions can be obtained *analytically.*
2. Testing of the cross-sectional smooth coefficient model against parametric alternatives is straightforward, as marginal likelihoods and Bayes factors can also be calculated analytically.
3. The appropriate amount of smoothing of the regression functions is determined by the data via an empirical Bayes approach. This data-based selection rule helps to mitigate concerns regarding subjectivity in the choice of smoothing parameters.

4. If the data-based selection rule is not used, then prior elicitation only requires that the researcher express beliefs about the degree of "smoothness" of the nonparametric regression function rather than beliefs regarding the values of the functions themselves.

5. Techniques are described for extending the standard cross-sectional smooth coefficient model to a panel data model, and a panel model with an endogeneity problem.

6. Finally, our approach provides exact finite sample results based on the given data, and thus avoids the use of complicated asymptotics (which are potentially inappropriate in modest samples) for inference.

In addition to the theoretical contributions, we provide an application showing how the estimation techniques can be used in practice. Specifically, we use the National Longitudinal Survey of Youth (NLSY) panel to explore the relationship between measured cognitive ability and log wages, and also to determine how returns to schooling vary with this measure of cognitive ability. While the issue of nonlinearities in ability has been documented in several previous studies (e.g., Cawley *et al* (1999), Heckman and Vytlacil (2001), DiNardo and Tobias (2001) and Tobias (2003)), fewer studies have investigated how observed ability affects the return to education. Among those that have (e.g., Blackburn and Neumark (1993), Heckman and Vytlacil (2001) and Tobias (2003)), individuals have either been classified into discrete educational groups to facilitate estimation within groups,[1] particular functional forms such as education-ability interactions have been assumed,[2] or the panel structure of the data has not been fully exploited.[3] The smooth coefficient model described in this paper can fully account for the panel structure of the NLSY, imposes virtually no structure on the way returns to schooling vary with ability, and uses the workhorse Mincerian linear-in-schooling model as the point of departure. We find that returns to education are essentially constant over the ability support, and also find that simpler (and widely-used) parametric models perform adequately in capturing key features of the NLSY data.

The outline of this paper is as follows. In the next section we introduce our class of smooth coefficient models and briefly describe our posterior simulators for fitting them. In section 3 we describe the NLSY data used in our empirical example, and section 4 provides generated data experiments and results from this application. The paper concludes with a summary in section 5, and remaining details regarding priors and the posterior simulators can be found in the appendix.

[1] Heckman and Vytlacil (2001) primarily consider three educational groups: high school dropouts, high school graduates and college graduates.

[2] Blackburn and Neumark (1995), for example, use ability-education-time interactions in their model. Heckman and Vytlacil (2001) relax many of these restrictions by using a linear regression spline with knots placed at each ability quartile.

[3] Tobias (2003), for example, primarily makes year-by-year comparisons, and categorizes individuals into two groups - those with 12 or fewer years of schooling, and those with some college education.

2 The Models

In this section we introduce three related *smooth coefficient models* and discuss Bayesian estimation and testing strategies for each model. Though these particular models are introduced with an eye toward our empirical example, the specifications we consider are quite general and can be used in a variety of applications. We begin in section 2.1 with a discussion of the cross-sectional smooth coefficient model. This is generalized in sections 2.2 and 2.3, where we take up the cases of a hierarchical smooth coefficient model and a hierarchical model with an endogeneity problem, respectively.

2.1 A Cross-Sectional Smooth Coefficient Model

To begin we consider the simplest case of a cross-sectional smooth coefficient model as in Li et al (2002) of the form

$$y_i = w_i\theta + f_1(A_i) + s_i f_2(A_i) + \varepsilon_i, \quad i = 1, 2, \cdots N \tag{1}$$

where, in this cross-sectional case, y_i is a scalar, w_i is a k_w vector of exogenous variables treated parametrically, s_i is an explanatory variable and $f_1(\cdot)$ and $f_2(\cdot)$ are unknown functions which depend on an exogenous variable A_i. We assume $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$.[4]

This model is termed a "smooth coefficient model" since the function f_2 acts as the "coefficient" on s_i, and we model this function as depending in a *smooth* way on an observed covariate A. To provide a concrete example of the potential usefulness of such a model, let us jump ahead to our empirical application. In the application of section 4 y_i will denote the log hourly wage received by individual i, A_i will be a continuous measure of cognitive ability, s_i will denote years of schooling completed, and w_i will denote a remaining set of characteristics affecting wages. Thus, the smooth coefficient model will enable us to investigate if there are possible nonlinearities in the ability-log wage relationships through f_1 (e.g., Cawley et al (1998), Heckman and Vytlacil (2001), DiNardo and Tobias (2001)) and additionally will enable us to flexibly estimate how returns to schooling vary with ability through the function f_2.

In this paper we develop a semiparametric framework similar to that described in Koop and Poirier (2003a,b) to model $f_1(A_i)$ and $f_2(A_i)$. Intuitively, we treat each point on the nonparametric regression lines as an unknown parameter. Specifically, let $\gamma_{ji} = f_j(A_i)$ for $j = 1, 2$ denote the N points on each nonparametric regression line and stack them into matrices as $\gamma_j = (\gamma_{j1}, .., \gamma_{jN})'$, $j = 1, 2$. Letting $\mu_i = w_i\theta + f_1(A_i) + s_i f_2(A_i)$ denote the value of the conditional mean function in (1), we can write

$$\mu = W\theta + I_N\gamma_1 + S\gamma_2 \equiv V\lambda, \tag{2}$$

[4] Note that this assumption can be easily relaxed by replacing it with the assumption that ϵ follows a finite mixture of Normals (e.g., McLachlan and Peel (2000)). To fix ideas on the estimation of f_1 and f_2 we do not describe in detail how this mixture of Normals extension could be done. Essentially, the algorithm we describe would be used within a given mixture component, and individuals can be ascribed to the various components of the mixture in a data augmentation step (e.g. Tanner and Wong (1987)).

where $\mu = (\mu_1, .., \mu_N)'$and W is an $N \times k_w$ matrix constructed from w_i in an analogous fashion. Furthermore, I_N is the $N \times N$ identity matrix, S is a diagonal matrix with i^{th} diagonal element given by s_i, $V = (W : I : S)$ and $\lambda = (\theta', \gamma_1', \gamma_2')'$.

Without imposing any additional structure to our model, we are plagued by the problem of *insufficient observations* in that we have more than twice as many parameters as observations. The complications caused by the high dimensionality of the resulting parameter space, however, *can be resolved through the use of prior information about the degrees of smoothness of the nonparametric regression lines.* The remainder of this section describes how we approach specifying this smoothing prior.

Without loss of generality, we assume the data are ordered so that $A_1 < A_2 < \cdots < A_N$. Define the $(N-2) \times N$ second-differencing matrix as:[5]

$$D = \begin{bmatrix} 1 & -2 & 1 & . & . & . & . & 0 \\ 0 & 1 & -2 & 1 & 0 & . & . & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & . & . & . & 1 & -2 & 1 \end{bmatrix}, \tag{3}$$

so that $D\gamma_j$ is the vector of second differences of points on the j^{th} nonparametric regression line, denoted $\Delta^2\gamma_{ji}$. For future reference, partition D as follows: $D = [D^* : D^{**}]$, where D^* is $(N-2) \times 2$ and define the 2×1 vector of initial conditions in f_1 and f_2 as γ_j^0 for $j = 1, 2$. Let γ_j^* be γ_j with these first two elements deleted. Rearranging the columns of V conformably into the matrix V^* and defining $\lambda^* = \left(\theta', \gamma_1^{0\prime}, \gamma_2^{0\prime}, \gamma_1^{*\prime}, \gamma_2^{*\prime}\right)'$ we can write (2) as

$$\mu = V^* \lambda^*. \tag{4}$$

Prior information about the degrees of smoothness in the nonparametric regression lines can be expressed in terms of $R\lambda^*$, where the $2(N-2) \times (k_w + 2N)$ matrix R is given as

$$R = \begin{bmatrix} 0 & D^* & 0 & D^{**} & 0 \\ 0 & 0 & D^* & 0 & D^{**} \end{bmatrix}. \tag{5}$$

For future reference, partition R as $R = [R_1 : R_2]$ where R_1 is an $2(N-2) \times (k_w + 4)$ matrix and R_2 is $2(N-2) \times 2(N-2)$.

It will prove to be useful to transform (4) to work directly with the parameter vector of second differences. Using standard transformations (see, e.g. Poirier (1995) pages 503-504), (4) can be written as:

$$\mu = X^{(1)}\beta_1 + X^{(2)}\beta_2 \equiv X\beta, \tag{6}$$

where $\beta = (\beta_1', \beta_2')'$, $\beta_1 = \left(\theta', \gamma_1^{0\prime}, \gamma_2^{0\prime}\right)'$, $\beta_2 = \left[(D\gamma_1)', (D\gamma_2)'\right]'$, $X^{(1)} = V^{(1)} - V^{(2)}R_2^{-1}R_1$ and $X^{(2)} = V^{(2)}R_2^{-1}$. In the previous expressions we have used the partition $V^* = \left[V^{(1)} : V^{(2)}\right]$ where $V^{(1)}$ is $N \times (k_w + 4)$. Note that β_2 is the vector of second differences of the points on the nonparametric regression lines and it is on this parameter vector that we place our smoothness prior.

[5] Note that other degrees of differencing can be handled by re-defining (3) as appropriate (see, e.g., Yatchew (1998) pages 695-698 or Koop and Poirier (2003a)).

To complete our Bayesian analysis of the cross-sectional smooth coefficient model, we specify a natural conjugate prior. Using the standard notation (e.g. Poirier (1995, p. 526)) for the Normal-Gamma (NG) prior, we write

$$\beta, \sigma_\varepsilon^{-2} \sim \text{NG}(\underline{\beta}, \underline{V}_\beta, \underline{s}^{-2}, \underline{\nu}). \tag{7}$$

Note that this prior implies $\beta|\sigma_\varepsilon^{-2} \sim N\left(\underline{\beta}, \sigma_\varepsilon^2 \underline{V}_\beta\right)$ and marginally, σ_ε^{-2} has a Gamma distribution with mean $\underline{s}^{-2}$ and variance $2/[\underline{\nu s}^4]$.

Of course, any values for the prior hyperparameters $\underline{\nu}$, $\underline{s}^{-2}$, $\underline{\beta}$ and $\underline{V}_\beta$ can be chosen, yet it is of interest to consider the use of suitably "diffuse" priors so that data information is predominant. In our empirical work we use a noninformative prior for the error variance (i.e. $\underline{\nu} = 0$ and, with this choice, $\underline{s}^{-2}$ is irrelevant) and add prior information on β to control the degree of smoothness of the nonparametric regression lines. To this end we set $\underline{\beta} = 0_{k_w+2N}$ so that the second differences of the regression functions (and coefficients on w) are centered over a prior mean of zero. Below we focus on the selection of the prior covariance matrix $\underline{V}_\beta$ which will govern the smoothness of the nonparametric regression curves.

We describe a particular strategy for selecting $\underline{V}_\beta$ that requires a minimal amount of subjective prior information. In particular, we assume

$$\underline{V}_\beta = \underline{V}_\beta(\eta_1, \eta_2) = \begin{bmatrix} \underline{V}_1 & 0 & 0 \\ 0 & V(\eta_1) & 0 \\ 0 & 0 & V(\eta_2) \end{bmatrix}, \tag{8}$$

where $\underline{V}_1$ is the prior covariance matrix for the parameters on the linear variables w and the initial conditions of our regression curves (i.e., the prior covariance matrix for β_1). Setting $\underline{V}_1^{-1} = 0$ yields the noninformative choice. The $(N-2) \times (N-2)$ matrices $V(\eta_j)$ are the prior covariance matrices placed over the second differences of f_j. Each of these depends on a scalar parameter η_j which will act as a *smoothing parameter*, similar in spirit to a bandwidth parameter in classical kernel-based methods.

Several sensible forms for $V(\eta_j)$ can be chosen, as discussed in Koop and Poirier (2003a). In this section we set $V(\eta_j) = \eta_j I_{N-2}$. This prior centers the second differences of the functions f_1 and f_2 around a mean of zero, and the scalar parameters η_1 and η_2 control the tightness around this mean and thereby the degree of smoothness of these functions. Our prior information about the smoothness of these curves is of the form: $\Delta^2\gamma_{ji} \sim N\left(0, \sigma_\varepsilon^2 \eta_j\right)$ for $i = 3, .., N,\ j = 1, 2.$[6] Intuitively, as η_1 and $\eta_2 \to \infty$, the prior becomes "diffuse," and with no additional structure placed on the model the resulting estimates will be undersmoothed. Conversely, as η_1 and $\eta_2 \to 0$, prior information will dominate, and will restrict the second differences to be identically zero (potentially oversmoothing).

[6] This approach to prior elicitation does not include any information in A_i other than order information (i.e. data is ordered so that $A_1 < ... < A_N$). If desired, the researcher could account for non-uniform spacing of the data by eliciting a prior of the form $\Delta^2\gamma_{ij} \sim N\left(0, \eta_j \Delta^2 A_i\right)$.

2.1.1 Estimation and Testing in the Cross-Sectional Smooth Coefficient Model

The approach of treating values of the functions $f_1(A_i)$ and $f_2(A_i)$ as parameters to be estimated proves to be quite convenient, since the resulting model fits into the framework of a linear regression model with a natural conjugate prior. As such, we can borrow from existing results for the analysis of such a model to address issues of estimation and testing in the cross-sectional smooth coefficient model.

Using standard Bayesian results for the Normal linear regression model with natural conjugate prior (e.g. Poirier (1995, p. 527)), it follows that the posterior for β and σ_ε^{-2} is

$$\beta, \sigma_\varepsilon^{-2}|\text{Data} \sim \text{NG}(\overline{\beta}, \overline{V}_\beta, \overline{s}^{-2}, \overline{\nu})$$

where

$$\overline{\beta} = \overline{V}_\beta \left(\underline{V}_\beta^{-1} \underline{\beta} + X'y \right), \tag{9}$$

$$\overline{V}_\beta = \left(\underline{V}_\beta^{-1} + X'X \right)^{-1}, \tag{10}$$

$$\overline{\nu} = \underline{\nu} + N \tag{11}$$

and

$$\overline{\nu}\overline{s}^2 = \underline{\nu}\underline{s}^2 + \left(y - X\overline{\beta}\right)' \left(y - X\overline{\beta}\right) + \left(\overline{\beta} - \underline{\beta}\right)' \underline{V}_\beta^{-1} \left(\overline{\beta} - \underline{\beta}\right). \tag{12}$$

The properties of the Normal-Gamma distribution also imply that it is trivial to transform back from the parameterization in (4) to the original parameterization given in (2).

Provided $\underline{V}_\beta^{-1}$ is non-singular, it can be verified that this is a proper posterior despite the fact that the number of explanatory variables exceeds the number of observations. In fact, we can go even further than this and show that a proper posterior exists even if $\underline{V}_1^{-1} = 0$, provided $W'W$ is nonsingular. Using properties of the Normal Gamma distribution, it also follows that the marginal posterior for β is multivariate-t (see, e.g., Poirier (1995), page 128) and thus *analytical* results for the Normal linear regression model with a natural conjugate prior can be used to carry out estimation and inference in the smooth coefficient model.

When *testing* the semiparametric smooth coefficient model against parametric alternatives or selecting values of the smoothing parameters, we calculate *log marginal likelihoods* for the models under consideration. Marginal likelihoods are widely used in Bayesian testing and their use arises from the observation that for any two competing models M_1 and M_2:

$$\frac{p(M_1|y)}{p(M_2|y)} = \left(\frac{p(y|M_1)}{p(y|M_2)} \right) \frac{p(M_1)}{p(M_2)}. \tag{13}$$

The left-hand side of (13) gives the *posterior odds* of Model 1 in favor of Model 2, and the ratio $p(M_1)/p(M_2)$ is the *prior odds ratio*, typically taken to be unity. The expression in parentheses following the equality in (13) is the *Bayes factor* or the ratio of marginal likelihoods, with $p(y|M_i)$ denoting the marginal likelihood for Model i. Thus, under equal prior odds, posterior odds ratios can be obtained by exponentiating the difference between the log marginal likelihoods.

The marginal likelihood associated with the linear regression model takes the form (Poirier (1995, p. 543)):

$$p\left(y\right)=c\left(\frac{\left|\overline{V}_{\beta}\right|}{\left|\underline{V}_{\beta}\right|}\right)^{\frac{1}{2}}\left(\overline{\nu s}^{2}\right)^{-\frac{\overline{\nu}}{2}}, \tag{14}$$

where

$$c\equiv\frac{\Gamma\left(\frac{\overline{\nu}}{2}\right)\left(\underline{\nu s}^{2}\right)^{\frac{\underline{\nu}}{2}}}{\Gamma\left(\frac{\underline{\nu}}{2}\right)\pi^{\frac{N}{2}}}. \tag{15}$$

Note that, formally, the marginal likelihood is not defined if we use a noninformative prior for the error variance (i.e. when $\underline{\nu}=0$) since the integrating constant is zero. However, insofar as we are using Bayes factors comparing models with the same noninformative prior for the error variance, the integrating constant cancels out and is irrelevant. Alternatively, setting $\underline{\nu}=0$ and ignoring c can be justified as being arbitrarily close to what would happen if you set $\underline{\nu}=\varepsilon$ for arbitrarily small ε.

There are several useful contexts in which one would be interested in calculating marginal likelihoods as in (14). First, posterior odds ratios can be used to provide an attractive and objective method for determining the appropriate degree of smoothing in our model. That is, M_1 and M_2 could both be smooth coefficient models, differing only in the values used for η_1 and η_2. If we calculate marginal likelihoods (using equation 14) for a variety of (η_1,η_2) combinations over a two dimensional grid, then we can find values of the smoothing parameters that are most supported by the data. Techniques such as this, which select prior hyperparameters which maximize the marginal likelihood, are referred to as *empirical Bayes methods*.[7]

Second, posterior odds ratios can be used to test the semiparametric smooth coefficient model against various parametric or semiparametric (e.g. the partial linear model) alternatives. As an example, let M_1 denote the smooth coefficient model with optimally chosen values of the smoothing parameters and let us consider a particular competitor, denoted M_2, which imposes the parametric restrictions $f_1(A)=\lambda_0+\lambda_1 A$ and $f_2(A)=\lambda_2$. In the context of our application, M_2 would denote the widely-estimated log wage equation with a linear ability term and a constant return to education.[8] To calculate the marginal likelihood associated with M_2, retain the specification $\mu=X\beta$ as in (6) where X and β are now defined as follows:

$$X=[W\ \iota_N\ A\ S] \quad \text{and} \quad \beta=[\theta'\ \lambda_0\ \lambda_1\ \lambda_2],'$$

with ι_N denoting a $N\times 1$ vector of ones. If we employ a natural conjugate prior, then the marginal likelihood for M_2 will be as in (14) (except with the new definition of X and the prior hyperparameters used in M_2 plugged in).

[7] As an aside, it is worth noting that the use of empirical Bayesian methods requires a small additional amount of prior information (relative to simply estimating the smooth coefficient model for a given choice of prior hyperparameters). As discussed in Koop and Poirier (2003b) use of noninformative priors over all the parameters other than β_2 is not possible since an improper posterior for η_1 and η_2 results. A proper prior is needed for either the error variance or the initial conditions. This motivates our choice of a proper (albeit virtually noninformative) prior for the initial conditions in the empirical section below. Note that we could be improper over these initial conditions as well if we were only interested in estimating the model for given values for η_1 and η_2.

[8] Note that this particular example is without loss of generality - one can impose any parametric restrictions, and conduct the model comparison in an identical manner.

Thus, testing the smooth coefficient model against a parametric or semiparametric alternative can also be done in a straightforward manner. *Prediction*, the other primary activity of the econometrician, can also be carried out rather simply using textbook results from the Normal linear regression model with natural conjugate prior (see, e.g., Poirier (1995), pages 551-558).

2.2 A Longitudinal (Hierarchical) Smooth Coefficient Model

In this section we generalize the cross-sectional model of Section 2.1 and show how such techniques can be used in a *longitudinal* or *hierarchical* data setting. We write

$$y_{it} = \alpha_i + z_{it}'\delta + \varepsilon_{it}, \quad i = 1, 2, \cdots N, \quad t = 1, 2, \cdots T_i, \tag{16}$$

where y_{it} is the dependent variable (in our application, it is the log of the hourly wage of individual i at time t), z_{it} is a vector of length k_z containing observations on exogenous explanatory variables which are time varying and $\varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2\right)$. The individual effects α_i are assumed to be independent of one another with

$$\alpha_i \sim N\left(\mu_i, \sigma_\alpha^2\right). \tag{17}$$

From a Bayesian point of view, (17) is a hierarchical prior for α_i, although a non-Bayesian may wish to interpret it as part of the likelihood function.

We assume that the conditional mean μ_i in (17) is of the same form as (2) and thus write

$$\mu_i = w_i\theta + f_1(A_i) + s_i f_2(A_i),$$

which can be re-parameterized as $\mu_i = x_i\beta$ as in (6) to work directly with second differences of the regression functions. Thus, we consider a specific hierarchical specification where the cross-sectional smooth coefficient model of section 2.1 applies to the mean function of the (time-invariant) second stage of our hierarchy.

The restriction that all of the nonparametric components enter through the second stage of the model is imposed with an eye toward our application. In this application we will want to treat our time-invariant measure of cognitive ability nonparametrically and allow for flexible interactions between ability and a linear schooling term (which is also time-invariant). In terms of estimation, however, this focus is essentially without loss of generality - similar methods to those described here can be used to fit this model when some (or all) of the nonparametric components appear at the first (time-varying) stage of the model.

Finally, to implement a Bayesian analysis we require priors for $\delta, \sigma_\varepsilon^{-2}, \sigma_\alpha^{-2}$ and β. For the first three of these we make standard choices:

$$\begin{aligned}
\delta &\sim N(\underline{\delta}, \underline{V}_\delta), & (18)\\
\sigma_\varepsilon^{-2} &\sim G(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon), & (19)\\
\sigma_\alpha^{-2} &\sim G(\underline{s}_\alpha^{-2}, \underline{\nu}_\alpha). & (20)
\end{aligned}$$

Noninformative priors for these parameters are the limiting cases $\underline{\nu}_\varepsilon = \underline{\nu}_\alpha = \underline{V}_\delta^{-1} = 0$.

As in the cross-sectional case, the prior for β is what governs the smoothing of the model, and we retain the same prior for this key parameter vector. Thus, we specify:

$$\beta|\sigma_\alpha^2 \sim N(\underline{\beta}, \sigma_\alpha^2 \underline{V}_\beta), \tag{21}$$

with prior hyperparameters selected as in Section 2.1 (i.e., the prior largely reflects the degree of smoothness in the nonparametric regression lines).[9]

2.2.1 Estimation and Testing in the Hierarchical Smooth Coefficient Model

Unlike the cross-sectional model of section 2.1, posterior distributions cannot be derived analytically in the hierarchical smooth coefficient model. However, posterior computation can be implemented using the *Gibbs sampler* - a widely used simulation-based algorithm which involves iteratively sampling from the posterior conditionals of the model. To facilitate convergence of the sampler, we employ a *blocking* or *grouping* step where the individual effects α and first-stage regression coefficients δ are drawn in a single block. We implement this blocking step by first sampling from the δ conditional marginalized over the individual effects, and then draw independently from the complete conditionals for the individual effects. Formally, this blocking step is completed by first sampling from

$$\delta|Data, \sigma_\alpha^2, \sigma_\varepsilon^2, \mu_i \sim N(D_\delta d_\delta, D_\delta), \tag{22}$$

where

$$D_\delta = (\sum_i Z_i' C_i^{-1} Z_i + \underline{V}_\delta^{-1})^{-1}, \quad d_\delta = \sum_i Z_i' C_i^{-1}(y_i - \iota_{T_i}\mu_i) + \underline{V}_\delta^{-1}\underline{\delta}.$$

In the previous equation, ι_{T_i} is a vector of ones, y_i and Z_i are $T_i \times 1$ and $T_i \times k_z$ matrices, respectively, containing observations for the i^{th} individual and $C_i \equiv \iota_{T_i}\sigma_\alpha^2 \iota_{T_i}' + \sigma_\varepsilon^2 I_{Ti}$. To complete the blocking step we then sample independently from the complete conditionals for the individual effects:

$$\alpha_i|Data, \sigma_\alpha^2, \sigma_\varepsilon^2, \delta, \mu_i \overset{ind}{\sim} N(D_{\alpha_i} d_{\alpha_i}, D_{\alpha_i}) \;\; i = 1, 2, \cdots, N, \tag{23}$$

where

$$D_{\alpha_i} \equiv (\frac{T_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\alpha^2})^{-1}, \quad d_{\alpha_i} \equiv \frac{\sum_{t=1}^{T_i}(y_{it} - z_{it}'\delta)}{\sigma_\varepsilon^2} + \frac{\mu_i}{\sigma_\alpha^2}.$$

For the remaining parameters of the model (i.e., σ_ε^{-2}, β and σ_α^2) we sample from their complete posterior conditionals. For the first-stage error precision we obtain:

$$\sigma_\varepsilon^{-2}|Data, \alpha, \delta \sim G\left(\overline{s}_\varepsilon^{-2}, \overline{\nu}_\varepsilon\right) \tag{24}$$

where

$$\overline{\nu}_\varepsilon = \sum_i T_i + \underline{\nu}_\varepsilon,$$

[9] Note that in the cross-sectional case, the prior analogous to the one in (7) was natural conjugate, here it is natural conjugate conditional upon $\alpha = (\alpha_1, .., \alpha_N)$.

$$\overline{s}_{\varepsilon}^{2} = \frac{\sum_i (y_i - \alpha_i - Z_i\delta)'(y_i - \alpha_i - Z_i\delta) + \underline{\nu}_{\varepsilon}\underline{s}_{\varepsilon}^{2}}{\overline{\nu}_{\varepsilon}},$$

and $\alpha = (\alpha_1, .., \alpha_N)'$.

The posterior conditional for β is given by:

$$\beta | Data, \alpha, \sigma_{\alpha}^{2} \sim N(\overline{\beta}, \sigma_{\alpha}^{2} D_{\beta}), \tag{25}$$

where

$$\overline{\beta} = D_{\beta} d_{\beta}, D_{\beta} = \left[X'X + \underline{V}_{\beta}^{-1}\right]^{-1}, \quad d_{\beta} = X'\alpha + \underline{V}_{\beta}^{-1}\underline{\beta}.$$

Finally, we sample the second-stage error precision σ_{α}^{-2} from

$$\sigma_{\alpha}^{-2} | Data, \alpha \sim G\left(\overline{s}_{\alpha}^{-2}, \overline{\nu}_{\alpha}\right) \tag{26}$$

where

$$\overline{\nu}_{\alpha} = N + \underline{\nu}_{\alpha}$$

and

$$\overline{s}_{\alpha}^{2} = \frac{\left(\alpha - X\overline{\beta}\right)'\left(\alpha - X\overline{\beta}\right) + \left(\overline{\beta} - \underline{\beta}\right)'\underline{V}_{\beta}^{-1}\left(\overline{\beta} - \underline{\beta}\right) + \underline{\nu}_{\alpha}\underline{s}_{\alpha}^{2}}{\overline{\nu}_{\alpha}}.$$

Thus, posterior analysis can be carried out using a Gibbs sampler which sequentially draws from (22), (23), (24), (25) and (26), and all of these densities are of standard forms.

Model comparison (e.g., testing the smooth coefficient model against a parametric alternative) or empirical Bayesian selection of η_1 and η_2 can be done using the same strategy involving marginal likelihoods as in Section 2.1. Unlike the cross-sectional smooth coefficient model of the previous section, however, marginal likelihoods for the hierarchical smooth coefficient model are not available in closed form, which poses a set of computational challenges. In particular, to use empirical Bayesian methods to estimate η_1 and η_2, the Gibbs sampler must be run for every (η_1, η_2) combination over a two-dimensional grid. This need for computational simplicity motivates our use of approximate methods for the necessary marginal likelihood calculation.

To this end, we use the Laplace-Metropolis estimator for the marginal likelihood (see, e.g., Raftery (1996), page 171). In particular, if we first integrate the individual effects α out of the likelihood function we obtain:

$$p\left(y|\delta, \sigma_{\varepsilon}^{2}, \beta, \sigma_{a}^{2}\right) \propto \prod_{i=1}^{N} |C_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(y_i - \iota_{T_i}\mu_i - Z_i\delta\right)' C_i^{-1}\left(y_i - \iota_{T_i}\mu_i - Z_i\delta\right)\right], \tag{27}$$

where $\mu_i = x_i\beta$. Then a highly accurate approximation to the marginal likelihood is given by:

$$p(y) \approx (2\pi)^{\frac{1}{2}(2N+k_w+k_z+2)} |\Psi|^{\frac{1}{2}} p\left(y|\widehat{\delta}, \widehat{\sigma_{\varepsilon}^{2}}, \widehat{\mu}, \widehat{\sigma_{a}^{2}}\right) p\left(\widehat{\delta}, \widehat{\sigma_{\varepsilon}^{2}}, \widehat{\mu}, \widehat{\sigma_{a}^{2}}\right), \tag{28}$$

where $\widehat{\cdot}$ denotes the posterior mode and Ψ is the posterior covariance matrix of $\left(\delta, \sigma_{\varepsilon}^{2}, \beta, \sigma_{a}^{2}\right)$.[10] Raftery (1996) remarks that the Laplace-Metropolis approximation is often much more accurate than other methods if the number of replications in the posterior simulator is relatively small.

[10] In practice, the simulated draws can be used to estimate the posterior mode and posterior covariance matrix.

2.3 A Longitudinal Smooth Coefficient Model with an Endogeneity Problem

Most of the model presented here is the same as the one described in the previous section, but additionally considers the issue where an explanatory variable at the second stage of the hierarchy is endogenous. To this end, we expand the model of section 2.2 to include a reduced form equation for the variable s_i[11]

$$\begin{aligned} y_{it} &= \alpha_i + z_{it}\delta + \varepsilon_{it} && (29) \\ \alpha_i &= w_i\theta + f_1(A_i) + s_i f_2(A_i) + u_i && (30) \\ s_i &= r_i\pi + v_i. && (31) \end{aligned}$$

In (31), r_i is a k_r vector of exogenous variables, possibly including w_i and A_i, but also including one or more instruments which are not present in (30). We consider a particular form of endogeneity problem wherein u_i and v_i are potentially correlated. In terms of our application, the parameter α_i is interpreted as an “individual” effect describing if a person earns log hourly wages that are higher or lower than expected, given the observable characteristics z. We can explain some of this variation in individual effects through time-invariant observables w_i (like parental education, etc.), ability A_i and schooling s_i. However, individual-level characteristics like “motivation” or “drive” also presumably affect one’s wages, and are captured in the error term u_i. It is also reasonable to assume that this unobserved motivation or drive is a factor that influences s_i, the quantity of schooling attained. As such it is seemingly reasonable to embrace the potential for correlation between u_i and v_i in an empirical analysis.

To intuitively describe how failure to account for this endogeneity problem may bias our results, first note that α_i is identifiable from the first stage of the hierarchy, and when T_i is reasonably large, the marginal posterior for α_i will tend to be dominated by data information from (29). To this end, let us suppose for the moment that the α_i are known. Equations (30) and (31) then constitute a triangular simultaneous equations model. If u_i and v_i are sufficiently correlated, a cross-sectional smooth coefficient analysis using only (30) (again treating α_i as known) will yield biased and inconsistent estimates of the parameters of interest. In this case, mean independence is violated as $E(u_i|s_i) \neq 0$. In short, if we ignore endogeneity, the coefficient (function) on the schooling variable s_i captures both the (structural) quantity of schooling return as well as the premium paid for unobserved “motivation” or “drive.” As such, when failing to account for this potential endogeneity problem, we might expect to obtain estimates of the return to education that are upward biased.

In the model in (29) - (31) we continue to assume $\varepsilon_{it} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ and now make the additional assumption

$$\left[\begin{array}{c} u_i \\ v_i \end{array} \right] \overset{iid}{\sim} N\left(0_2, \Sigma\right) \text{ where } \Sigma \equiv \left[\begin{array}{cc} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{array} \right].$$

[11] We consider this particular model specification as we will treat the schooling variable s_i as endogenous in our application. Of course, the algorithm described in the appendix can be easily adapted to treat an element of w_i as endogenous, or to treat an element of z_{it} as endogenous.

To complete this model, we add the following priors:

$$\begin{aligned}\Sigma^{-1} &\sim W\left([\underline{\rho}\underline{A}]^{-1}, \underline{\rho}\right) &(32)\\ \pi &\sim N(\underline{\pi}, \underline{V}_{\pi}) &(33)\\ \beta &\sim N(\underline{\beta}, \underline{V}_{\beta}), &(34)\end{aligned}$$

where $W(\cdot,\cdot)$ denotes the Wishart density (e.g., Poirier 136-138), the typical conjugate prior for the inverse covariance matrix. We retain the same priors for δ, and $\sigma_{\varepsilon}^{-2}$, as given in (18) and (19).

2.3.1 Estimation and Testing in the Hierarchical Model with Endogeneity

Since estimation of models like this one, with a simultaneous equations format involving the individual effects, has not received much attention in the literature, we spend a bit of time to derive the associated likelihood function and joint posterior distribution. We defer specific details regarding the posterior simulator to the appendix.

By Bayes Theorem, the joint posterior distribution is proportional to the product of the prior times the likelihood:

$$p(\alpha, \delta, \beta, \pi, \sigma_{\varepsilon}^{-2}, \Sigma^{-1}|y, s) \propto p(y, s|\alpha, \delta, \beta, \pi, \sigma_{\varepsilon}^{2}, \Sigma^{-1})p(\alpha, \delta, \beta, \pi, \sigma_{\varepsilon}^{-2}, \Sigma^{-1}), \qquad (35)$$

where the likelihood function is the joint density of y and the endogenous schooling variable s conditioned on the model parameters. As for the prior specification given on the right hand side of (35), our specifications in (29)-(34) imply:

$$\begin{aligned}p(\alpha, \delta, \beta, \pi, \sigma_{\varepsilon}^{-2}, \Sigma^{-1}) &= p(\alpha|\beta, \pi, \Sigma^{-1})p(\delta)p(\beta)p(\pi)p(\sigma_{\varepsilon}^{-2})p(\Sigma^{-1})\\ &= \left[\prod_{i=1}^{n} p(\alpha_i|\beta, \pi, \Sigma^{-1})\right] p(\delta)p(\beta)p(\pi)p(\sigma_{\varepsilon}^{-2})p(\Sigma^{-1}),\end{aligned}$$

where the last line follows from the assumed conditional independence across observations. The density $p(\alpha_i|\beta, \pi, \Sigma^{-1})$ can be obtained by substituting out s_i from (30) using the reduced form equation in (31).

As for the likelihood function in (35), first let Γ denote all the parameters in the model and note

$$\begin{aligned}p(y, s|\Gamma) &= \prod_{i=1}^{N} p(y_{i1}, \cdots, y_{iT_i}, s_i|\Gamma) &(36)\\ &= \prod_{i=1}^{N} \left(\prod_{t=1}^{T_i} p(y_{it}|\alpha_i, \delta, \sigma_{\varepsilon}^{-2})\right) p(s_i|\alpha_i, \pi, \Sigma^{-1}, \beta), &(37)\end{aligned}$$

where the first line follows by the assumed independence across individuals, and the last equation follows by noting that the density for y_{it} does not depend on s_i given α_i and that y_{it} is assumed to be independent over time given α_i. When appropriate, we have also dropped irrelevant parameters from the conditioning.

Substituting the likelihood and prior into the expression in (35) we obtain the unnormalized joint posterior

$$p(\Gamma|y,s) \propto \left[\prod_{i=1}^{N}\left(\prod_{t=1}^{T_i} p(y_{it}|\alpha_i,\delta,\sigma_\varepsilon^2)\right) p(s_i,\alpha_i|\pi,\Sigma^{-1},\beta)\right] p(\delta)p(\beta)p(\pi)p(\sigma_\varepsilon^2)p(\Sigma^{-1}). \tag{38}$$

Despite the fact that this model handles both nonparametric estimation and endogeneity concerns simultaneously, estimation is relatively-straight forward as all of the complete posterior conditionals derived from (38) can be easily sampled, as described in detail in the appendix. Finally, testing and empirical Bayesian selection of smoothing parameters can be conducted using the same procedure described in Section 2.2 using the Laplace-Metropolis approximation.

3 The Data

The data used in our empirical work are taken from the National Longitudinal Survey of Youth (NLSY). The NLSY is a panel data set providing a wealth of information on the earnings experience, educational histories, and family backgrounds of a sample of young men and women in the U.S. The NLSY panel used in this analysis begins in 1979, conducts annual interviews until 1994, and then reports information from biannual surveys until 2000.

As discussed throughout this paper, we are primarily interested in flexibly exploring the relationship between ability and log wages and the manner in which returns to education vary with ability. Fortunately for this purpose, the NLSY reports scores on the 10 tests comprising the Armed Services Vocational Aptitude Battery (ASVAB), which is administered to the NLSY participants in 1980 and serves as a reasonable instrument for cognitive ability. To reduce the dimensionality of these ability measures and fix our attention on a scalar "ability" variable, we follow Cawley, Conneely, Heckman and Vytlacil (1997) and purge the 10 test scores of a linear age effect and then use the first principal component of the resulting ten vectors of (standardized) residuals as our ability measure.[12]

Our time-varying characteristics consist of quadratics in total weeks of actual labor market experience (denoted EXP and EXP^2) and tenure on the current job (TENURE and TENURE^2), an indicator for residence in an urban area (URBAN) and a time trend (TREND). The actual weeks of labor market experience variable is constructed by aggregating reported weeks of work between interview dates. The dependent variable employed in the analysis is the log hourly wage. Our time-invariant characteristics consist of our ability measure (ABILITY), highest grade completed by the respondent (EDUC), highest grade completed by the respondent's mother (MOMED) and father (DADED) and number of siblings (NUMSIBS).

[12] Since students varied in age at the time the tests were administered, we regressed each test score on age and then obtained and standardized the 10 residual vectors from these regressions. The eigenvector corresponding to the largest eigenvalue of the residual correlation matrix serves as the weighting vector, and the ability measure we use is the product of this weighting vector times the standardized residual scores. This resulting ability measure is then standardized to have mean zero and unit variance for interpretation purposes. Finally, the 10 component tests of the ASVAB battery are general science, arithmetic reasoning, word knowledge, paragraph comprehension, coding speed, numerical operations, auto and shop information, mathematics knowledge, mechanical comprehension and electronics information.

In keeping with the majority of this literature, we restrict our attention to the analysis of white males in the NLSY, and specifically, we focus only on those white males from the cross-sectional samples. We exclude observations when the hourly wage (in real 2000 dollars) is less than \$1 or greater than \$50, when the respondent reports to be currently enrolled in high school or college in the given year and when the quantity of schooling completed varies over time even after conditioning on those not enrolled in school. We delete observations when the number of weeks worked since the last interview is less than the reported increase in tenure with the given employer between interview dates, and when parental or the individual's own education is less than 9. We also require that each individual is observed for at least 5 years throughout the sample period. This sample selection procedure yields a total of $3,980$ observations from 359 individuals. Thus, on average, each individual is observed for approximately ten years of the panel, and for some individuals, we have as many as eighteen observations.

When dealing with endogeneity concerns, we require an instrument. This instrument must affect the quantity of schooling attained by the individual, but not be correlated with the person-specific random effect given the other controls we employ. Our choice in this regard is to use the quantity of schooling obtained by the respondent's oldest sibling (SIBED). Our argument for the use of this instrument is that sibling's education should be strongly correlated with one's own education, as it proxies both familial preferences toward the importance of education, and potentially, resources constraints faced by the family. However, siblings education itself should play no structural role in the wage equation, *conditioned on one's own schooling and our other controls for family background.* When estimating the model that controls for endogeneity, we impose further restrictions in our sample. Specifically, we now consider only those individuals in the NLSY with a sibling, and also require that the oldest sibling be at least 24 years of age in 1979[13] so that he/she is likely to have completed his/her schooling. This sample selection scheme produced a total of 1,203 observations from 98 individuals. Finally, in the model with endogenous schooling, we include ABILITY, MOMED, DADED, NUMSIBS and SIBED in the reduced form schooling equation.

4 Empirical Results

In this section we illustrate how our methods can be applied in practice using both generated data and data from the NLSY. We recognize that the NLSY data is longitudinal in nature and thus use it to obtain results for the longitudinal smooth coefficient models with and without endogeneity, as described in sections 2.2 and 2.3. For the cross-sectional smooth coefficient model, we choose to perform generated data experiments to illustrate estimation, testing and smoothing parameter selection when the data generating process is known.

[13] The question is asked to the NLSY respondents in 1979, the base year of the survey.

4.1 Cross-Sectional Model

We perform two generated data experiments, the first of which generates artificial data from a highly nonlinear model while the second generates data from a linear model. For the nonlinear data set, we extend the data generating mechanism used in Yatchew (1998, Figure 3). In particular, for $i = 1, .., 200$ we generate

$$y_i = 2w_i + A_i \cos(4\pi A_i) + \sin(2\pi A_i)\, s_i + \varepsilon_i, \tag{39}$$

where ε_i, A_i and w_i i.i.d. random variables drawn from a $N(0, 1)$ distribution and $s_i \overset{iid}{\sim} U[0, 10]$, with U denoting the uniform distribution.

In addition to carrying out posterior inference on all the parameters in the smooth coefficient model, we calculate the Bayes factor comparing the smooth coefficient model to a particular parametric model where all explanatory variables enter linearly:

$$y_i = w_i\theta + \lambda_0 + \lambda_1 A_i + \lambda_2 s_i + \varepsilon_i. \tag{40}$$

Our general prior elicitation strategy is to use the smoothness prior for the nonparametric regression lines (with smoothness parameters η_1 and η_2 chosen using empirical Bayesian methods), to select proper but relatively noninformative priors for parameters which are present in one model but not the other, and to choose noninformative (or virtually noninformative) priors for all parameters which are common to both models. In terms of the NG$(\underline{\beta}, \underline{V}_\beta, \underline{s}^{-2}, \underline{\nu})$ natural conjugate prior described in Section 2.1 we set $\underline{\nu} = 0$ (and, with this value, $\underline{s}^{-2}$ is irrelevant). For θ (which is common to both models) we use a virtually noninformative prior by setting the appropriate diagonal elements of $\underline{V}_1$ to 10^{20} (see (8)). For the initial conditions (which appear only in the smooth coefficient model), we set the appropriate diagonal elements of $\underline{V}_1$ to 100. For the parametric model, the prior mean of $\lambda = (\lambda_0, \lambda_1, \lambda_2)$ is set to zero and the prior covariance matrix to $100\sigma_\varepsilon^2 I_3$. We use empirical Bayesian methods to select η_1 and η_2 which yields $\eta_1 = 2 \times 10^{-5}$ and $\eta_2 = 3 \times 10^{-5}$.

Empirical results from this first generated data experiment are sensible. The log of the Bayes factor in favor of the smooth coefficient model is 221.79, indicating strong support for the smooth coefficient model over the linear parametric model (which is far from the data generating process). In Figures 1 and 2 we present posterior means (which can be obtained *analytically*) of the two nonparametric regression lines (solid) as well as the true relationships (dashed) that were used to generate the data. It can be seen that the fitted nonparametric regression lines track the true lines quite well, despite the fact that our data set is fairly small and we have included a fairly large random error component. To see the latter, note that Figure 1 also plots the "Data" defined as $y_i - w_i E(\theta|Data) - E(\gamma_{2i}|Data)\, s_i$. Despite a large scattering of these "data" points, the smooth coefficient model picks up the main pattern very effectively.

We also generated a second cross-sectional data set that used (40) as the data generating process. To be precise, this second data set is exactly equal to our first one, except the terms $A_i \cos(4\pi A_i)$ and $\sin(2\pi A_i)\, s_i$ are deleted. All other modeling details, including the prior, are as described above. For the sake of brevity, we do not present detailed results from this data set. Suffice it to note here that

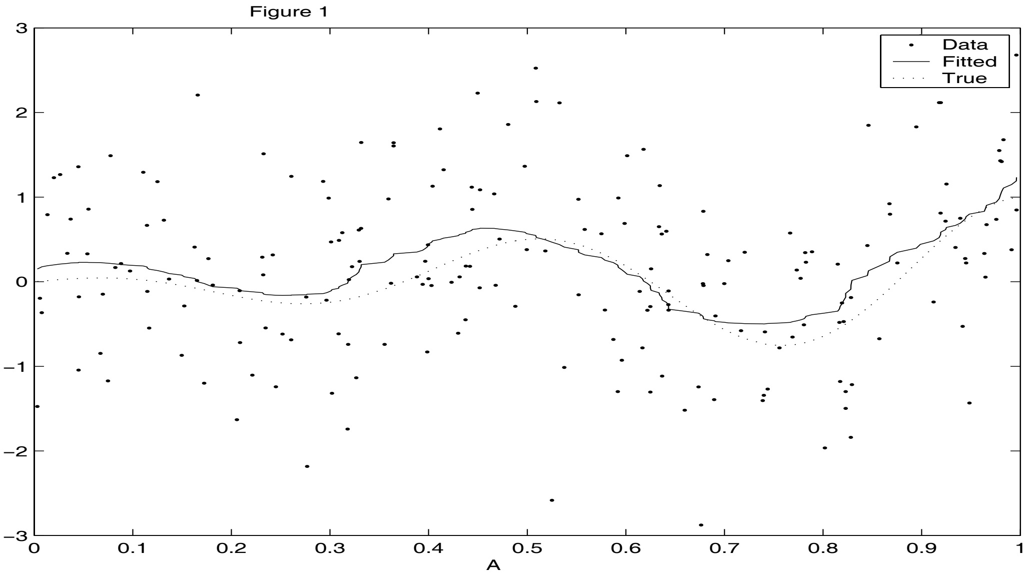

Figure 1: Fitted and True Regression lines for $f_1(A) = A\cos(4\pi A)$.

the log of the Bayes factor in favor of the smooth coefficient model over the linear model given in (40) is -12.91, indicating support for the (true) linear model. It is also worth mentioning that even in this case the smooth coefficient model does a good job of fitting the data. That is, the empirical Bayesian methods select very small values for η_1 and η_2 and the resulting fitted nonparametric lines are very close to simply being horizontal lines at zero (as they should be). Of course, if one does have a parsimonious parametric model that is known to fit the data well, then the use of nonparametric methods may be superfluous. However, it is reassuring to see that our nonparametric methods work well in both designs, and our flexible methods have the advantage that they are *adaptable* in applied situations where the design is unknown.

4.2 Longitudinal Model without Endogeneity

In this section we turn to our application and present empirical results using the NLSY data described in Section 3 and the model discussed in Section 2.2. The explanatory variables in the first (i.e. the elements of z_{it} in (16)) and second (i.e. the elements of w_i following (17)) stages of the hierarchy are discussed in Section 3. We are particularly in interested in investigating the following questions: (1) What is the relationship between expected log wages and our measure of cognitive ability? (2) Do returns to schooling vary with ability? and (3) Are standard parametric models adequate for describing the relationships in the NLSY data?[14]

To investigate the last of these questions we calculate the Bayes factor comparing the smooth coefficient

[14] Given our focus on the hierarchical smooth coefficient model and its potential applicability, we do not address the issue of time-varying returns to ability and/or education over this period. See Blackburn and Neumark (1993), Heckman and Vytlacil (2001) and Taber (2001) for more on this particular issue.

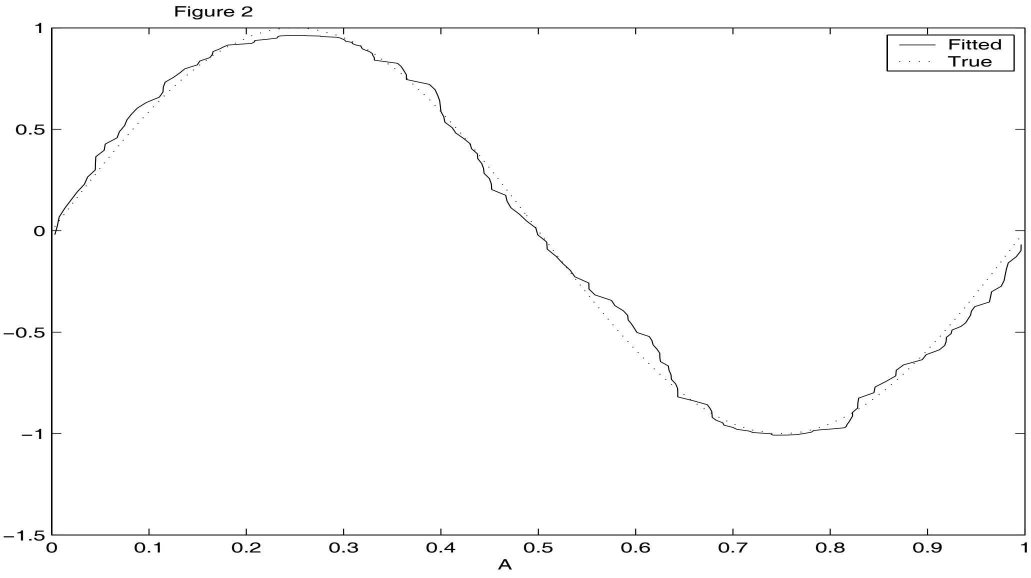

Figure 2: Fitted and True Regression Lines for Smooth Coefficient Term $f_2(A) = \sin(2\pi A)$.

model to a parametric model where all explanatory variables enter linearly. That is, the first and second stages of the hierarchy for the parametric model are given by (16) and (17) but the mean in the second stage is now assumed to be *linear* in A and S:[15]

$$\mu_i = w_i\theta + \lambda_0 + \lambda_1 A_i + \lambda_2 s_i. \tag{41}$$

Our general prior elicitation strategy remains the same as that described in section 4.1. That is, we use empirical Bayesian methods to elicit the smoothness prior for the nonparametric regression lines, we select proper but relatively noninformative priors for parameters which are present in one model but not the other and we choose noninformative (or virtually noninformative) priors for all parameters which are common to all models. Accordingly, using the notation of (18) - (20) we set $\underline{\delta} = 0$, $\underline{V}_{\delta}^{-1} = 0$, $\underline{\nu}_{\varepsilon} = 0$ (and with this value, $\underline{s}_{\varepsilon}^{-2}$ is irrelevant), $\underline{s}_{\alpha}^{-2} = 1$ and $\underline{\nu}_{\alpha} = .01$ for both models. The structure of $\underline{\beta}$ and $\underline{V}_{\beta}$ for the parametric and nonparametric models is exactly as in the cross-sectional empirical illustration (see Section 4.1).

Posterior results are produced using the MCMC algorithm described in (22) through (26).[16] A two dimensional grid search over values for η_1 and η_2 indicates support for small values of the smoothing parameters, and specifically, our empirical Bayesian strategy yields $\eta_1 = \eta_2 = 10^{-12}$ as the optimal choice. This indicates that the nonparametric regression lines f_1 and f_2 are very smooth and can be taken as informal evidence that a linear model is supported by the data. A more formal test in this regard is provided by the Bayes factor comparing the smooth coefficient model to the parametric model with the mean of the second stage of the hierarchy given by (41). The log of this Bayes factor is -948.6, indicating *strong* evidence in favor of the parametric model. The reason for this is clear.

[15] An alternative specification would include an interaction between A_i and s_i. We do not include this specification since it receives little support from the data. However, it is worth stressing that this and other similar extensions are trivial to handle in our framework.

[16] We run all MCMC algorithms for $11,000$ replications and discard the initial $1,000$ as burn-in replications. Our results pass standard convergence diagnostics.

The maximum value of the log likelihoods for these two models are the same (to one decimal place), indicating the two models fit the data roughly equally. However, Bayes factors also have a reward for parsimony built in, and this strongly favors the more parsimonious parametric model (note that the smooth coefficient model has almost $2N$ more coefficients than the parametric model). Larger values of the smoothing parameters yield a slight increase in the maximum of the likelihood function, but this is more than counterbalanced by the increasing penalty associated with added parameterization (i.e. larger values of η_1 and η_2 imply that the model can fit a wider range of data sets and, thus, is less parsimonious). This is a general property of our approach and, we feel, a sensible one. Nonparametric models are non-parsimonious, so receive little support unless the parametric alternative fits the data very poorly.

For this application we do not find that parametric models fit the data poorly, and specifically, we find that simpler parametric specifications provide an adequate description of the ability and ability-schooling relationships in the NLSY data.[17] Of course, the power and lure of nonparametric methods is their ability to adapt to capture the shape of unknown regression functions. The fact that these functions can be recovered by simpler methods for this application does not invalidate the use of the smooth coefficient model; we obtain similar empirical results from both approaches, and our ability to declare the parametric specification as "satisfactory" depends on the availability of more flexible modeling alternatives.

Table 1: Posterior Results for First and Second Stage Parameters:
Hierarchical Model Without Endogeneity

		Smooth Coefficient Model		Parametric Model	
	Variable/	Post.	Post.	Post.	Post.
	Parameter	Mean	St. Dev.	Mean	St. Dev.
First Stage	EXP	1.24×10^{-3}	1.30×10^{-4}	1.26×10^{-3}	1.30×10^{-4}
	EXP^2	-4.12×10^{-7}	6.72×10^{-8}	-4.14×10^{-7}	6.81×10^{-8}
	TENURE	6.67×10^{-4}	8.15×10^{-5}	6.70×10^{-4}	8.04×10^{-5}
	TENURE^2	-6.82×10^{-7}	1.03×10^{-7}	-6.87×10^{-7}	1.02×10^{-7}
	URBAN	0.079	0.017	0.077	0.017
	TREND	-6.68×10^{-3}	4.68×10^{-3}	-7.52×10^{-3}	4.66×10^{-3}
	σ_ε^2	0.090	0.003	0.090	0.003
Second Stage	MOMED	-7.17×10^{-3}	1.17×10^{-2}	-6.11×10^{-3}	1.17×10^{-2}
	DADED	0.021	0.009	0.022	0.009
	NUMSIBS	0.020	0.010	0.020	0.010
	INTERCEPT	$--$	$--$	0.904	0.155
	ABILITY	$--$	$--$	0.041	0.020
	EDUC	$--$	$--$	0.064	8.73×10^{-3}
	σ_α^2	0.101	0.008	0.103	0.009

[17] The graph in Figure 3 does suggest some slight nonlinearities in the ability log-wage relationship. We thus added a quadratic ability term to our fully parametric model and found evidence that the quadratic term was "significant," as the quadratic coefficient had a high posterior probability of being positive. This finding of some nonlinearities is consistent with the results in Cawley, Heckman and Vytlacil (1999) and Tobias (2003). In formal testing, both the linear and quadratic models are strongly preferred over our smooth coefficient model given the high degree of parameterization in the smooth coefficient model.

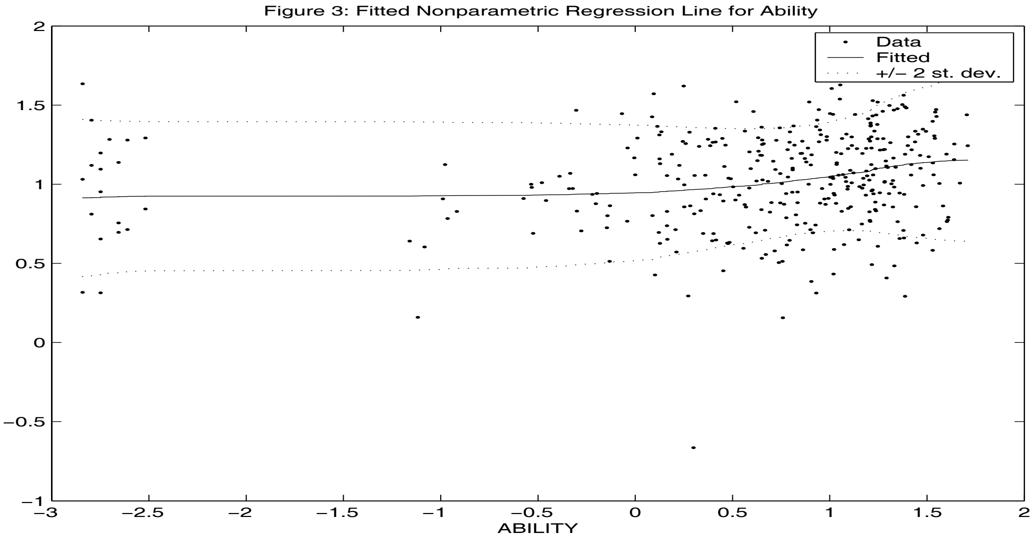

Figure 3: Fitted Regression Line for Ability Term $f_1(A)$ in Hierarchical Model

To show the similarity between smooth coefficient results and those from a linear model, Table 1 presents coefficient posterior means and standard deviations from the smooth coefficient and parametric models. For all parameters which are common to both models, it can be seen that the posteriors are virtually identical to one another. In Figure 3 and Figure 4 we present evidence relating to the parameters which are not common to both models and plot estimates of $f_1(A)$ and $f_2(A)$ from the smooth coefficient model. Figure 3 plots the posterior mean of $f_1(A)$ (solid) and $E(f_1(A)|\text{Data})\pm 2\text{Std}(f_1(A)|\text{Data})$ (dashed), while Figure 4 similarly plots the posterior mean and posterior standard error bands associated with f_2.

These figures again reveal that the nonparametric and linear models are basically telling the same story. We see some slight nonlinearities in Figure 3, though Figure 3 is not far different from a linear model and results obtained from linear specification fall comfortably within the plotted standard error bands. From Table 1, the parametric model says that an added year of schooling increases hourly wages by about 6.5 percent with the ±2 Std. interval given by $[0.049, 0.081]$. By construction, the parametric model imposes the same returns to schooling on individuals with different level of ability. Though our flexible smooth coefficient model has the potential of allowing returns to schooling to vary across individuals of differing ability, it estimates the return to education as being roughly constant across individuals of varying ability.[18] The smooth coefficient model yields a slightly lower point estimate (roughly six percent), but relative to the size of posterior standard deviations, this difference is negligible. The posterior standard deviations for returns to schooling in the smooth coefficient model are also slightly larger than in the parametric model, which is to be expected given our agnostic stance regarding the specification of the model. Despite these small differences, the parametric and nonparametric models yield roughly the same posterior inferences for returns to schooling and measured cognitive ability.

[18] In a related analysis that did not make use of the nonparametric techniques described here and required time-varying schooling, Koop and Tobias (2002) found little evidence that measured ability played a significant role in explaining variation in returns to schooling across individuals.

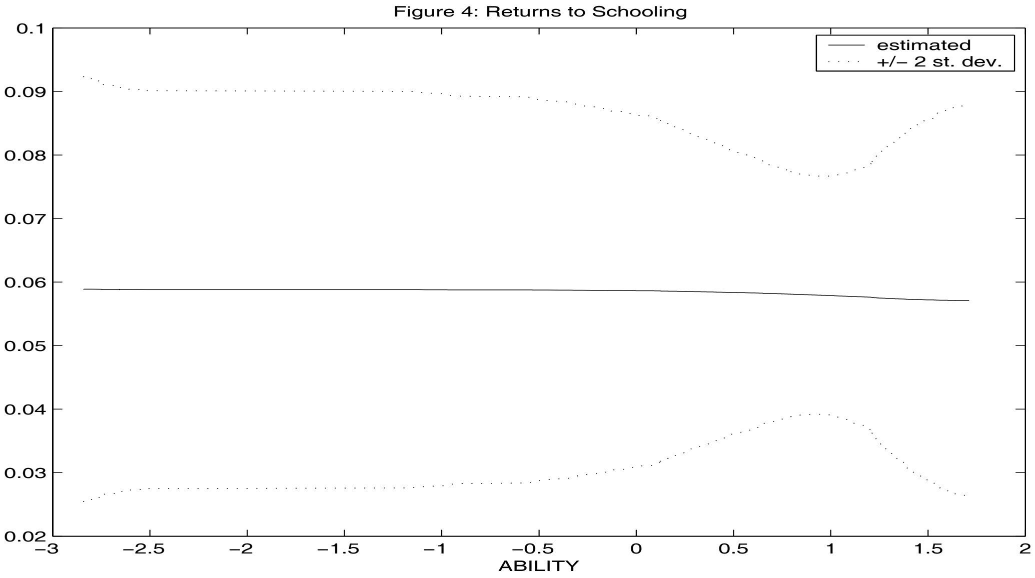

Figure 4: Fitted Regression Line for Return to Schooling Term $f_2(A)$ in Hierarchical Model.

4.3 Longitudinal Model with Endogeneity

The longitudinal model with endogeneity is given in (29) through (31). Equations (29) and (30) retain the same specifications that were used in section 4.2 with no endogeneity concerns. Equation (31) includes an intercept, MOMED, DADED, NUMSIBS, ABILITY and SIBED as explanatory variables, where sibling's education (SIBED) is the instrument used to identify the model. As in the previous section, we carry along a fully parametric specification and compare results from that specification to those obtained from our smooth coefficient analysis. The first two equations of this parametric model are the same as were used previously in the model without endogeneity (see (41)). The third equation (equation 31) is the same in the parametric and smooth coefficient models.

For parameters that are common to both the models in section 4.2 and 4.3, we retain the same prior specifications that were used in section 4.2. We use a noninformative prior for the coefficients in reduced form schooling equation (i.e.,. we set $\underline{V}_{\pi}^{-1} = 0$ and, with this value, $\underline{\pi}$ is irrelevant). For the prior in (32) we use relatively noninformative values[19] of $\underline{\rho} = 9$ and $\underline{A} = I_2$.

We again find little support for the flexibility afforded by the smooth coefficient model for this application, since empirical Bayesian estimation selects very small values for the smoothing parameters, $\eta_1 = \eta_2 = 10^{-12}$. Like the previous section, this result suggests that our semiparametric estimates will be sufficiently smooth so that simpler parametric models will do an adequate job at reproducing the shapes of the regression curves. We do not view this preference for the simpler model as problematic in any way, but instead are pleased with the fact that our smooth coefficient analysis successfully recreates the findings of a well-fitting parametric model without requiring the assumption of a particular

[19] The value for $\underline{\rho}$ is the smallest which guarantees that prior means, variances and covariances exist. See Poirier (1995, page 138) for related discussion.

parametric form.

Table 2: Posterior Results for First and Second Stage Parameters $(\delta, \sigma_\varepsilon^2)$

		Smooth Coefficient Model		Parametric Model	
	Variable/ Parameter	Post. Mean	Post. St. Dev.	Post. Mean	Post. St. Dev.
First Stage	EXP	1.91×10^{-3}	2.94×10^{-4}	1.84×10^{-3}	2.92×10^{-4}
	EXP^2	-1.11×10^{-7}	1.71×10^{-7}	-1.07×10^{-6}	1.74×10^{-7}
	TENURE	7.39×10^{-4}	1.70×10^{-4}	7.64×10^{-4}	1.71×10^{-4}
	TENURE^2	-8.67×10^{-7}	2.24×10^{-7}	-8.68×10^{-7}	2.28×10^{-7}
	URBAN	0.280	0.049	0.274	0.050
	TREND	6.71×10^{-3}	1.13×10^{-2}	6.97×10^{-3}	1.14×10^{-3}
	σ_ε^2	0.126	0.014	0.127	0.014
Second Stage	First Equation (Individual effect is dependent variable)				
	MOMED	−0.017	0.029	−0.012	0.029
	DADED	0.019	0.025	-4.35×10^{-3}	0.024
	NUMSIBS	−0.028	0.026	−0.025	0.026
	INTERCEPT	−−	−−	1.014	0.439
	ABILITY	−−	−−	-9.22×10^{-3}	0.045
	EDUC	−−	−−	0.062	0.028
	Second Equation (Schooling is dependent variable)				
	INTERCEPT	6.404	0.392	6.458	0.396
	MOMED	0.029	0.031	0.048	0.030
	DADED	0.337	0.028	0.317	0.025
	NUMSIBS	−0.075	0.030	−0.056	0.027
	ABILITY	0.273	0.049	0.245	0.049
	SIBED	0.179	0.021	0.170	0.020
	ρ_{uv}	−0.237	0.398	−0.163	0.489

We present coefficient posterior means and standard deviations from both the parametric and smooth coefficient models in Table 2. For parameters which are common to both models, Table 2 again shows that the posteriors are very similar. It is also worth mentioning that the coefficient on our instrument SIBED is reasonably large in magnitude and clearly "significant" in both the parametric and smooth coefficient models, suggesting it plays an important role in determining the quantity of schooling attained. For those quantities which are not common to both specifications (namely the specification of ability and the ability-education interaction in the individual effect equation (30)) we plot in Figures 5 and 6 our estimates of $f_1(A)$ and $f_2(A)$ from the smooth coefficient model.

The general shapes in Figures 5 and 6 differ from those reported in the model without endogeneity in Figures 3 and 4. Specifically, Figure 6 suggests that returns to schooling are increasing with ability and very low ability individuals receive virtually no returns to education. The downward slope of Figure 5 also appears to offset the upward slope of Figure 6, so that the overall shape of the relationship between ability and log wages remains a bit unclear. To get a better sense of how log wages change with changes in ability, we calculated the function $f_1(A) + f_2(A)Ed$ for $Ed = 12$ and $Ed = 16$.[20] These

[20] We do not present these figures here, though they are available upon request.

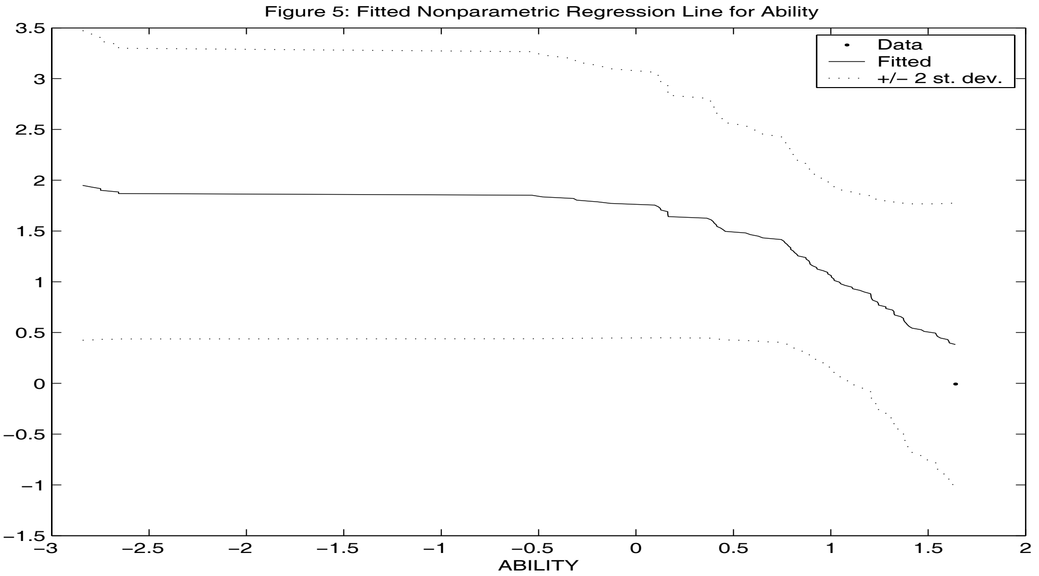

Figure 5: Fitted Regression Line for Ability Term $f_1(A)$ in Hierarchical Model with Endogeneity

calculations revealed that log wages are clearly increasing in ability for those with a college education ($Ed = 16$), but the ability-log wage relationship for those with just a high school degree was rather flat with a small downturn at the far right-tail of the ability distribution.

Before we make too much of these differences, it is important to note that the disparity in results obtained from sections 4.2 and 4.3 can arise from two sources. First, the endogeneity of schooling could be an empirically important issue, and when formally controlling for endogeneity we simply obtain different results. Second, the analysis of section 4.3 is based on a different and strictly smaller data set owing to the necessary reduction in observations due to the use of SIBED as our instrument. To distinguish between these two stories, we first note that the results of Table 2 do not suggest strong evidence that endogeneity of schooling is an empirically important issue. The posterior mean of the correlation coefficient between the errors in the two second stage equations - the source of the endogeneity problem - is slightly negative, though its posterior standard deviation is huge. This suggests that with this relatively small sample size[21] we are not able to pin down this key correlation parameter, and thus learn little regarding the empirical importance of endogeneity.

To investigate this issue a bit further, we took this smaller data set and re-estimated the model in section 4.2 which does not control for endogeneity.[22] Results from this analysis were similar to those reported in this section - returns to schooling were increasing in ability and the ability-log wage relationship in f_1 was essentially flat and then dipped down around $A = 0$. Since analysis of this restricted data set produces similar results when controlling for endogeneity or when abstracting from the endogeneity problem, it seems that differences between Figures 3 and 4 and Figures 5 and 6 arise from the use of different data sets. Given no clear evidence regarding the empirical importance of the endogeneity of schooling in this application, we fall back on the the simpler specification in section 4.2

[21] Recall that we have only 98 individuals to estimate the bivariate relationship in (30) and (31).

[22] Of course, equation (31) is then omitted from the analysis.

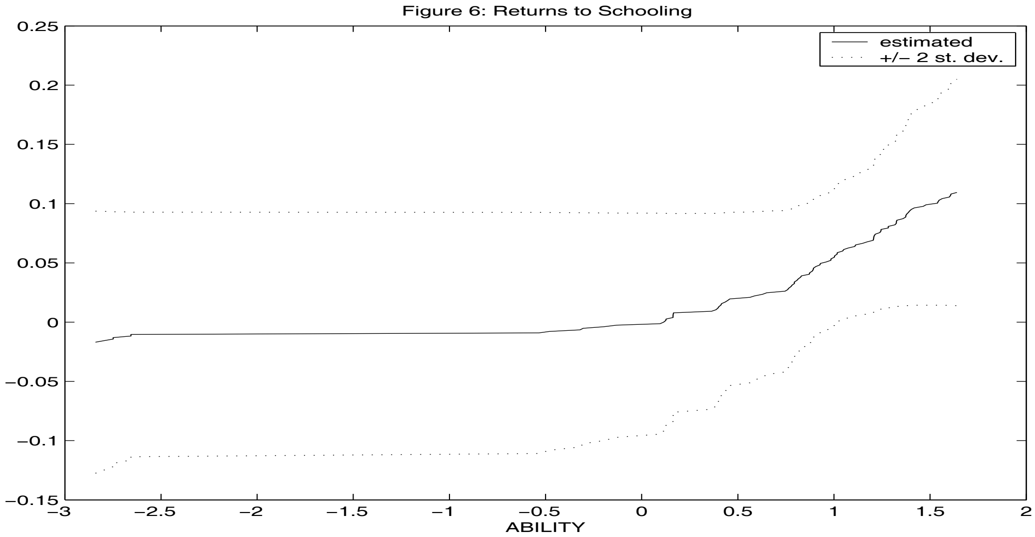

Figure 6: Fitted Regression Line for Ability Term $f_2(A)$ in Hierarchical Model with Endogeneity

which abstracts from the endogeneity problem and provides us with more observations to estimate the bivariate relationship in (30) and (31). Finally, it is important to recognize that the model described in this section and section 2.3 is quite operational and can be used in other situations where researchers want both flexibility in the specification of the conditional mean function and the ability to control for endogeneity problems.

5 Conclusion

In this paper we have described Bayesian procedures for estimation and testing in cross sectional and longitudinal data smooth coefficient models. In the cross sectional model, estimation, testing and smoothing parameter selection can be carried out *analytically*, thus making analysis of the smooth coefficient model a simple yet flexible option for practitioners. In the hierarchical smooth coefficient model and the hierarchical model with endogeneity concerns, estimation only requires iterative simulation from standard distributions.

We illustrated the flexibility and practicality of our methods in generated data experiments and in an application using data from the National Longitudinal Survey of Youth (NLSY). Using the NLSY we investigated the issue of nonlinearities in the relationship between log wages and measured cognitive ability and also flexibly modeled the dependence of returns to education on this ability measure. Our results suggested that returns to education were roughly constant throughout the ability support and that simpler (and often used) parametric specifications provide an adequate description of these relationships.

5.1 Posterior Simulator: Longitudinal Model with Endogeneity

The model is given as:

$$y_{it} = \alpha_i + z_{it}\delta + \varepsilon_{it} \tag{42}$$

$$\alpha_i = w_i\theta + f_1(A_i) + s_i f_2(A_i) + u_i \tag{43}$$

$$s_i = r_i\pi + v_i, \tag{44}$$

where again, we let

$$\mu_i = w_i\theta + f_1(A_i) + s_i f_2(A_i) = x_i\beta,$$

as described in section 2.1. The likelihood function, joint posterior distribution and priors used are presented in Section 2.3. As a notational convention, we use Γ to denote all the parameters in the model and, e.g., $\Gamma_{-\delta}$ to denote all the parameters except for δ.

To fit this model we again employ a blocking step to sample first from the conditional for δ marginalized over the random effects, and then to sample from the complete conditionals for the random effects. We obtain:

$$\delta | Data, \Gamma_{-\alpha,\delta} \sim N(D_\delta d_\delta, D_\delta), \tag{45}$$

where

$$D_\delta = \left(\sum_i Z_i' \Omega_i^{-1} Z_i + \underline{V}_\delta^{-1} \right)^{-1},$$

$$d_\delta = \sum_i Z_i' \Omega_i^{-1} \left(y_i - \iota_{T_i} \left[\mu_i + \frac{\sigma_{uv}}{\sigma_v^2}(s_i - r_i\pi) \right] \right) + \underline{V}_\delta^{-1} \underline{\delta}$$

and

$$\Omega_i = \sigma_u^2(1-\rho_{uv}^2)\iota_{T_i}\iota_{T_i}' + \sigma_\varepsilon^2 I_{T_i}.$$

We also obtain the complete conditional for the random effects:

$$\alpha_i | \Gamma_{-\alpha_i}, \text{Data} \overset{ind}{\sim} N\left(D_{\alpha_i} d_{\alpha_i} D_{\alpha_i}\right), \quad i = 1, 2, \cdots N, \tag{46}$$

where

$$D_{\alpha_i} = \left[T_i/\sigma_\varepsilon^2 + \sigma_u^{-2}(1-\rho_{uv}^2)^{-1} \right]^{-1}$$

and

$$d_{\alpha_i} = \left[\sum_y (y_{it} - z_{it}\delta)/\sigma_\varepsilon^2 \right] + \sigma_u^{-2}(1-\rho_{uv}^2)^{-1} \left(\mu_i + \frac{\sigma_{uv}}{\sigma_v^2}(s_i - r_i\pi) \right).$$

In equations (45) and (46), ι_x denotes an $x \times 1$ vector of ones, Z_i and y_i have been stacked over t within i conformably, ρ_{uv} denotes the correlation between u and v, and x_i and β are defined as in section 2.1. These conditionals are derived from (38) after factoring the distribution for $(s_i \ \alpha_i)$ into the conditional for α_i given s_i times the marginal for s_i and applying the result of Lindley and Smith (1972).

We obtain the following posterior conditional for σ_ε^{-2}:

$$\sigma_\varepsilon^{-2} | Data, \alpha, \delta \sim G\left(\overline{s}_\varepsilon^{-2}, \overline{\nu}_\varepsilon\right) \tag{47}$$

where

$$\overline{\nu}_{\varepsilon} = \sum_i T_i + \underline{\nu}_{\varepsilon},$$

$$\overline{s}^2_{\varepsilon} = \frac{\sum_i (y_i - \alpha_i - Z_i\delta)'(y_i - \alpha_i - Z_i\delta) + \underline{\nu}_{\varepsilon}\underline{s}^2_{\varepsilon}}{\overline{\nu}_{\varepsilon}}.$$

As for the complete conditionals for the second-stage regression parameters and inverse covariance matrix, let us first stack the triangular system in (43) and (44) together and introduce some new notation. We stack the time-invariant components together as follows:

$$\begin{bmatrix} \alpha \\ s \end{bmatrix} = \begin{bmatrix} X & 0 \\ 0 & R \end{bmatrix} \begin{bmatrix} \beta \\ \pi \end{bmatrix} + \begin{bmatrix} u \\ v \end{bmatrix},$$

or equivalently

$$\tilde{\alpha} = \tilde{X}\tilde{\beta} + \tilde{u},$$

where $\tilde{\alpha} = [\alpha' \ S']'$, $\tilde{\beta} = [\beta' \ \pi']'$, $\tilde{u} = [u' \ v']'$ and $\tilde{X}$ is defined as the blocked diagonal design matrix with X and R on the diagonals. We also note that $E(\tilde{u}\tilde{u}|\tilde{X}) = \Sigma \otimes I_n$. Given this, it follows that

$$\tilde{\beta}|\Gamma_{-\tilde{\beta}}, \text{Data} \sim N\left(D_{\tilde{\beta}} d_{\tilde{\beta}}, D_{\tilde{\beta}}\right), \tag{48}$$

where

$$D_{\tilde{\beta}} = \left(\tilde{X}'(\Sigma^{-1} \otimes I_N)\tilde{X} + \underline{\tilde{V}}_{\tilde{\beta}}^{-1}\right)^{-1},$$

$$d_{\tilde{\beta}} = \tilde{X}'(\Sigma^{-1} \otimes I_N)\tilde{\alpha} + \underline{\tilde{V}}_{\tilde{\beta}}^{-1}\underline{\widetilde{\beta}},$$

$$\underline{\tilde{V}}_{\tilde{\beta}} = \begin{bmatrix} \underline{V}_{\beta} & 0 \\ 0 & \underline{V}_{\pi} \end{bmatrix} \quad \text{and} \quad \underline{\widetilde{\beta}} = \begin{bmatrix} \underline{\beta} \\ \underline{\pi} \end{bmatrix}.$$

Finally, consider the complete posterior conditional for Σ^{-1} and let

$$\psi_i = \psi_i(\alpha_i, \beta, \pi) = \begin{bmatrix} \alpha_i - \mu_i \\ s_i - r_i\pi \end{bmatrix} = \begin{bmatrix} u_i \\ v_i \end{bmatrix}.$$

Thus, conditioned on β and π, the second stage errors are effectively "known." Given this, we obtain the following posterior conditional

$$\Sigma^{-1}|\Gamma_{-\Sigma^{-1}}, \text{Data} \sim W\left[\left(\sum_{i=1}^{N} \psi_i\psi_i' + \underline{\rho}\underline{A}\right)^{-1}, \underline{\rho} + N\right]. \tag{49}$$

Posterior analysis can be performed by sequentially drawing from (45), (46), (47), (48) and (49).

References

[1] Blackburn, M. and D. Neumark, 1993, Omitted ability bias and the increase in the return to schooling, Journal of Labor Economics 11(3), 521-544.

[2] Casella, G. and E. George, 1992, Explaining the Gibbs Sampler, The American Statistician 46, 167-174.

[3] Cawley, J., Conneely, K., Heckman, J. and E. Vytlacil, 1997, Cognitive ability, wages and meritocracy, in: S. Fienberg, D. Resnick and K. Roeder eds., Intelligence, genes and success: Scientists respond to the Bell Curve (Springer Verlag, New York).

[4] Cawley, J ., Heckman, J. and E. Vytlacil, 1999, On policies to reward the value added by educators, Review of Economics and Statistics 81(4), 720-728.

[5] DiNardo, J. and J.L. Tobias, 2001, Nonparametric density and regression estimation, Journal of Economic Perspectives 15(4), 11-28.

[6] Heckman, J. and E. Vytlacil, 2001, Identifying the role of cognitive ability in explaining the level of and change in the return to schooling, Review of Economics and Statistics 83(1), 1-12.

[7] Koop, G. and D.J. Poirier, 2003a, Bayesian variants of some classical semiparametric regression techniques, Journal of Econometrics, forthcoming.

[8] Koop, G. and D.J. Poirier, 2003b, Empirical Bayesian inference in a nonparametric regression model, to appear in a volume from the Conference in Honour of Professor J. Durbin on State Space Models and Unobserved Components..

[9] Koop, G. and J.L. Tobias, 2002, Learning about heterogeneity in returns to schooling, Journal of Applied Econometrics, forthcoming.

[10] Li, Q., Huang, C., Li, D. and T. Fu, 2002, Semiparametric smooth coefficient models, Journal of Business and Economic Statistics 20, 412-422.

[11] Lindley, D. and A.F.M. Smith, 1972, Bayes estimates for the linear model, Journal of the Royal Statistical Society, Series B 34, 1–41.

[12] McLachlan, G. and D. Peel, 2000, Finite mixture models (John Wiley & Sons Inc., New York).

[13] Poirier, D.J., 1995, Intermediate Statistics and Econometrics (The MIT Press, Cambridge).

[14] Raftery, A., 1996, Hypothesis testing and model selection, in: Markov chain monte carlo in practice, W. Gilks, S. Richardson and D. Spiegelhalter, eds., (Chapman and Hall, Boca Raton) 163-188.

[15] Taber, C., 2001, The rising college premium in the eighties: Return to college or return to unobserved ability?, Review of Economic Studies 68(3), 665-691.

[16] Tanner, M.A. and W.H. Wong, 1987, The calculation of posterior distributions by data augmentation, Journal of the American Statistical Association 82, 528-549.

[17] Tobias, J.L., 2003, Are returns to schooling concentrated among the most able? A semiparametric analysis of the ability-earnings relationships, Oxford Bulletin of Economics and Statistics 61(1), 1-29.

[18] Yatchew, A., 1998, Nonparametric regression techniques in economics, Journal of Economic Literature 36, 669-721.