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
Iowa State University
tobiasj@iastate.edu

October 2004

Abstract

We describe procedures for Bayesian estimation and testing in cross sectional, panel data and nonlinear *smooth coefficient models.* 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 - for example, estimation, testing and smoothing parameter selection can be carried out *analytically* in the cross-sectional smooth coefficient model.

We apply our methods using data from the National Longitudinal Survey of Youth (NLSY). Using the NLSY data we first explore the relationship between ability and log wages and flexibly model how returns to schooling vary with measured cognitive ability. We also examine model of female labor supply and use this example to illustrate how the described techniques can been applied in nonlinear settings.

JEL Classifications: C11, C15, C51

Keywords: Bayesian econometrics, Semiparametric Regression, Smooth Coefficient Models.

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* (see, 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 to be constant and equal to a parameter β, the smooth coefficient model writes this marginal effect as an unknown function of some explanatory variable, say z. 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 with an endogeneity problem, and finally, in an ordered probit model. This last model illustrates how the described methods generalize to nonlinear models which can be regarded as linear in an equivalent latent variable representation.

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 univariate cross-sectional smooth coefficient model to a panel data model with endogeneity concerns, and some nonlinear models.

6. Finally, our approach avoids the use of complicated asymptotics (which are potentially inappropriate in modest samples) for inference.

In addition to the theoretical contributions, we provide two different applications showing how the estimation techniques can be used in practice. Specifically, we first 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.

We then introduce a second application (also using NLSY data) and continue to investigate the role of measured cognitive ability on outcomes of interest. This application also illustrates how the smooth coefficient model can be used in nonlinear models that can be regarded as linear in suitably-defined latent data. We take a cross section of NLSY outcomes in 2000 and model the labor supply choices made by a sample of married white females.[4] Specifically, we model the decisions made by these females to remain out of the labor force, to work part time or to work full time, and thus introduce

[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.

[4] See Gangadharan and Rosenbloom (1996), Angrist and Evans (1998), Buchmueller and Valetta (1999), Anderson and Levine (1999) and Chou and Staiger (2001), among many others, for previous work on female labor supply issues. To our knowledge, no studies involving female labor supply have estimated a model with the flexibility of the smooth coefficient specification described in this paper.

a three choice ordered probit model. Following Nandram and Chen (1996), we employ a rescaling transformation to improve the mixing of our chain and show how the smooth coefficient model can be applied in this nonlinear (and reparameterized) setting. In this application we clearly see the importance of an ability-spousal income interaction in our smooth coefficient ordered probit model. As a result, we obtain predicted probabilities of each employment state that differ significantly from those obtained from a default linear specification. Finally, marginal likelihood calculations reveal a preference for the generalized smooth coefficient model over this baseline linear alternative.

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. We begin with a univariate cross-sectional smooth coefficient model. Section 2.2 takes up the case of a longitudinal smooth coefficient model with an endogeneity problem while section 2.3 introduces an ordered probit model. In section 3 we provide some generated data experiments to illustrate the performance of our methods. Section 4 describes the NLSY data used in our empirical analyses, and results from our two applications are provided in sections 5 and 6. The paper concludes with a summary in section 7, and remaining details regarding priors and posterior simulators can be found in the appendix.

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 primarily introduced with an eye toward our empirical examples, the specifications we consider are quite general and can be used in a variety of applications.

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):

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

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 $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$.[5]

[5] 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), Albert and Chib (1993)].

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 our returns to schooling application 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.

We develop a semiparametric framework similar to that described in Koop and Poirier (2004a,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

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

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:[6]

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

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

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

[6] 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 (2004a)).

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

$$R = \left[\begin{array}{ccccc} 0 & D^* & 0 & D^{**} & 0 \\ 0 & 0 & D^* & 0 & D^{**} \end{array} \right]. \tag{5}$$

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.

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 $2/[\underline{\nu}\underline{s}^4]$.

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

$$\underline{V}_\beta = \underline{V}_\beta(\eta_1, \eta_2) = \left[\begin{array}{lll} \underline{V}_1 & 0 & 0 \\ 0 & V(\eta_1) & 0 \\ 0 & 0 & V(\eta_2) \end{array} \right], \tag{8}$$

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 (2004a). 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 $i = 3, .., N$, $j = 1, 2$.[7] 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 (oversmoothing).

Koop and Poirier (2004a,b) discuss the relationships between this general approach (i.e. treating points on the nonparametric regression lines as unknown parameters in a Normal linear regression models and using priors to smooth) and other approaches to non- and semiparametric regression. They stress that an important advantage of this approach over others is the simplicity and clarity obtained by staying within the familiar framework of the Normal linear regression model. The reader interested in related approaches is referred to (among many others) Green and Silverman (1994) for a discussion of the penalized likelihood approach, Silverman (1985) or Wahba (1983) for a discussion of splines and Smith and Kohn (1996) for a clever implementation of the spline approach using Bayesian model averaging. It is worth noting that, unlike our paper, these papers do not consider the smooth coefficient model, nor panel data, nor allow for endogeneity.

As stressed in Koop and Poirier (2004b) semiparametric regression models of the sort we consider can also be considered as state space models simply by changing i subscripts to t and treating nonparametric components such as $f_1(A_i)$ as states, our smoothness priors as state equations and our empirical Bayesian procedure (outlined below) as being analogous to estimating the signal-noise ratios. Thus, all the contributions made in this paper add to the state space literature as well. Koop and Poirier (2004b) only consider the simplest nonparametric regression model, so one can interpret the present paper as extending this in the direction of the smooth coefficient model (in both linear and nonlinear settings) with panel data and endogeneity.

At a more general level, the relationship between state space models and nonparametric regression has been noted before (see, e.g., Ansley and Kohn (1985) and Harvey and Koopman (2000)). There is a large literature related to Bayesian analysis of state space models (see, among many others, Carter and Kohn (1994), DeJong and Shephard (1995), Fruhwirth-Schnatter (1994a,b), Hickman and Miller (1981), Koop and van Dijk (2000), Shively and Kohn (1997) and West and Harrison (1997)). With the partial exception of Koop and van Dijk (2000), prior elicitation has not been a central focus of this literature. Hence, another contribution of our paper is to develop new ways of thinking about prior elicitation in sophisticated state space models (including extensions for panel data and endogeneity).

[7] 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), p. 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):

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

where

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

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.*[8]

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.[9] 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).

[8] 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 (2004b) 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.

[9] 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 take up the case of a *longitudinal* or *hierarchical* smooth coefficient model.[10] In this panel 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 of section 5, 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 $\varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2\right)$. Our model is completed by specifying the following two equations:

$$\begin{aligned} \alpha_i &= w_i\theta + f_1(A_i) + s_i f_2(A_i) + u_i & (17) \\ s_i &= r_i\pi + v_i. & (18) \end{aligned}$$

In terms of our application, A_i represents a continuous measure of individual i's cognitive ability, s_i denotes the number of years of schooling completed and w_i are a remaining set of time-invariant characteristics. Thus, the model described above permits (within a panel setting) baseline nonlinearities between ability and log wages through f_1 and also permits returns to education to depend nonparametrically on ability through f_2. Finally, note that $\mu_i = w_i\theta + f_1(A_i) + s_i f_2(A_i)$ can be reparameterized as $\mu_i = x_i\beta$ as in (6) to work directly with second differences of the regression functions. In what follows, we assume that the mean function has been transformed in this way, and thus will work directly with the vector β. This parameter vector includes θ, the initial points on the regression curves, and the second differences of the regression functions.[11]

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.

[10] For some discussion and applications of Bayesian hierarchical modeling, see, for example Laird and Ware (1982), Gelfand et al (1990) and Lange et al (1992), among many others.

[11] From a Bayesian point of view, we have specified a hierarchical prior for α_i, although a non-Bayesian may wish to interpret it as part of the likelihood function. There is a huge Bayesian literature on parametric panel data models or extensions to models such as probit (see, among many others, Chib and Greenberg (1995), Chib (2004), Geweke, Keane and Runkle (1997), Koop, Osiewalski and Steel (1997) and McCulloch and Rossi (1994)) and most of these papers emphasize the hierarchical prior interpretation. Our treatment of longitudinal smooth coefficient model draws upon standard, well-known results from this literature.

Finally, we also wish to consider the case where the quantity of schooling attained, s_i, is endogenous, leading us to specify the reduced form schooling equation in (18). In (18), 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 (16) or (17). 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 (16), and when T_i is reasonably large, the marginal posterior for α_i will tend to be dominated by data information from (16). To make our argument clear, let us suppose for the moment that the α_i are known. Equations (17) and (18) then constitute a triangular simultaneous equations model. If u_i and v_i are sufficiently correlated, a cross-sectional smooth coefficient analysis using only (17) (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.

To account for the potential endogeneity of schooling, we make the assumption

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

and to implement a Bayesian analysis we require priors for δ, σ_ε^2, Σ^{-1}, π and β. For the first four of these we make standard choices:

$$\begin{aligned} \delta &\sim N(\underline{\delta}, \underline{V}_\delta), & (19)\\ \sigma_\varepsilon^{-2} &\sim G(\underline{s}_\varepsilon^{-2}, \underline{\nu}_\varepsilon), & (20)\\ \Sigma^{-1} &\sim W\left([\underline{\rho}\underline{A}]^{-1}, \underline{\rho}\right) & (21)\\ \pi &\sim N(\underline{\pi}, \underline{V}_\pi) & (22) \end{aligned}$$

where $W(\cdot,\cdot)$ denotes a Wishart density (e.g., Poirier (1995), pages 136-138), the typical conjugate prior for the inverse covariance matrix. To complete the prior specification, we place a smoothing prior over the elements of β (as described in section 2.1) and thus specify:

$$\beta \sim N(\underline{\beta}, \underline{V}_\beta), \tag{23}$$

where $\underline{V}_\beta = \underline{V}_\beta(\eta_1, \eta_2)$ as defined in (8).

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. Despite its seeming complexity, estimation of the model in (16)-(18) requires only the ability to sample from standard distributions. Complete details behind this procedure are described in the appendix.

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 approximation[12] for the marginal likelihood (see, e.g., Raftery (1996), page 171). Let $\Gamma = [\delta, \beta, \pi, \sigma_{uv}, \sigma_\epsilon^2, \sigma_v^2, \sigma_{uv}]$ denote the parameters of the model other than α. We first integrate the individual effects α out of the likelihood function[13] to obtain:

$$
\begin{aligned}
p\left(y, s|\Gamma\right) \quad \propto \quad & p(s|\Gamma) \prod_{i=1}^{N} |C_i|^{-\frac{1}{2}} * \\
& \exp\left[-\frac{1}{2} \left(y_i - \iota_{T_i} \left[\mu_i + \frac{\sigma_{uv}}{\sigma_v^2}(s_i - r_i\pi) \right] - z_i\delta \right)' C_i^{-1} \left(y_i - \iota_{T_i} \left[\mu_i + \frac{\sigma_{uv}}{\sigma_v^2}(s_i - r_i\pi) \right] - z_i\delta \right) \right],
\end{aligned} \tag{24}
$$

where $\mu_i = x_i\beta$ and $C_i = \sigma_\epsilon^2 I_{T_i} + \sigma_\alpha^2(1-\rho_{uv}^2)\iota_{T_i}\iota_{T_i}'$. Then a highly accurate approximation to the marginal likelihood is given by:

$$
p\left(y, s\right) \approx (2\pi)^{\frac{1}{2}(2N+k_w+k_z+k_r+4)} \left|\Psi\right|^{\frac{1}{2}} p\left(y, s|\widehat{\Gamma}\right) p\left(\widehat{\Gamma}\right), \tag{25}
$$

where $\widehat{\cdot}$ denotes the posterior mode and Ψ is the posterior covariance matrix of Γ.[14] 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. Finally, empirical Bayesian selection of smoothing parameters can be conducted using the same procedure described in Section 2.1 using the Laplace-Metropolis approximation.

[12] Of course, a variety of methods could be employed to calculate the marginal likelihood. In this paper we do not take up the issue of comparing the performances of various methods within the class of smooth coefficient models and defer this as a subject for future work.

[13] We first obtain the joint distribution of (y, s) given the parameters and then integrate out α by marginalizing over the conditional distribution of $\alpha|s$. See the appendix for related discussion.

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

2.3 A Smooth Coefficient Ordered Probit Model

Our second application will illustrate how our smooth coefficient approach can also be applied in many nonlinear models. With an eye toward our female labor supply application, we consider the following smooth coefficient version of an ordered probit model:

$$z_i = w_i\theta + f_1(A_i) + f_2(A_i)s_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0,1), \tag{26}$$

and

$$y_i = \begin{cases} 1 & \text{if } \alpha_0 < z_i \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < z_i \leq \alpha_2 \\ \vdots & \vdots \\ L & \text{if } \alpha_{L-1} < z_i \leq \alpha_L \end{cases},$$

where (y_i, A_i, w_i, s_i) denotes the observed data and z_i represents an unobservable latent variable. For identification purposes we set $\alpha_0 = -\infty$, $\alpha_1 = 0$ and $\alpha_L = \infty$. Finally, note again that the mean function $\mu_i = w_i\theta + f_1(A_i) + f_2(A_i)s_i$ can be written as $x_i\beta$, as described in section 2.1. In what follows, we work with this parameterization and write (26) as $z_i = x_i\beta + \varepsilon_i$.

Given that the model is linear in the latent data z (but nonlinear when marginalized over z), data augmentation (e.g., Tanner and Wong (1987)) can be used to greatly simplify the required calculations. That is, we augment the parameter space with the latent data z, obtain the complete data likelihood function (e.g. Albert and Chib (1993)) and augmented joint posterior distribution.

We will proceed in the same manner as we did for the cross sectional model (see Section 2.1) and first sort the data by values of A. We will impose a degree of smoothing on the regression curves by placing a smoothing prior over the elements of β. Specifically, we let $\alpha = [\alpha_2 \; \alpha_3 \; \cdots \; \alpha_{L-1}]$, and specify

$$p(\beta, \alpha) = p(\alpha)p(\beta) \propto \phi[\beta; 0, \underline{V}_\beta(\eta_1, \eta_2)],$$

with $\underline{V}_\beta$ as defined in (8) and $\phi(x; \mu, \sigma^2)$ denoting a Normal density for x with mean μ and variance σ^2. The augmented joint posterior distribution is then obtained:

$$p(\alpha, \beta, z|y) \propto \left[\prod_{i=1}^{N} \exp\left(-\frac{1}{2}(z_i - x_i\beta)^2\right) I(\alpha_{y_i - 1} \leq z_i < \alpha_{y_i})\right] \phi(\beta; 0, \underline{V}_\beta),$$

with $I(\cdot)$ denoting the standard indicator function. It has been noted (e.g. Cowles (1996)) that the standard Gibbs algorithm for the ordered probit suffers from slow mixing, particularly in larger samples. To surmount this problem, we follow Nandram and Chen (1996) and introduce a rescaling transformation. To fix ideas and remain consistent with our application, we also focus below on the case of a three-level ordered probit model and thus set $L = 3$.[15]

For this transformation, we let

$$\delta = (1/\alpha_2), \;\; \alpha^* = \delta\alpha, \;\; \beta^* = \delta\beta, \;\; z^* = \delta z.$$

[15] In models with more than three levels (choices), an additional step will be required to sample the unknown transformed cutpoints α^*. Nandram and Chen (1996) suggest a Metropolis-Hastings step using a Dirichlet proposal density, which can also be adopted in this smooth coefficient extension.

We then obtain the joint posterior distribution of $(\delta^2, \beta^*, z^*|y)$ by a change of variables. Letting k_θ denote the dimension of θ and noting that the Jacobian of this transformation is $(\delta^2)^{-(N+k+3)/2}$ with $k = 2N + k_\theta$, we obtain

$$p(\delta^2, \beta^*, z^*|y) \propto (\delta^2)^{-3/2} \left[\prod_{i=1}^{N} \phi(z_i^*, w_i\beta^*, \delta^2) I[\alpha_{y_{i-1}}^* < z_i^* \leq \alpha_{y_i}^*] \right] \phi_k(\beta^*, 0, \delta^2 \underline{V}_\beta(\eta_1, \eta_2)). \tag{27}$$

Note that in this transformed model there are no unknown cutpoints since $\alpha_2^* = 1$. As described in the appendix, this model can be fit quite simply using the Gibbs sampler, and in our experience, the reparameterization described above significantly improves the mixing of the parameter chains. This example also reveals how our smooth coefficient model can be employed in a variety of other models that are nonlinear, but can be equivalently represented as a model that linear is in some latent data. For instance, smooth coefficient probit, tobit, generalized tobit, multinomial probit and stochastic frontier models can be handled in the same way.

3 Generated Data Experiments

In this section we perform several generated data experiments to illustrate the performance of our methods in recovering the true shape of the regression functions. We first consider the univariate smooth coefficient model (section 2.1) and perform two experiments using this model. The first experiment 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

$$y_i = 2w_i + A_i \cos(4\pi A_i) + \sin(2\pi A_i) s_i + \varepsilon_i, \tag{28}$$

where ε_i, A_i and w_i are 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:

$$y_i = w_i\theta + \lambda_0 + \lambda_1 A_i + \lambda_2 s_i + \varepsilon_i. \tag{29}$$

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 (29) as the data generating process. To be precise this second data generating process 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 the log of the Bayes factor in favor of the smooth coefficient model over the linear model given in (29) 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[16] 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 The Data

The data used in our empirical work are taken from the National Longitudinal Survey of Youth (NLSY). The NLSY is a rich 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.

As discussed throughout this paper, we are primarily interested in flexibly exploring the relationship between measured cognitive ability and economic outcomes. Specifically, we investigate the rela-

[16] Because of problems associated with calculating $|\underline{V}_\beta|^{-1/2}$ at $\eta_1 = \eta_2 = 0$, we perform a grid search bounding the elements of η_1 and η_2 slightly above zero. The linear model is considered as a separate specification. Empirical Bayes methods seek to drive these hyperparameters toward zero in this experiment, as one might expect given the results of our marginal likelihood calculations and the fact that the linear model is the "true" specification.

tionship between ability and log wages in a smooth coefficient hierarchical (panel) model, and also investigate how ability impacts the labor supply decisions of married white females in a smooth coefficient ordered probit model. Fortunately for this purpose, the NLSY reports scores on 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.[17]

Sample Restrictions in the Hierarchical Smooth Coefficient Model

For the hierarchical application we exploit the panel structure of the NLSY. The NLSY panel begins in 1979 at which point the respondents range from 14 to 22 years of age. Information from annual interviews are collected until 1994, and after 1994 we are able to obtain data from biannual surveys until 2000.

The time-varying characteristics we include as determinants of log wages (z_{it}) 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).

In keeping with the majority of this literature, we restrict the sample used in our hierarchical analysis to 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 five 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

[17] 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.

eighteen observations.

In order to deal with the potential endogeneity of schooling in the hierarchical model, 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.* This particular instrument choice imposes 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[18] 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, we include ABILITY, MOMED, DADED, NUMSIBS and SIBED in the reduced form schooling equation.

Sample Restrictions in the Smooth Coefficient Ordered Probit Model

To estimate the ordered probit model of section 2.3, we extract information on the labor supply decisions of a cross-sectional sample of married white females in the NLSY. Specifically, we use our ordered probit specification to model the decision made by a sample of women to remain out of the labor force ($y = 1$), to work part time ($y = 2$) or to work full time ($y = 3$) in 2000. We extract data on measured cognitive ability (ABILITY), education (ED), spousal income (SPINC), and number of children (NUMKIDS) in the household and include these variables in our ordered probit specification. Our smooth coefficient extension of this ordered probit model permits a flexible baseline function of ability in the latent variable equation (f_1 in (26)). Perhaps most importantly, our smooth coefficient ordered probit also allows for a flexible interaction between spousal income and ability. Our prior expectation is that when the income of the husband is small, women of all ability types will have a large incentive to participate in full-time employment. Conversely, when the income of the husband is large, then we will see more differentiation in female labor supply decisions. Specifically, high ability women will work full time with significantly higher probability than lower ability women, and women of all ability types will have a lower probability of full time employment when spousal income is high. After restricting the sample to married white females from the cross-sectional sample with non-missing information for the required variables, we obtain a final sample of $N = 655$ observations.

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

5 Empirical Results: Hierarchical Smooth Coefficient Model

In this section we turn to our first application and present empirical results using the NLSY data 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 also discussed in Section 4. We are particularly in interested in using the NLSY data to investigate 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 these relationships in the NLSY data?[19]

5.1 Hierarchical Model without Endogeneity

To fix ideas we first consider the case of a hierarchical model with no endogeneity problems and thus impose $\sigma_{uv} = 0$[20] in the model of Section 2.2. Under this assumption we no longer require an instrument in the schooling equation and thus do not need to restrict our attention to only those individuals with older siblings in the NLSY. This provides us with more observations to explore the relationship between ability and log wages and to determine how returns to education vary with measured cognitive ability. Finally, under the assumption of exogenous schooling, our model is described by (16) and (17) and the reduced form schooling equation in (18) does not need to be specified.

To investigate the adequacy of often-used parametric models in this literature (which often argue that potential endogeneity problems can be ignored upon including a rich set of covariates in the wage equation), we calculate the Bayes factor comparing the smooth coefficient 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:[21]

$$\mu_i = w_i\theta + \lambda_0 + \lambda_1 A_i + \lambda_2 s_i. \tag{30}$$

Our general prior elicitation strategy remains the same as that described in section 2.1. 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 (19) - (23) 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 3).

[19] 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.

[20] This assumption is relaxed in the following section.

[21] 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.

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

		Smooth Coefficient Model		Parametric Model	
	Variable/ Parameter	Post. Mean	Post. St. Dev.	Post. Mean	Post. 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

Posterior results are produced using the MCMC algorithm described in the appendix.[22] 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 (30). The log of this Bayes factor is -948.6, indicating *strong* evidence in favor of the parametric model. The reason for this is clear. With very small values of η_1 and η_2 the two models fit the data roughly equally.[23] 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). 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.

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 Figures 3 and 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 $E(f_1(A)|\text{Data}) \pm 2\text{Std}(f_1(A)|\text{Data})$

[22] 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.

[23] As evidence of this, we used the simulated posterior output from each model to calculate the log likelihoods at each iteration of the sampler. The maxima of these log likelihoods were equal to one decimal place.

(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 the 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.[24] 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.

5.2 Longitudinal Model with Endogeneity

The longitudinal model with endogeneity is given in (16) through (18). The reduced form schooling equation (18) is now added to the analysis. This equation 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. The third equation, (18), is the same in the parametric and smooth coefficient models.

For parameters that are common to both the models in this and the previous section, we retain the same prior specifications that were used in the previous section. 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 (21) we use relatively noninformative values[25] of $\underline{\rho} = 9$ and $\underline{A} = I_2$.

[24] 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.

[25] 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.

Table 2: Posterior Results for First and Second Stage Parameters: Model with Endogeneity

		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 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.

For parameters which are common to both models, we again find that results obtained from the parametric and smooth coefficient models are strikingly similar. The posterior mean (and standard deviation) of the correlation between the errors in two second stage equations (i.e. the source of the endogeneity problem) was found to be $-.24$ (.40) in the smooth coefficient model and $-.16$ (.49) in the parametric model. This suggests that with this relatively small sample size[26] we are not able to pin down this key correlation parameter, and thus learn little regarding the empirical importance of endogeneity. Finally, 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 and serves as an adequate instrument.

For those quantities which are not common to both specifications (namely the specification of ability and the ability-education interaction), we found no convincing evidence of nonlinearities.[27] This again suggests that standard parametric models - which are the workhorses of this literature - appear to

[26] Recall that we have only 98 individuals to estimate the bivariate relationship in (17) and (18).

[27] These results are not reported here, but are available upon request.

be adequate for analysis of these issues using the NLSY data. This result both strengthens the conclusions of previous studies which have been based on simpler specifications and also suggests that simpler models are adequate for future work addressing similar topics using the NLSY. 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 parametric form.

6 Empirical Results: An Ordered Probit Model of Female Labor Supply

We now turn to our second application involving the NLSY data and model the labor supply decisions of a sample of married white females in 2000. Specifically, we model their decision to remain out of the labor force, to work part time or to work full time in the given year. We estimate our three-choice ordered probit model using the Gibbs sampling algorithm described in the appendix.

In this application we pay particular attention to how changes in spousal income affect changes in our three employment states for women of varying ability. Our smooth coefficient ordered probit model in (26) permits a flexible interaction between ability and spousal income and thus is well-suited for this investigation. As a competing specification we also consider the performance of a fully parametric model that is linear in all the explanatory variables.

Using our empirical Bayesian methods, we settle on $\eta_1 = 1.0 \times 10^{-7}$ and $\eta_2 = 1.0 \times 10^{-11}$ as optimal values of the smoothing parameters. We present in Table 3 below estimation results for the parametric portion of the model using this value of η_1 and η_2.

Table 3: Posterior Means and Standard Deviations on Parametric Coefficients in Ordered Probit Model

Variable	Post Mean	Post Std.
Num-Kids	-.23	.044
Education	.033	.027
Unemp-rate	.035	.033
Cutpoint (α_2)	.032	.045

As can be seen from Table 3 (and as one certainly might expect), the number of children in the household plays a strong role in the model, with more children in the home clearly decreasing the probability that a woman works full time and increasing the probability that she remains out of the labor force.

To investigate how changes in spousal income impact labor supply decisions across the ability distribution, we calculate the probabilities of each employment state over the ability support. Specifically, we compute

$$\Pr(y = j|s, w, A = A_0, \alpha, \gamma, \theta) = \Phi\left(\alpha_j - w\theta - f_1(A_0) - f_2(A_0)s\right) - \Phi\left(\alpha_{j-1} - w\theta - f_1(A_0) - f_2(A_0)s\right),$$

where A_0 is a particular point in the ability distribution. We set the variables in w equal to their sample means and evaluate these probabilities at various values of spousal income, s. We calculate these probabilities for each employment state ($j = 1, 2, 3$) and for each observed ability value A_i in the sample. The draws obtained from the posterior distribution of α, γ and θ, are then used to obtain a point estimate of the desired probability at each ability value. Finally, we also calculate these effects for the competing parametric (linear) specification and compare these estimates to those obtained from the smooth coefficient model.

Results of this exercise are presented in Figures 5 and 6. In these figures we plot results for both the smooth coefficient and parametric models when spousal income is set at \$120,000 (approximately the 95th percentile of the spousal income distribution), and thus investigate the impact of ability on female labor supply decisions for those with high levels of family income. As one can see, the figures suggest important differences between the smooth coefficient and linear models. First, the smooth coefficient model suggests that very low ability women are more likely to stay out of the labor force than to work full time at this high level of spousal income. In contrast, the parametric model predicts that the probability of full time work exceeds the probability of no work at all values of the ability distribution. Second, both models suggest that the probability of full time (no) work is increasing (decreasing) with ability, but the slopes of these changes differ considerably across specifications. Specifically, our smooth coefficient model shows a sharp increase in the probability of full time employment for ability values greater than one standard deviation above the mean and a corresponding sharp decrease in the probability of no work over this region. The fully parametric model can not account for these features of the data and simply predicts a nearly linear trend through the ability support.

At the other end of the spousal income distribution, results were very similar across model specifications. When spousal income is \$10,000 (approximately the 4th percentile of the income distribution), for example, both the smooth coefficient and linear models predict a high probability of full-time employment and a low probability of remaining out of the labor force regardless of ability level.[28] Taken together, these results show there is little evidence of or role for baseline nonlinearities in ability (through f_1), but there is an important ability- spousal income interaction (through f_2). When evaluating these probabilities at low levels of spousal income, the contribution of the interaction term is minimal and predictions are found to be similar across the smooth coefficient and parametric models. However, for larger values of spousal income (s), the important role of the ability-income interaction becomes clear. Finally, the log of the Bayes factor in favor of the smooth coefficient model over the parametric alternative was found to be 3.82, indicating preference for the more general smooth coefficient specification despite its added parameterization.

[28] We do not present a full set of results for the sake of brevity. Estimates of f_1, f_2 and probabilities of each state with $s = 10,000$ are available upon request.

7 Conclusion

In this paper we have described Bayesian procedures for estimation and testing in cross sectional, longitudinal data and nonlinear 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 nonlinear models that can regarded as linear latent variable models, estimation only requires iterative simulation from standard distributions.

We illustrated the flexibility and practicality of our methods in generated data experiments and in applications 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. In an ordered probit model of female labor supply, we preferred our flexible smooth coefficient model over a linear parametric alternative and found that in some cases, predictions regarding quantities of interest were quite different across models. These results are of substantive interest and also illustrate how our approach to the estimation of smooth coefficient models can be applied in a variety of settings.

7.1 Appendix A: Posterior Simulator for Longitudinal Model with Endogeneity

The model is given as:

$$\begin{aligned} y_{it} &= \alpha_i + z_{it}\delta + \varepsilon_{it} && (31) \\ \alpha_i &= w_i\theta + f_1(A_i) + s_i f_2(A_i) + u_i && (32) \\ s_i &= r_i\pi + v_i, && (33) \end{aligned}$$

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.

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

$$p(\alpha, \delta, \beta, \pi, \sigma_\varepsilon^{-2}, \Sigma^{-1}|y, s) \propto p(y, s|\alpha, \delta, \beta, \pi, \sigma_\varepsilon^2, \Sigma^{-1})p(\alpha, \delta, \beta, \pi, \sigma_\varepsilon^{-2}, \Sigma^{-1}). \qquad (34)$$

the likelihood function is the joint density of y and the endogenous schooling variable s conditioned on the model parameters. The prior assumptions in (19)-(23) 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 (17) using the reduced form equation in (18).

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

$$\begin{aligned} p(y, s|\Gamma) &= \prod_{i=1}^{N} p(y_{i1}, \cdots, y_{iT_i}, s_i|\Gamma) && (35) \\ &= \prod_{i=1}^{N} \left(\prod_{t=1}^{T_i} p(y_{it}|\alpha_i, \delta, \sigma_\varepsilon^{-2})\right) p(s_i|\alpha_i, \pi, \Sigma^{-1}, \beta), && (36) \end{aligned}$$

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.

Combining the likelihood in (36) with our prior we obtain the unnormalized joint posterior

$$p(\Gamma|y, s) \propto \left[\prod_{i=1}^{N} \left(\prod_{t=1}^{T_i} p(y_{it}|\alpha_i, \delta, \sigma_\varepsilon^2)\right) p(s_i, \alpha_i|\pi, \Sigma^{-1}, \beta)\right] p(\delta)p(\beta)p(\pi)p(\sigma_\varepsilon^2)p(\Sigma^{-1}). \qquad (37)$$

To carry out posterior simulation in this model we 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:

$$\delta|Data, \Gamma_{-\alpha,\delta} \sim N(D_\delta d_\delta, D_\delta), \tag{38}$$

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:

$$\alpha_i | \Gamma_{-\alpha_i}, \text{Data} \overset{ind}{\sim} N\left(D_{\alpha_i} d_{\alpha_i} D_{\alpha_i}\right), \;\; i = 1, 2, \cdots N, \tag{39}$$

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 these equations, 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 (37) 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}:

$$\sigma_\varepsilon^{-2} | Data, \alpha, \delta \sim G\left(\overline{s}_\varepsilon^{-2}, \overline{\nu}_\varepsilon \right) \tag{40}$$

where

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

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

As for the complete conditionals for the second-stage regression parameters and inverse covariance matrix, let us first stack the triangular system in (32) and (33) 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

$$\tilde{\beta}|\Gamma_{-\tilde{\beta}}, \text{Data} \sim N\left(D_{\tilde{\beta}} d_{\tilde{\beta}}, D_{\tilde{\beta}}\right), \tag{41}$$

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

$$\Sigma^{-1}|\Gamma_{-\Sigma^{-1}}, \text{Data} \sim W\left[\left(\sum_{i=1}^{N}\psi_i\psi_i' + \underline{\rho}\underline{A}\right)^{-1}, \underline{\rho} + N\right]. \tag{42}$$

Posterior analysis can be performed by sequentially drawing from (38), (39), (40), (41) and (42).

7.2 Appendix B: Posterior Simulator for the Ordered Probit Model

From the joint posterior in the ordered probit model given in (26), the following complete conditionals are obtained:

$$\beta^*|z^*, \delta^2, y \sim N(D_\beta d_\beta, D_\beta)$$

where

$$D_\beta = \delta^2\left(X'X + \underline{V}_\beta^{-1}\right)^{-1}, \quad d_\beta = X'z^*/\delta^2.$$

As for the conditional for the latent data,

$$z_i^*|\beta^*, \delta^2, y \overset{ind}{\sim} TN_{[\alpha^*_{y_i-1}, \alpha^*_{y_i})}(x_i\beta^*, \delta^2),$$

where $TN_{(a,b)}(\mu, \sigma^2)$ denotes a Normal density with mean μ and variance σ^2 truncated to the interval (a, b). Finally,

$$\delta^2|\beta^*, Z^*, y \sim IG\left(\frac{N+k+1}{2}, \left[\frac{1}{2}(z^* - X\beta^*)'(z^* - X\beta^*) + \frac{1}{2}\beta^{*\prime}\underline{V}_\beta^{-1}\beta^*\right]^{-1}\right),$$

where IG denotes the inverted Gamma distribution (see Poirier (1995), p. 111). To recover the original coefficient vector β and cutpoint α_2, simply use the inverse transformations $\beta = (1/\delta)\beta^*$ and $\alpha_2 = (1/\delta)$. It is also with noting that the computationally expensive term $(X'X + \underline{V}_\beta^{-1})^{-1}$ in D_β can be calculated outside the Gibbs loop for given η_1 and η_2.

References

[1] Albert, J. and S. Chib, 1993, Bayesian analysis of binary and polychotomous response data, Journal of the American Statistical Association 88, 669-679.

[2] Anderson, P.M. and P.B. Levine, 1999, Child care and mothers' employment decisions, NBER Working Paper #7058.

[3] Angrist, J. and W. Evans, 1998, Children and their parents' labor supply: Evidence from exogenous variation in family size, American Economic Review 92, 307-322.

[4] Ansley, C. and R. Kohn, 1985, Estimation, filtering and smoothing in state space models with incompletely specified initial conditions, Annals of Statistics 13, 1286-1316.

[5] 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.

[6] Buchmueller, T.C. and R.G. Valetta, 1999, The Effect of health insurance on married female labor supply, Journal of Human Resources 43, 42-70.

[7] Carter, C. and R. Kohn, 1994, On Gibbs sampling for state space models, Biometrika 81, 541-553.

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

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

[10] 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.

[11] Chib, S., 2004, Panel data modeling and inference: A Bayesian primer, in The Econometrics of Panel Data - A Handbook of the Theory and Applications, 3rd edition, edited by Laszlo Matyas and Patrick Sevestre (Kluwer, Boston).

[12] Chib, S. and E. Greenberg, 1995, Hierarchical analysis of SUR models with extensions to correlated serial errors and time varying parameter models, Journal of Econometrics 68, 339-360.

[13] Chou, Y.J. and D. Staiger, 2001, Health Insurance and Female Labor Supply in Taiwan, Journal of Health Economics 20, 187-211.

[14] Cowles, M.K., 1996, Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models, Statistics and Computing 6, 101-111.

[15] DeJong, P. and N. Shephard, 1995, The simulation smoother for time series models, Biometrika 82, 339-350.

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

[17] Fruhwirth-Schnatter, S., 1994a, Data augmentation and dynamic linear models, Journal of Time Series Analysis 15, 183-202.

[18] Fruhwirth-Schnatter, S., 1994b, Bayesian model discrimination and Bayes factors for linear Gaussian state space models, Journal of the Royal Statistical Society, Series B 56, 237-246.

[19] Gangadharan, J. and J. Rosenbloom, 1996, The Effects of child-bearing on married women's labor supply and earnings: Using twin births as a natural experiment, NBER Working Paper #5647.

[20] Gelfand, A.E., Hills, S., Racine-Poon, A., and A.F.M Smith, 1990, Illustration of Bayesian inference in Normal data models using Gibbs sampling, Journal of the American Statistical Association 85, 972-985.

[21] Geweke, J., M. Keane, and D.E. Runkle, 1997, Statistical inference in the multinomial multiperiod probit model, Journal of Econometrics 80, 125-165.

[22] Green, P. and B. Silverman, 1994, Nonparametric regression and generalized linear models (Chapman and Hall: London).

[23] Harvey, A. and S. J. Koopman, 2000, Signal extraction and the formulation of unobserved components models, Econometrics Journal 3, 84-97.

[24] 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.

[25] Hickman, J. and R. Miller, 1981, Bayesian bivariate graduation and forecasting, Scandinavian Actuarial Journal, 129-150.

[26] Koop, G., J. Osiewalski and M.F.J Steel, 1997, Bayesian efficiency analysis through individual effects: Hospital cost frontiers, Journal of Econometrics 76, 77-105.

[27] Koop, G. and D.J. Poirier, 2004a, Bayesian variants of some classical semiparametric regression techniques, Journal of Econometrics, forthcoming.

[28] Koop, G. and D.J. Poirier, 2004b, 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.

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

[30] Koop, G. and H. van Dijk, 2000. Testing for integration using evolving trend and seasonals models: A Bayesian approach, *J*ournal of Econometrics 97, 261-291.

[31] Laird, N.M. and J. Ware, 1982, Random-effects models for longitudinal data, *Biometrics* 38, 963-974.

[32] Lange, N., Carlin, B.P. and A.E. Gelfand, 1992, Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-Cell numbers (with discussion) *Journal of the American Statistical Association* 87, 615-632.

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

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

[35] McCulloch, R. and P. Rossi, 1994, An exact likelihood analysis of the multinomial probit model, Journal of Econometrics 64, 207-240.

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

[37] Nandram, B. and M.-H. Chen, 1996, Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence, Journal of Statistical Computation and Simulation 54, 129-144.

[38] Poirier, D.J., 1995, Intermediate statistics and econometrics (The MIT Press, Cambridge).

[39] 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.

[40] Shively, T. and R. Kohn, 1997, A Bayesian approach to model selection in stochastic coefficient regression models and structural time series models, Journal of Econometrics 76, 39-52.

[41] 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.

[42] Smith, M. and Kohn, R., 1996, Nonparametric regression using Bayesian variable selection, Journal of Econometrics 75, 317-343.

[43] 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.

[44] 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.

[45] 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.

[46] Wahba, G., 1983, Bayesian confidence intervals for the cross-validated smoothing spline, Journal of the Royal Statistical Society, Series B 45, 133-150.

[47] West, M. and J. Harrison, 1997, Bayesian forecasting and dynamic models, second edition. (Springer Verlag, Berlin).

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

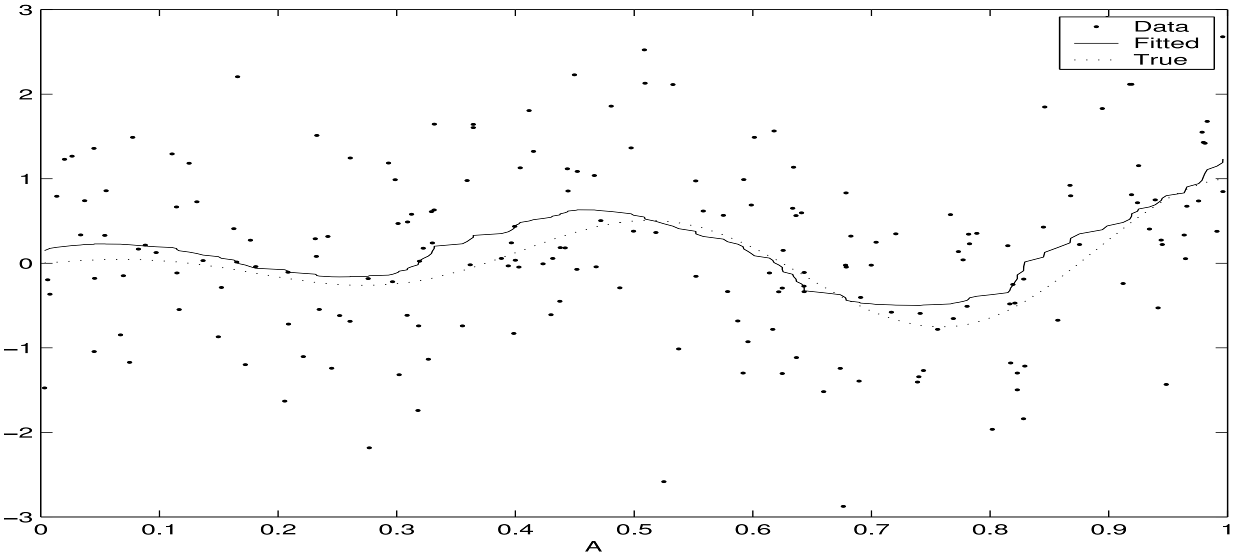

Figure 1: Fitted and True Regression lines for $f_1(A) = A\cos(4\pi A)$.

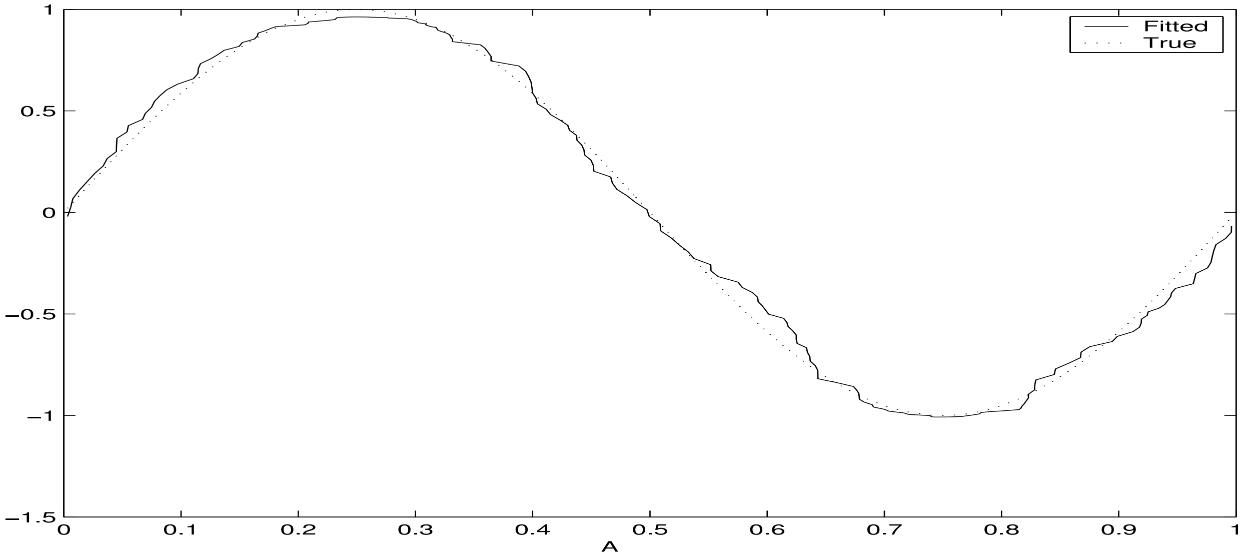

Figure 2: Fitted and True Regression Lines for Smooth Coefficient Term $f_2(A) = \sin(2\pi A)$.

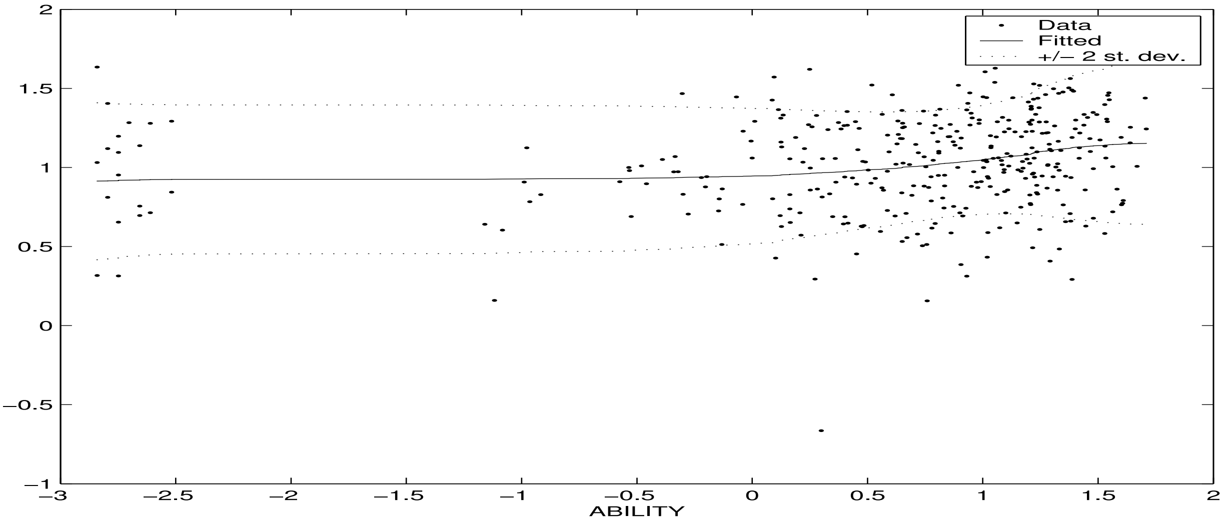

Figure 3: Fitted Regression Line for Ability Term $f_1(A)$ in Hierarchical Model

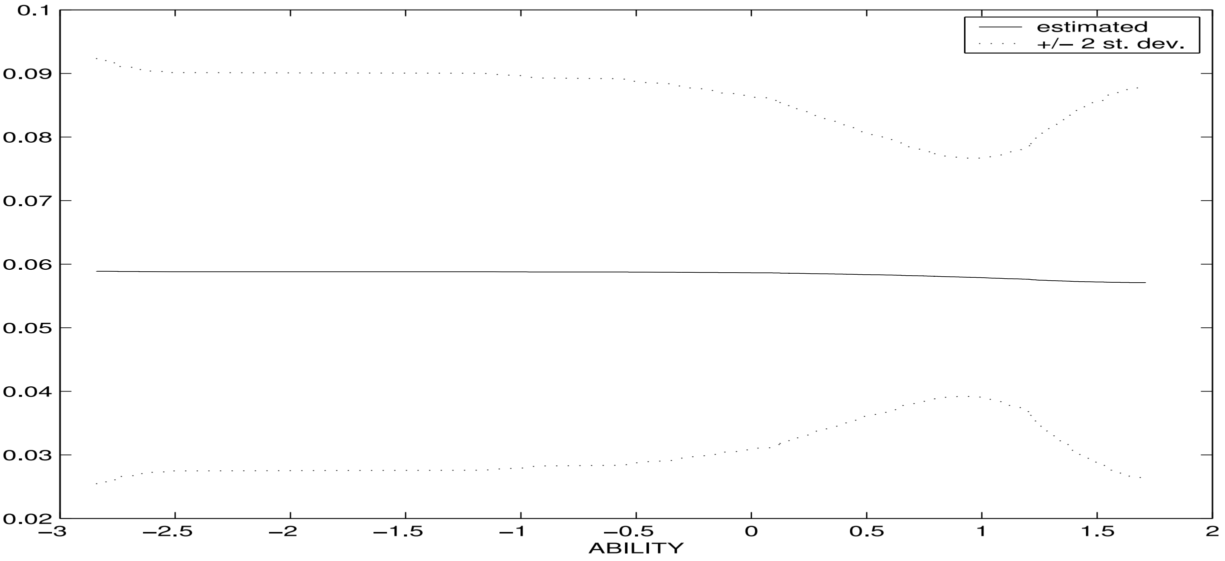

Figure 4: Fitted Regression Line for Return to Schooling Term $f_2(A)$ in Hierarchical Model.

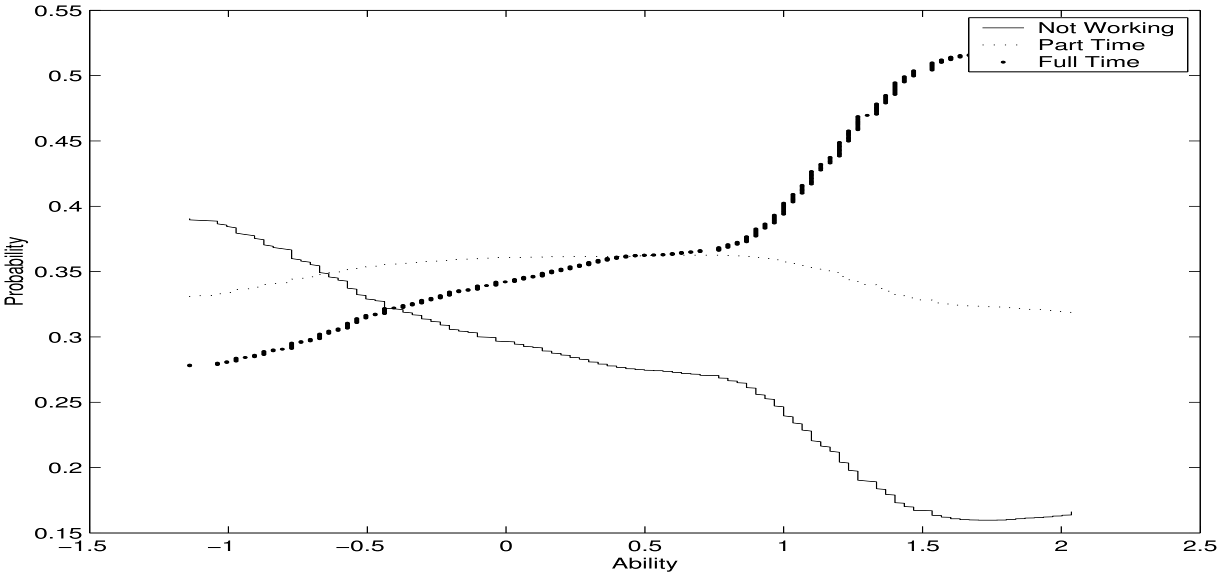

Figure 5: Probabilities of Each Employment State Across Ability Levels: Spousal Income = 120K, Smooth Coefficient Model

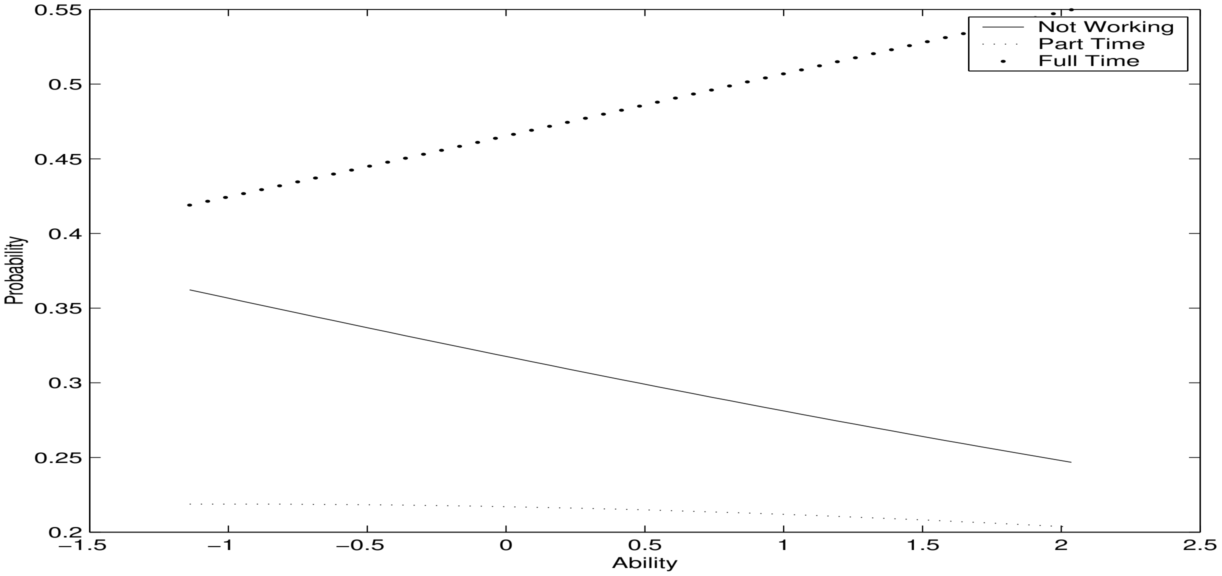

Figure 6: Probabilities of Each Employment State Across Ability Levels: Spousal Income = 120K, Parametric Model