FLEXIBLE CORRELATED RANDOM EFFECTS ESTIMATION IN PANEL MODELS WITH UNOBSERVED HETEROGENEITY

C. ALAN BESTER AND CHRISTIAN HANSEN†

Abstract. In this paper, we consider identification in a correlated random effects model for panel data. We assume that the likelihood for each individual in the panel is known up to a finite dimensional common parameter and an individual specific parameter. We allow the distribution of unobserved individual specific effects to depend on observed explanatory variables and make no assumptions about the particular functional form of this dependence. This leads to a semiparametric problem where the parameters include a finite dimensional common parameter, θ and an infinite dimensional conditional density, q, that describes the distribution of unobserved individual specific effects. For a given likelihood, we establish restrictions on the space of functions $\mathcal{H}$ for the distribution of unobserved heterogeneity under which $\{\theta, q\}$ are identified. We show the model parameters may be consistently estimated by sieve maximum likelihood for a fixed panel length, T. The conditions on $\mathcal{H}$, which include assumptions about the support of explanatory variables and smoothness of q in its arguments, are relatively mild and are similar to those required for nonparametric density estimation.

Keywords: identification, mixture models, nonparameteric ML

1. Introduction

Heterogeneity is an important feature of many economic models. In panel data settings, individual specific effects, α_i, are unobservable and often thought of as nuisance parameters that may be correlated with observed explanatory variables. Many papers in the recent literature have considered these models under asymptotics where N and T go to infinity; see, for example, Arellano and Hahn (2005) for a survey as well as Hahn and Kuersteiner (2002),

Date: This draft: December 14, 2007. First draft: April 7, 2006.

†University of Chicago, Graduate School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637. Email: chansen1@gsb.uchicago.edu.

Hahn and Kuersteiner (2004), and Hahn and Newey (2004) among others. Hueristically, in this setting each of the α_i are separately identified and may be treated as parameters to be estimated because the econometrician will eventually have an arbitrarily large amount of data for each individual. These estimation schemes are broadly referred to as fixed effects estimators. However, estimation and inference based on asymptotics where $T \to \infty$ may perform poorly in practice. In many settings, T is small relative to N, and error introduced by estimating each α_i individually may induce large finite sample biases in estimates of the common parameters, θ, and other quantities of interest.

Instead one may suppose that T is fixed and each α_i is drawn from a density $q(\alpha_i|\cdot)$. Estimates of θ based on a particular parametric specification for q, and/or restrictions such as independence between observables and unobservables, are generally referred to as random effects estimators. Examples include Hausman and Taylor (1981) for the linear model, and **?** for the Poisson count model. The tradeoff is that if one is willing to make assumptions about the structure of unobserved heterogeneity, one may obtain consistency and derive the asymptotic distributions of parameters of interest for fixed T.[1] This results in much better inference properties when T is small relative to N, but will generally result in an inconsistent estimator of θ if the assumptions about the distribution of unobserved heterogeneity are violated. Because economic theory often provides little guidance about the distribution of

[1]For certain specific structural models, fixed effects estimators have appeared in the literature which are consistent with T fixed. These estimators generally rely on transforming the data such that structural model is separable in θ and α_i, which is usually not possible in general nonlinear settings. These transformations are well-known for the linear model, Logit model, and Poisson model; see Wooldridge (2002). In addition, there are a variety of such approaches that apply to semiparametric panel data models under the assumption that the structural model depends on x_{it} and α_i only through a linear index but is otherwise unknown. Examples include Manski (1987) in the static binary choice model, Honoré and Kyriazidou (2000b) in dynamic binary choice, Honoré and Kyriazidou (2000a) in censored models, Horowitz and Lee (2004) in duration models, and Kyriazidou (1997) for selection models.

individual specific effects, fixed effects approaches are dominant in empirical economics in spite of their potentially poor finite sample performance in nonlinear models.

In this paper, we propose a flexible correlated random effects estimator for likelihood-based panel models. We place no restriction on q beyond assuming it is an element of a space of functions, $\mathcal{H}$. In particular, we allow the distribution of unobserved heterogeneity to depend on the observed explanatory variables and do not assume a particular functional form for this dependence. This leads to a semiparametric problem where the parameters, $\psi \equiv (\theta, q)$, include a finite dimensional vector θ that will enter in the objective function for each individual and an infinite dimensional function q that describes the distribution of unobserved individual specific effects.

Given this setting, we then ask how much flexibility we can accomodate in the distribution of individual specific heterogeneity while still retaining identification for our estimator of θ. We argue that the answer to this question depends on the amount of variation in the observed data: As the observed data become richer, the restrictions that need to be placed on q to achieve identification become weaker. For example, we argue that identification of θ and q among a general class of functions $\mathcal{H}$ will be possible when the observed outcome has continuous and unbounded support. In this paper, we consider using sieve maximum likelihood for estimation and demonstrate that it consistently estimates ψ.

The remainder of the paper is as follows. In the next section, we provide further details regarding the flexible random effect model. Identification arguments are presented in Section 3. Section 4 outlines the sieve maximum likelihood estimator and verifies consistency given the identification established in Section 3, and Section 5 contains the results from a simulation study. Section 6 concludes.

2. Models

We consider likelihood-based panel models with time invariant individual specific effects. Let the observed data be $z_i = (y_i, x_i)$, $i = 1, \ldots, N$ where $y_i \in \mathcal{Y} \subseteq \mathbb{R}^T$ is a vector of outcomes for individual i and $x_i \in \mathcal{X} \subseteq \mathbb{R}^{T \times d_x}$ is a $T \times d_x$ matrix of explanatory variables. We consider models defined by a vector of common parameters, $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, which also include time invariant individual specific heterogeneity that may be represented for each i by a vector, $\alpha_i \in \mathcal{A} \subset \mathbb{R}^{d_\alpha}$. We begin with a parametric likelihood model for the density of y_i given x_i, α_i, and θ, $g(y_i|x_i, \alpha_i, \theta)$, which we take as given and refer to as the structural model. We assume that $g(y_i|x_i, \alpha_i, \theta)$ corresponds to a well-defined parametric model that a researcher would actually use in practice; that is, if the individual specific effect α_i were observable, the common parameters θ would be identified and could be estimated by maximum likelihood.

Individual specific effects are drawn from a density, $q_0(\alpha|x)$ that may in general depend on the full matrix of explanatory variables and is assumed to lie in a space of functions, $\mathcal{H}(\mathcal{A} \times \mathcal{X})$. For each $\theta \in \Theta$, a conditional density $f(y|x, \theta)$ of the observables is obtained by integrating out over the distribution of unobserved heterogeneity,

$$f(y|x, \theta) = \int_{\mathcal{A}} g(y|x, \alpha, \theta) q_0(\alpha|x) d\alpha, \tag{2.1}$$

where we have supressed the i subscripts for notational convenience. We assume that for each $\theta \in \Theta$, the conditional density f is itself a member of a function space, $\mathcal{F}(\mathcal{Y} \times \mathcal{X})$. One may then estimate the model parameters by maximizing the sample analogue of the integrated likelihood,

$$L(\psi) = E_{y,x}\left[\log \int_{\mathcal{A}} g(y|x, \alpha, \theta) q(\alpha|x) d\alpha\right] \tag{2.2}$$

where $\psi = (\theta, q)$. When q_0 is known, the integral (2.1) defines a parametric family of distributions and θ may be estimated by conventional ML with the integral performed 'outside' the estimation procedure.

Many standard random effects estimators assume that q is known up to a finite dimensional vector of parameters, $\gamma \in \Gamma$, in which case $\mathcal{H}$ consists of functions of the form $\{q(\alpha|x,\gamma)|\gamma \in \Gamma\}$. Though the integral (2.1) must now be performed inside the estimation procedure, these models are also estimable by ML with the sample analogue of (2.2) being maximized over $\Theta \times \Gamma$. Unfortunately, it is rare for economic theory to dictate a particular parametric form for q, and estimates of θ will almost always fail to be consistent when $q_0 \notin \mathcal{H}$.

The inconsistency of the ML estimator of θ with an incorrectly specified parametric model for q has led to the prevalence of fixed effects approaches in econometrics. These approaches impose no restrictions on $\mathcal{H}$, the space of functions for q. Without such restrictions, a natural way to proceed is to treat the α_i as parameters to be estimated along with θ using the conditional likelihood $g(y_i|x_i, \alpha_i, \theta)$. While this approach will consistently estimate θ without restricting q when $T \to \infty$, it suffers from the well-known incidental parameters problem illustrated in Neyman and Scott (1948): When T is small, error from estimating α_i affects the estimates of all parameters and may lead to large biases and poor finite-sample performance of large T asymptotic approximations. The recent econometrics literature contains a number of papers that offer potential solutions to this problem using asymptotics where N and T go to infinity; see, for example, Hahn and Kuersteiner (2002), Hahn and Kuersteiner (2004), Hahn and Newey (2004), and Woutersen (2005). The basic idea of these approaches is to use higher-order bias corrections to remove the first order (in T) bias of the fixed effects estimators. Another approach considered in Honoré and Tamer (2006) and Chernozhukov, Hahn, and

Newey (2004) imposes no restrictions on q but does not worry about point identification and instead focuses on the identified set for θ.

In this paper, we consider a flexible random effects model where $\mathcal{H}$ is allowed to be an infinite dimensional function space. Our approach is semiparametric in that the parameter vector $\psi = (\theta, q)$ includes a finite dimensional vector of common parameters θ and an infinite dimensional function, q, with unspecified dependence on x. Ours is still fundamentally a random effects approach in that we will impose restrictions on the the space $\mathcal{H}$ to achieve identification, though the restrictions that will need to be imposed are generally much weaker than assuming a parametric model for q or independence between α and x. We emphasize that the semiparametric nature of the problem arises from the distribution of individual specific effects, not the distribution of observation-specific disturbances, which conditional on α and x is determined completely by the structural model g.

We still wish to estimate ψ by maximizing the sample analogue of (2.2), which leads to several concerns. First, because the explanatory variables x enter in both the distribution of unobserved effects and in the structural model along with θ, the model parameters may not be identified when T is fixed. Second, even when ψ is identified, maximizing the integrated likelihood may result in an inconsistent estimate of ψ or one that converges arbitrarily slowly. Both of these concerns have to do with the size of $\mathcal{H}$, and we address them in turn in the next two sections of the paper. In the next section, we introduce identification conditions that play similar roles to order and rank conditions for linear models. Given identification, estimation may proceed simply using the method of sieves. In section 4, we follow the sieve literature by replacing $\mathcal{H}$ with a sieve space $\mathcal{H}_n$ that becomes dense in $\mathcal{H}$ as n gets large.

3. Identification

The "true" model is defined by two conditional densities. The density of the dependent variable y conditional on the (observed) explanatory variable x and the unobserved, individual-specific effect α is $g(y|x, \alpha, \theta_0)$ and is referred to the structural model. We treat g as specified by the researcher up to a finite dimensional parameter θ that is common across individuals in the panel. The density of the unobserved effect conditional on the explanatory variable is $q_0(\alpha|x) \in \mathcal{H}$. For simplicity, we focus on the case where y is a T-vector for each individual. The conditional density of y given x, $f(y|x, \theta_0) \in \mathcal{F}$ is given by 2.1. We assume that $\int g dy = 1$ for each $(x, \alpha, \theta) \in \mathcal{X} \times \mathcal{A} \times \Theta$ and $\int q d\alpha = 1$ for each $x \in \mathcal{X}$, from which it follows that $\int f dy = 1$ for each $(x, \theta) \in \mathcal{X} \times \Theta$.

For identification, we make use of the standard high-level assumption employed in maximum likelihood estimation:[2]

Assumption 1. For $\theta \in \Theta$, $q \in \mathcal{H}$, and $\psi = (\theta, q)$, $\psi \neq \psi_0$ implies $\int_{\mathcal{A}} g(y|x, \alpha, \theta) q(\alpha|x) d\alpha \neq \int_{\mathcal{A}} g(y|x, \alpha, \theta_0) q_0(\alpha|x) d\alpha$.

This assumption is standard in the literature, and it is easy to verify that if it is satisfied the solution to the maximum likelihood problem defined in (2.2) will be unique. As with many high-level assumptions, Assumption 1 provides little intuition into the conditions that actually make identification feasible. In addition, identification will depend sensitively on the definition of $\mathcal{H}$ in many situations.

An alternate way to view the identification problem, which may be helpful to illustrate why restrictions on $\mathcal{H}$ will generally be necessary, is to view (2.1) as a nonlinear operator

[2]For random variables $h_1(Y, X)$ and $h_2(Y, X)$, let $h_1(Y, X) \neq h_2(Y, X)$ mean $Pr_{Y,X}(h_1(Y, X) \neq h_2(Y, X)) > 0$.

that maps a parameter θ and a function q into a conditional density $f(y|x)$ that is identified and consistently estimable from the observed data. Identification is therefore equivalent to asking for a unique solution to the nonlinear operator equation $v = h(u)$, where $u \in \Theta \times \mathcal{H}$ and $v \in \mathcal{F}$ where $\mathcal{H}$ is a space of functions that map $\mathcal{A} \times \mathcal{X}$ to $\mathbb{R}$ and $\mathcal{F}$ is a space of functions that map $\mathcal{Y} \times \mathcal{X}$ to $\mathbb{R}$. While finding primitive conditions for unique solvability of a nonlinear operator equation is extremely difficult, a necessary condition for invertability of $h(u)$ is that the domain of h be no larger than the range of h, which is intuitively an order type condition akin to those imposed in the classical analysis of linear simultaneous equations models. A minimal condition for this to be satisfied will be that $T > d_\alpha$, and in cases where $\mathcal{Y} = \mathbb{R}^T$ and $\mathcal{A} = \mathbb{R}^{d_\alpha}$, this condition will likely be sufficient for identification subject to sensibly parameterizing the model. However, in cases where the support of y is restricted, restrictions on the complexity of $\mathcal{H}$ in either the x direction, the α direction, or both will typically be required for identification to be possible.

To provide additional intuition into when Assumption 1 may be satisfied and to provide further insight into the types of restrictions on $\mathcal{H}$ that may be useful for identification, we provide several cases below in which more primitive and intuitive conditions for local identification may be stated. The key in each case will be an "order" type condition that guarantees that the space $\mathcal{H}$ is sufficiently simple relative to the space $\mathcal{F}$ to allow q_0 and θ_0 to be recovered from the system of equations defined in (2.1). The intuition is easiest to see when both y and x are discrete and finite, in which case the order condition reduces simply to a counting condition. In our paper, we characterize the space $\mathcal{H}$ in which θ_0 and q_0 may be point identified. In many cases, one will have to place substantial restrictions on the space $\mathcal{H}$ to achieve point identification in which case a bounds analysis may be more

appealing. On the other hand, when either x or y is continuous, the restrictions on $\mathcal{H}$ will generally be quite mild and one may prefer the approach considered in the present paper.

It is also worth noting that by asking for joint identification of both q_0 and θ_0 we may be asking for too much. It is well-known that there are a number of cases; e.g. Manski (1987) in the static binary choice model, Honoré and Kyriazidou (2000b) in dynamic binary choice, Honoré and Kyriazidou (2000a) in censored models, Horowitz and Lee (2004) in duration models, and Kyriazidou (1997) for selection models as well as linear, poisson, and logit models; for which θ_0 may be identified and estimated without reference to q_0; and in many cases, θ_0 and perhaps a set of functionals of q_0, which may potentially be identified independently of q_0, are the parameters of interest. One criticism of such approaches is that they are generally very problem specific, and it is unclear how they may be extended to other contexts.

The approach we consider also offers a justification to the use of flexible parametric random effects models as a simple approximation to the nonparametrically identified mixing distribution over α in cases where q_0 is identified. While it would be interesting to examine the general question of when θ_0 and relevant functionals of q_0 could be identified without identification of q_0, that lies beyond the scope of the present paper, though the conditions would almost certainly be weaker than those outlined below. It would also be interesting to examine the question of identification while relaxing the tight parametric specification of the structural likelihood; we also leave this for future research.

3.1. **Example 1. Discrete y and x.** For our first example, we consider the simplest case in which both y and x are discrete; i.e. $\mathcal{Y} = \{\mathrm{y}_1, \ldots \mathrm{y}_J\}$ and $\mathcal{X} = \{\mathrm{x}_1, \ldots \mathrm{x}_R\}$ with each y_j and x_r a T-vector. For simplicity, we further suppose that at each time period t, y_t may take

on M possible values so that $J = M^T$. For example, in the binary choice model with $T = 2$, we have $\mathcal{Y} = \{(0,0)', (0,1)', (1,0)', (1,1)'\}$. For $x = x_r$, we also have that the conditional distribution $f(y|x = x_r)$ is a J-vector, $\mathbf{f}_r$, whose jth element is $\Pr\big(y = \mathrm{y}_j | x = x_r, \theta\big)$.

The fact that conditional probabilities must sum to one imples that one element of $\mathcal{Y}$ will be redundant for each $x \in \mathcal{X}$, so equation (2.1) defines a system of $R(J-1)$ potentially nonredundant equations. Point identification of (θ_0, q_0) then requires that the solution to this system of equations for θ and q be unique. In this case, we can immediately see that identification of q will require substantial restrictions on $\mathcal{H}$. In particular, we can identify at most $R(J-1)$ parameters which obviously precludes identification of q outside of the class of distributions characterized uniquely by at most $R(J-1) - d_\theta$ parameters.

The discrete y case with no covariates or unknown common parameters, θ, was analyzed extensively in Lindsay (1983a, 1983b) who showed that the nonparametric maximum likelihood problem

$$\arg\max_q L(q) = E_{y,x}\left[\log \int_{\mathcal{A}} g(y|\alpha) q(\alpha) d\alpha\right] \tag{3.1}$$

is solved by J-point discrete distributions. Of course, one could also parameterize $q = q(\alpha|\gamma)$ with $\dim(\gamma) \leq J-1$ and, subject to rank conditions, identify γ. The discrete y and x case is also the basic framework considered in Chernozhukov, Hahn, and Newey (2004) who provide bounds on the set of θ that are identified from the data without imposing any restrictions on q though their results would undoubtedly generalize to models with continuous y or x.

The classic nonidentification result for the panel probit with $Pr(y_t = 1|x_t, \alpha, \theta) = \Phi(x_t'\theta + \alpha)$ of Chamberlain (1992) and Hahn (2001) with $T = 2$ and $\mathcal{X} = (0,1)$ also corresponds to this case. Looking at the model through the lens of random effects provides a similar

nonidentification result for panel binary choice when $T = 2$ and $R = 1$, namely that joint identification of θ_0 and q_0 will not be possible unless q_0 belongs to a known 2-parameter parametric class such as 3-point discrete distributions or a normal location-scale family.[3]

While we do not have nonparametric identification of q_0 in this case, it is worth considering further as it is simple and yet adds substantial intuition about the general problem. In particular, it illustrates the differences between increasing the complexity of $\mathcal{Y}$ and $\mathcal{X}$ and provides substantial intuition into why we will generally obtain nonparametric identification of q in the presence of continuous y or x.

We suppose initially that $R = 1$ (i.e., the explanatory variable is a constant) and that α is discrete with known support $\mathcal{A} = \{a_1, \dots a_K\}$. The distribution of unobserved effects is a K-vector π with kth element equal to $\Pr(\alpha = a_k)$. Equation (2.1) may then be rewritten as

$$\mathbf{f} = \mathrm{A}_\theta \pi$$

where A_θ is a $J \times K$ matrix with $[\mathrm{A}_\theta]_{jk} = \Pr\big(y = \mathrm{y}_j | x = \mathrm{x}^*, \alpha = a_k, \theta\big) = g(\mathrm{y}_j | \mathrm{x}^*, a_k, \theta)$. The parameter vector $\psi = (\theta, \pi)$ is now finite dimensional and must be recovered from the 'observed probabilities' in $\mathbf{f}$. Letting ι_K denote a K-vector of ones, note that we have two constraints, $\iota_K' \pi = 1$ and $\iota_J' \mathrm{A}_\theta = \iota_K'$, for each $\theta \in \Theta$ that together ensure the implied conditional probabilities for y given $x = \mathrm{x}^*$ sum to one, $\iota_J' \mathbf{f} = 1$.

We state conditions for local identification. For a given θ^*, define the $J \times d_\theta$ matrix D_{θ^*} as

$$D_{\theta^*} = \left[\left(\frac{\partial \mathrm{A}_\theta}{\partial \theta_1}\Big|_{\theta^*} \right) \pi_0, \dots, \left(\frac{\partial \mathrm{A}_\theta}{\partial \theta_{d_\theta}}\Big|_{\theta^*} \right) \pi_0 \right]$$

[3]This case is also interesting as it provides an illustration as to why asking for joint identification of θ and q may be too strong as it is well-known that the parameter θ_0 is identified in the panel logit in this situation without needing to identify q.

where each $\frac{\partial \mathrm{A}_\theta}{\partial \theta_j}$ is the matrix A_θ differentiated component by component with respect to the jth element of θ. The resulting matrix contains the derivatives of the structural model, g, averaged over the true distribution of unobserved effects.

Proposition 1. *Suppose that $\mathcal{A} = \{a_1, \ldots a_K\}$ where the a_j are known. Suppose that (i) $d_\theta + K \leq J$ and (ii) rank$([D_{\theta_0} \; B_{\theta_0}]) = d_\theta + K - 1$ where $B_\theta = A_\theta \begin{bmatrix} I_{K-1} \\ -\iota'_{K-1} \end{bmatrix}$. Then (θ_0, π_0) is identified in a neighborhood of θ_0 and π_0, $\mathcal{N}((\theta'_0, \pi'_0)')$.*

Under the assumed conditions and imposing the constraint that $\iota'_K \pi = 1$, the conclusion of Proposition 1 follows immediately from the implicit function theorem. The hypotheses in Proposition 1 are quite natural rank and order conditions. Notice that here the space $\mathcal{H}$ is finite dimensional by necessity: We cannot identify an infinite dimensional distribution of unobserved effects with only a finite number of possible observed (x, y) outcomes. In this case, we could state the analogue of Assumption 1 as a sufficient condition for global identification. However, finding simple sufficient primitive conditions for global identification would be difficult in general as such conditions would be equivalent to conditions guaranteeing unique solution of a general nonlinear system of equations.

This setting is simplistic as we are interested in cases where $\mathcal{H}$ is infinite dimensional, but provides several important insights. The operator A_θ describes how the probabilities of observable outcomes, f, change with a small change in the distribution of unobserved effects. The rank condition (ii) amounts to saying that in a neighborhood of (θ_0, π_0), changing θ affects the probabilities of observed (x, y) outcomes in a way that cannot be "canceled out" by perturbing π.

Although the length of the panel, T, does not appear explicitly in our assumptions, it becomes apparent how the length of the panel provides additional information when we

recall that $J = M^T$. For example, in the panel probit discussed above with $R = 1$ and $T = 2$, we can identify the q only if we are willing to assume that the support of α consists of at most three, known points. If, on the other hand, we had $T = 10$, we could potentially identify q with the number of support points of α equal to $2^{10} - 1 = 1023$. While this is certainly not equivalent to nonparametric identification, it does suggest that the restrictions that need to be imposed on q may be relaxed very quickly as either T or, more generally, the complexity of $\mathcal{Y}$ increases.

To generalize this result for $R > 1$, we stack the observed probabilities for y given each possible value of x into a $JR \times 1$ vector, $\mathbf{f} = (\mathbf{f}_1', \ldots, \mathbf{f}_R')'$, where each $\mathbf{f}_r$ is a J vector with jth element is $\Pr\big(y = \mathrm{y}_j | x = \mathrm{x}_r, \theta\big)$. We then have $\mathbf{f}_r = \mathrm{A}_{\theta,r}\pi_r$ where $[\mathrm{A}_{\theta,r}]_{jk} = g(\mathrm{y}_j|\mathrm{x}_r, a_k, \theta)$ and we may write $\mathbf{f} = \mathrm{A}_\theta \pi$ with A_θ a $JR \times KR$ block diagonal matrix.

We first consider the case where we are unwilling to restrict the dependence of q on x. In this case, we have a different set of probabilities associated with each x_r, π_r. Proposition 1 then holds with (i) and (ii) replaced by (i') $d_\theta + KR \leq JR$ and (ii') $\operatorname{rank}([D_{\theta_0}\ \mathrm{B}_{\theta_0}]) = d_\theta + (K-1)R$ where $\mathrm{B}_\theta = \begin{bmatrix} B_{\theta,1} \\ \vdots \\ B_{\theta,R} \end{bmatrix}$ and $\mathrm{B}_{\theta,r} = \mathrm{A}_{\theta,r} \begin{bmatrix} I_{K-1} \\ -\iota_{K-1}' \end{bmatrix}$. That is, we can identify θ_0 and a set of x_r specific probabilities $\pi_{r,0}$ where the number of support points for each x_r, K, satisfies $K \leq J - d_\theta/R$.

Now note that when the explanatory variable takes on multiple values one may potentially identify q which are more complicated in the α direction if one is willing to restrict the complexity of the dependence of q on X. As the most extreme case, one could assume that q does not depend on X. In this case, one could potentially identify q among the class of

$(R(J-1)-d_\theta+1)$-point discrete distributions. Alternatively, one could parameterize the probabilities associated with each mass point as $\pi_j(X,\gamma_j)$ where $\dim(\gamma_j)=S$ and identify q among the class of $\frac{R(J-1)-d_\theta+1}{S}$-point discrete distributions while allowing limited dependence of q on X. There are, of course, many other ways of trading off complexity of q in α and X that could be considered.

For concreteness, suppose again that we are in the binary choice model described above with $T=2$ and now suppose that the single x variable may take on 2 possible values in each time period. If we impose no smoothness on the dependence of q on x, we may identify a three-point distribution for α for each of our 4 potential x values. On the other hand, if we assume α does not depend on x, we may identify θ and potentially identify q among the class of 12-point discrete distributions. There are, of course, intermediate cases. For example, we could assume that $\mathcal{A}=\{a_1,...,a_4\}$ with $\pi_j=\Phi(\gamma_{0j}+x_1\gamma_{1j}+x_2\gamma_{2j})$, potentially identifying a four point distribution that allows some, though not arbitrary, dependence on X.

While the preceding discussion ignores the rank condition and focuses purely on the counting condition, it highlights the important features that will lead to nonparametric identification when either y or x is continuous. As the discussion suggests, we may quickly relax restrictions on q as we let the space $\mathcal{Y}$ become richer. Indeed, we will show that in an important class of models, θ_0 and $q_0(\alpha|x)$ are identified without imposing any smoothness of q in x when y is continuous. On the other hand, the discussion suggests that increasing the complexity of $\mathcal{X}$ when $\mathcal{Y}$ is finite may help us to identify $q_0(\alpha|x)$ in a rich space but that this will come at the cost of restricting the dependence of q on x.

3.2. **Parametric random effects.** We now consider random effects models where the distribution of unobserved effects may depend on x but has known functional form up to a finite

dimensional vector of parameters, $q(\alpha|x,\gamma)$. The argument here is similar to the previous section. Here, however, we have assumed that $\mathcal{H} = \mathcal{H}_\Gamma = \left\{q(\alpha|x,\gamma)\middle|\gamma \in \Gamma \subset \mathbb{R}^K\right\}$ is finite dimensional while x and y may be continuous variates. This case is a natural lead in to the more general case where we do allow $\mathcal{H}$ to be infinite dimensional. In particular, our strategy in the case where $\mathcal{Y}$ is finite but x is continuous will build heavily on the argument in this section.

We assume that the support of the unobservable α does not depend on γ and impose that q and the structural model g are densities, i.e., $\int_{\mathcal{A}} q(\alpha|\cdot,\gamma)d\alpha = 1$ for each $\gamma \in \Gamma$ and $\int g(y|\cdot,\alpha,\theta)dy = 1$ for each $(\alpha,\theta) \in \mathcal{A}\times\Theta$. Note that if $\mathcal{A}$ (or $\mathcal{Y}$) is discrete, q is a vector (A_θ is a matrix) and this restriction becomes $\iota' q = 1$ as in section 1. In this case we have a fully parametric problem, $\psi = (\theta,\gamma) \in \mathbb{R}^{d_\theta+K}$ with the parametric family of densities for the observables given by

$$f(y|x,\theta,\gamma) = \int_{\mathcal{A}} g(y|x,\alpha,\theta)\, q(\alpha|x,\gamma)d\alpha.$$

Similarly to the previous section, we may the identification question as asking when the nonlinear equation defined in (2.1) is uniquely solved at (θ_0,γ_0). Because the parameter vector ψ is finite dimensional, for local identification we need only that enough points $\{(x^*,y^*)\} \subset \mathcal{X}\times\mathcal{Y}$ can be found such that the matrix of derivatives of the integrated likelihood (2.1) evaluated at those points and $\theta = \theta_0, \gamma = \gamma_0$ is full rank. In particular, for a set of points $\left\{\mathrm{y}_j^*, \mathrm{x}_r^*\right\}$ with $j = 1,\ldots,J$ and $r = 1,\ldots,R$, define a $JR \times d_\theta + K$ matrix, $\mathcal{D}_{\psi^*}$, whose typical row is

$$\left[\int \left(\frac{\partial g}{\partial \theta_1}\Big|_{\theta^*}\right) q_0 d\alpha, \ldots, \int \left(\frac{\partial g}{\partial \theta_{d_\theta}}\Big|_{\theta^*}\right) q_0 d\alpha, \int g\left(\frac{\partial q}{\partial \gamma_1}\Big|_{\gamma^*}\right) d\alpha, \ldots, \int g\left(\frac{\partial q}{\partial \gamma_K}\Big|_{\gamma^*}\right) d\alpha\right]\Bigg|_{y_j^*, x_r^*}$$

Proposition 2. *Assume that $\mathcal{A}$ is a known subset of a finite dimensional Euclidean space. Let J and R be positive integers. Suppose that (i) $d_\theta + K \leq JR$ and (ii) there exist $\left\{y_j^*, x_r^*\right\}$ with $j = 1, \ldots, J$ and $r = 1, \ldots, R$ such that the matrix $\mathcal{D}_{\psi_0}$ is well-defined, continuous in (θ, γ), and has full column rank. Then ψ_0 is identified in $\mathcal{N}(\psi_0)$.*

As before, the conclusion of Proposition 2 follows immediately from the implicit function theorem. Assumption 1 obviously provides a sufficient condition for global identification that, given the parametric assumptions on g and q, could be verified on a model by model basis. The local identification condition here simply requires that the model has been sensibly parameterized such that none of the scores of the integrated likelihood are a linear combination of the others evaluated at ψ_0; that is, there are no redundant parameters.

Note that the observed y and x may be discrete or continuous so long as one can find enough points in $\mathcal{Y} \times \mathcal{X}$ to make the matrix $\mathcal{D}$ full rank. This condition simply says that varying x and y should produce variation in the likelihood. For example, if there were a single x variable that at ψ_0 did not enter the structural likelihood g, there would be no points x_r^* available and all identification would have to come from finding a sufficient number of y values. Also note that if y or x is continuous, then assumed continuity of g in y or x will guarantee that the condition above holds on a set of positive measure in $\{y, x\}$.

Our strategy below, when $\mathcal{H}$ is an infinite dimensional space of functions and y is finite, will be to consider basis function approximations of the form $q(\alpha|x) = \sum_{k=1}^{K} \gamma_k p_k(\alpha, x) + R_K$, where $\{p_k\}$ is a set of orthogonal polynomials or B-splines and R_K is a remainder term that goes to zero as K gets large. Proposition 2 may then be applied by considering a sequence of matrices $\mathcal{D}_K$, similar to the matrix $\mathcal{D}$ defined above, with strengthened conditions that prevent the matrices $\mathcal{D}_K$ from coming arbitrarily close to violating the rank condition as K

gets larger. This argument is developed in the next section. In particular, we show that in the presence of a continuous x variable, q may be identified in an infinite dimensional space $\mathcal{H}$ even when the number of possible outcomes for y is finite.

3.3. **General Case.** We now consider models where the space of distributions for unobserved effects is infinite dimensional but y is finite. We consider approximating $q \in \mathcal{H}$ by a sequence of finite dimensional approximations $q(\alpha|x) \approx \sum_{k=1}^{K_m} \gamma_k p_k(\alpha, x)$, where $\{p_k\}$ is a set of orthogonal polynomials or B-splines and the approximation error goes to zero as K_m gets large. For each K_m, the problem is a parametric random effects model as described in the previous section. When x is continuous with known compact support, identification follows from applying Proposition 2 to a sequence of matrices $\mathcal{D}_{K_m}$ along with additional regularity conditions on the eigenvalues of the sequence of matrices $\mathcal{D}_{K_m}$ and the rate at which the error of approximating $q(\alpha|x)$ by a K_m term series goes to zero.[4]

Define the functions

$$H_K^{(k)}(y, x, \theta) \quad = \quad \int_{\mathcal{A}} g(y|x, \alpha, \theta) p_k(\alpha, x)\, d\alpha \qquad k = 1, \ldots, K,$$

and let

$$D_\theta^{(j)}(y, x, \theta^*) = \int \left(\frac{\partial g}{\partial \theta_j}\Big|_{\theta^*} \right) q_0 d\alpha.$$

The function $H^{(k)}$ is the structural likelihood integrated against the kth basis function and is thus proportional to the derivative of the integrated likelihood (2.2) with respect to the kth series coefficient, γ_k. As in the previous section we require the integrated likelihood evaluated

[4]This approach is simply one way of reducing the complexity of $\mathcal{H}$ relative to the complexity of $\mathcal{F}$. There are, of course, numerous other avenues that could be pursued. One approach which we conjecture would be quite useful would be to restrict the dependence of q on X by imposing that X enter the model through an index.

at the true distribution of unobserved effects be differentiable in θ. Again take a set of points $\left\{\mathrm{y}_j^*, \mathrm{x}_r^*\right\}$ with $j = 1, \ldots, J_{K_m}$ and $r = 1, \ldots, R_{K_m}$, and define a $J_{K_m} R_{K_m} \times (d_\theta + K_m)$ matrix, $\mathcal{D}_{K_m}(\widetilde{\theta})$, whose typical row is

$$\begin{aligned}\left[\mathcal{D}_{K_m}(\widetilde{\theta})\right]_{jr} &= \left[D_\theta^{(1)}(\mathrm{y}_j^*, \mathrm{x}_r^*, \theta_{jr}), \ldots, D_\theta^{(d_\theta)}(\mathrm{y}_j^*, \mathrm{x}_r^*, \theta_{jr}), H_{K_m}^{(1)}(\mathrm{y}_j^*, \mathrm{x}_r^*, \widetilde{\theta}), \ldots, H_{K_m}^{(K_m)}(\mathrm{y}_j^*, \mathrm{x}_r^*, \widetilde{\theta})\right] \\ &= \left[\quad D_\theta(\mathrm{y}_j^*, \mathrm{x}_r^*, \theta_{jr})\,,\; H_{K_m}(\mathrm{y}_j^*, \mathrm{x}_r^*, \widetilde{\theta}) \quad\right] \qquad (3.2)\end{aligned}$$

where θ_{jr} is an intermediate value on the line between θ_0 and $\widetilde{\theta}$ such that $\int_{\mathcal{A}} g(y_j^*, x_r^*, \alpha, \widetilde{\theta}) q_0 d\alpha = \int_{\mathcal{A}} g(y_j^*, x_r^*, \alpha, \theta_0) q_0 d\alpha + D_\theta(y_j^*, x_r^*, \theta_{jr})(\widetilde{\theta} - \theta_0)$. For identification we require that $\mathcal{D}_{K_m}$ be full rank for each K_m and that the eigenvalues of $\mathcal{D}_{K_m}'\mathcal{D}_{K_m}$ go to zero sufficiently slowly. Note that $J_{K_m} \leq J - 1$, where J is the number of possible outcomes for y, while $R_{K_m} \to \infty$ as $K_m \to \infty$. Note that in a model with multiple x's that includes a single continuous x variable that satisfies the conditions of the proposition below, the argument of the proposition may be applied pointwise in the remaining x's.

Proposition 3. *Assume that $\mathcal{A} \subset \mathbb{R}^{d_\alpha}$ is known and compact, x is continuous, g and q are bounded away from zero and infinity, and g is continuously differentiable in θ. Let $\mathcal{H}'(\mathcal{A} \times \mathcal{X})$ be a separable Hilbert space of functions and $\mathcal{P} = \{p_k\}_{k=1}^{\infty}$ an orthonormal system in $\mathcal{H}'$. Let $\{K_m\}$ be a given nondecreasing sequence of positive integers. Suppose that, for each $m > 0$ and $\widetilde{\theta} \in \mathcal{N}(\theta_0)$: (i) there exist positive integers R_m and $J_m \leq J - 1$ such that $d_\theta + K_m \leq J_m R_m$, and (ii) on every set $E \subseteq \mathcal{Y} \times \mathcal{X}$ with $Pr_{X,Y}(E) = 1$ there exist points $\left\{y_j^*, x_r^*\right\}$ with $j = 1, \ldots, J_m$ and $r = 1, \ldots, R_m$ such that the matrix $\mathcal{D}_{K_m}(\widetilde{\theta})$ has full column rank with minimum eigenvalue $\min\left[\mathrm{eig}\left(\frac{1}{K_m}\mathcal{D}_{K_m}'\mathcal{D}_{K_m}\right)\right] \geq \delta_{K_m}$ where $\delta_{K_m} \to 0$ as $K_m \to \infty$. Suppose $q_0 \in \mathcal{H}$ where*

$$\mathcal{H} = \left\{ q \in \mathcal{H}' \;\middle|\; \int_{\mathcal{A}} q\, d\alpha = 1 \text{ and } \sup_{\alpha \in \mathcal{A}, x \in \mathcal{X}} |q - \gamma_K' p_K| = o(\delta_{K_m}^{-1/2}) \right\},$$

Then ψ_0 is identified in $\mathcal{N}(\theta_0) \times \mathcal{H}$.

The proposition is proven in the appendix. The argument is essentially the same as the argument applied in the previous section with one important modification: We impose that the uniform approximation error of function $q(\alpha|x)$ goes to zero faster than the eigenvalues of the matrix $\mathcal{D}_{K_m}$. That such an assumption would be necessary was heuristically motivated in Sections 3.1 and 3.2 where we saw that additional variation in x was helpful only if we were willing to impose restrictions on the complexity of the function $q(\alpha|x)$ in x. Intuitively, the assumption requires that we restrict the set of possible q to consist of functions that are smoother than the function g. Results for the assumption about the uniform approximation rates are readily available; see Newey (1997). For example, if q has s derivatives, it is well-known that the uniform approximation rate of $q(\alpha|x)$ by a power series is $O(K^{-s/(d_\alpha+T)})$.

The additional assumption on the eigenvalues may also be related to the smoothness and variation of the functions g and $\frac{\partial g}{\partial \theta_j}$, $j = 1, ..., d_\theta$. Intuitively, if these functions are very smooth in x, it will be hard to pick values of x that induce substantial variation in the rows of $\mathcal{D}_{K_m}$ which will evidence itself in rapid decay of the eigenvalues. In the extreme case where g and $\frac{\partial g}{\partial \theta_j}$ have power series approximations that terminate after a finite number of terms, say K, the problem reverts back to the parametric random effects case studied above; in particular, joint identification of q and θ will require that $\mathcal{H}$ consist of functions q with power series expansions that terminate after at most $K - d_\theta$ terms.

The smoothness condition bears some relation to other approaches to identification in nonlinear panel models which assume the existence of variables that are independent of the unobserved effect α. Butler and Louis (1997) shows that q and θ are identified with continuous x's if α is independent of all of the x's, which may be viewed as uniform approximability in the x direction by a constant. Similarly, Honoré and Lewbel (2002) and Lewbel (2005) use

a "special" regressor that is absolutely continuous, enters the model with non-zero coefficient, and is independent of α to achieve identification of θ in a panel binary choice model. Altonji and Matzkin (2005) suppose the existence of instrumental variables that are independent of the individual effects α to achieve identification of θ.

4. Estimation

Given identification of θ and q, there are many estimation strategies one could pursue. In this paper, we estimate the model parameters θ and q using sieve ML; see, for example, Gallant and Nychka (1987), Shen (1997), Chen and Shen (1998), and Ai and Chen (2003). We show that the sieve maximum likelihood estimator is consistent under fairly standard regularity conditions following arguments from Newey and Powell (2003). Specifically, we have

$$(\widehat{\theta}, \widehat{q}) = \arg \max_{(\theta,q)\in\Theta\times\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^{n} \ln \int_{\mathcal{A}} g(y_i|x_i,\alpha,\theta)q(\alpha|x_i)\mathrm{d}\alpha$$

where we assume data are i.i.d. across individuals and $\mathcal{H}_n$ is a sequence of approximating sieve spaces that contain progressively more flexible parametric models for the density q. Consistency of $(\widehat{\theta}, \widehat{q})$ is then obtained in the following result.

Proposition 4. *Let $\psi_0 = (\theta_0, q_0)$, and suppose that Assumption 1 is satisfied. Further suppose that (i) $\psi_0 \in \Psi = \Theta\times\mathcal{H}$ where Ψ is a compact subset of a space with a norm $\|\psi\| = \|\theta\| + \int_{\mathcal{A}} |q|d\alpha$, (ii) $\{y_i, x_i\}$ are i.i.d. across i, (iii) $\sup_{\psi\in\Psi} \mathrm{E}[|\ln \int_{\mathcal{A}} g(y_i|x_i,\alpha,\theta)q(\alpha|x_i)\mathrm{d}\alpha|] < \infty$, (iv) $\mathrm{E}[\sup_{\psi\in\Psi} |\frac{\partial}{\partial\theta} \ln \int_{\mathcal{A}} g(y_i|x_i,\alpha,\theta)q(\alpha|x_i)\mathrm{d}\alpha|] < \infty$, $\mathrm{E}[\sup_{\psi\in\Psi} |\int_{\mathcal{A}} g(y_i|x_i,\alpha,\theta)q(\alpha|x_i)\mathrm{d}\alpha|^{-1}] < \infty$, and $|g(y|x,\alpha,\theta)| < \Delta$ for all y, x, α, and θ, and (v) $\Theta\times\mathcal{H}_n$ are compact subsets of Ψ such that for any $\psi \in \Psi$ there exists $\widetilde{\psi} \in \Theta\times\mathcal{H}_n$ such that $\widetilde{\psi} \to \psi$. Then $\widehat{\psi} \xrightarrow{p} \psi_0$.*

Proposition 4 shows that the sieve ML estimates of (θ_0, q_0) are consistent under fairly standard regularity conditions; see, e.g., Newey and Powell (2003) or Ai and Chen (2003).

Condition (i) imposes compactness of the parameter space, and condition (ii) imposes that data are i.i.d. across i but does not imposes intertemporal restrictions within i. Assumptions (iii) and (iv) are analogous to standard dominance conditions. Condition (v) imposes conditions on the sieve space $\mathcal{H}_n$ and will be satisfied by common sets of approximating functions such as power series, splines, and fourier series; see Gallant and Nychka (1987) and Ai and Chen (2003).

As in Newey and Powell (2003), we give only consistency results. For some models, one may be able to adopt the general results of Shen (1997) and Chen and Shen (1998) for semiparametric sieve estimation to establish rates of convergence of the estimator of q and $\sqrt{n}$ convergence of θ. However, at this level of generality, it seems unlikely that one would be able to establish the $n^{1/4}$ rate of convergence of q used to derive $\sqrt{n}$ convergence of θ. In particular, the structural likelihood g will be extremely smooth in most commonly used models which would suggest very slow rates of convergence for the estimator of q.

The intuition for the difficulty is that the density of the observed data is obtained by integrating the object of interest, the density of α given x, against the structural likelihood g. When g is very smooth, as would be the case for many models of interest such as the panel probit, the integration of q against g effectively irons out many of the features of q. While we expect that broad features of the distribution of α given x may be readily distinguished, we expect that it will take a large amount of data to allow one to sort out more local sorts of behavior such as curvature and tail behavior. In this sense, the problem is very similar to standard deconvolution where, for example, it is well-known that rates of convergence for estimates of the density of v where $y = u + v$ and u is Gaussian are extremely slow.

In spite of potentially very slow convergence of estimates of q, it may still be possible to obtain rates for many quantities of interest to applied researchers. For example, on might be interested in the derivative of the probability that $y = 1$ given x with α held fixed with respect to x in a simple probit model integrated against the density of α given x: $\int_{\mathcal{A}} \theta\phi(x\theta + \alpha)q(\alpha|x)\mathrm{d}\alpha$ where ϕ is the standard normal density. Since these quantities are often defined as integrals of the conditional density of α against functions which are almost as smooth as g itself, it seems likely that one should be able to estimate these quantities quite well even if q itself can only be estimated very poorly. Simulation evidence presented in the following section also seems to support this conjecture. It does seem that further exploration of the asymptotic properties of the sieve ML estimator in this context as well as other potential estimation procedures for both θ, q, and marginal effects is an important direction for additional research.

5. Simulation Example

To illustrate the potential of flexible random effects models, we present results from a simple simulation example. In the simulation, we consider estimation of a common parameter, θ, and various average derivatives of interest. We also present some evidence regarding estimates of the density of unobserved effects, q, though given the arguments in the previous section it seems unlikely that these estimates will be well-behaved in reasonably sized samples.

For our simulations, we consider a panel probit with $T = 2$ and $N = 1000$. We generate the data as

$$y_{it} = \mathbf{1}[y_{it}^* > 0] \text{ where}$$

$$y_{it}^* = x_{it}\theta + \alpha_i + u_{it},$$

$$u_{it} \sim N(0,1),\ x_{it} \sim N(0,1),$$

and we consider two different specifications for α_i:

$$\alpha_i | x_{i1}, x_{i2} \sim 4\left[BETA\left(3 + 6\frac{\exp\{4\bar{x}\}}{1+\exp\{4\bar{x}\}}, 9 - 6\frac{\exp\{4\bar{x}\}}{1+\exp\{4\bar{x}\}}\right) - .5\right],\ \text{or} \tag{5.1}$$

$$\alpha_i | p_i, x_{i1}, x_{i2} \sim p_i N(.75 + .5x_{i1} + .5x_{i2}, 9/64) + (1-p_i)N(-.75 + .5x_{i1} - .5x_{i2}, 9/64) \tag{5.2}$$

where $p_i | x_{i1}, x_{i2} \sim BERNOULLI(\Phi(\frac{x_{i1}+x_{i2}}{3}))$ and $\Phi(\cdot)$ is the standard normal distribution function. We refer to the design where α_i is generated using (5.1) as the "beta design" and the design where α_i is generated using (5.2) as the "normal mixture design". In both cases, the designs were chosen to induce a nontrivial relationship of the distribution of α_i on the observed x's which should be challenging for the estimation procedure. In the beta design, the distribution of α_i is symmetric when $\bar{x} = 0$ and becomes increasingly right-skewed (left-skewed) as $\bar{x}$ becomes more negative (positive) though it always remains unimodal with support $[-2, 2]$. In the normal mixture design, the dependence of the distribution of α_i on x is far more complicated with some values of x leading to bimodality and others to essential unimodality. Different values of x also nontrivially affect many features of the shape of the distribution and the set of α's which occur with nontrivial probability. In all of the simulations, we set θ equal to one, and all results are based on 500 simulation replications.

We use sieve maximum likelihood to obtain the semi-parametric estimation results in the simulations. We specify the conditional density of α given x using a variant of the SNP basis of Gallant and Nychka (1987) and Gallant and Tauchen (1989). Specifically, we specify $q_K(\alpha|x) = \frac{[P_K((\alpha-m(x))/(\sqrt{2}s),x_1,x_2)]^2 \frac{1}{s}\phi((\alpha-m(x))/s)}{\int [P_K((\alpha-m(x))/(\sqrt{2}s),x_1,x_2)]^2 \frac{1}{s}\phi((\alpha-m(x))/s)\mathrm{d}\alpha}$ where $m(x) = \mu_0 + \mu_1 x_1 + \mu_2 x_2$, $\phi(\cdot)$ is the standard normal density function, and $P_K(u,v,w)$ is a polynomial of the form $P_K(u,v,w) = \sum_{j=1}^{K} \gamma_j p_{a_\alpha(j)}(u) p_{a_{x_1}(j)}(v) p_{a_{x_2}(j)}(w)$ with $p_m(u)$ the m^{th} term in an Hermite polynomial orthonormal with respect to weight function $\frac{1}{\sqrt{\pi}}\exp(-u^2)$ and $a_\alpha(j)$, $a_{x_1}(j)$, and $a_{x_2}(j)$ functions indicating which order term to use for term j in the series for α, x_1, and x_2 respectively.[5] In estimation, we thus estimate s, μ_0, μ_1, μ_2, and the K-dimensional vector of polynomial coefficients γ along with the common parameter θ. This parameterization is convenient in that it includes the familiar Chamberlain's (1980) random effects model which models the distribution of the unobserved effect given the x's as Gaussian with constant variance and mean that depends linearly on x as the leading term.

Estimation results are reported in Tables 1-3. For point estimation of θ and marginal effects, we consider estimates based on a standard probit that ignores unobserved heterogeneity (Probit), fixed effects probit (FE), bias-reduced fixed effects probit (FE-BC) using Bester and Hansen (2007), random effects assuming $\alpha \sim N(0,\sigma^2)$ (RE), Chamberlain's random effects which assumes $\alpha|X \sim N(\beta_1 x_1 + \beta_2 x_2, \sigma^2)$ (CRE), flexible RE SNP with linear terms (SNP-linear), flexible RE SNP with quadratic terms (SNP-quadratic), and flexible RE

[5]Note that this allows a closed form expression for the integral in the denominator of q_K.

SNP with cubic terms (SNP-cubic).[6] We use Gauss-Hermite quadrature to evaluate the integrated likelihood for all of the random effects estimators. For estimating marginal effects, we also consider the linear model where the binary outcome variable is regressed on x and individual specific dummy variables (OLS) and the linear model where the binary outcome variable is regressed on x and individual specific dummy variables in the subsample where the outcome is not constant over the two time periods (OLS-FE). When considering estimation of the conditional density of α, we include the infeasible kernel conditional density estimates, which treat values of α as observed, as a point of comparison (Kernel).

We first consider results for integrated mean square error (IMSE) of the conditional density estimate $\widehat{q}(\alpha|x_1, x_2)$ of $q(\alpha|x_1, x_2)$, $\int_{\mathcal{A}}(\widehat{q}(\alpha|x_1, x_2) - q_0(\alpha|x_1, x_2))^2 d\alpha$, for various values of x_1 and x_2 given in Table 1. Panels A and B in Table 1 respectively give mean and median IMSEs from the beta design, and Panels C and D report mean and median IMSEs from the normal mixture design. The first row of each panel reports results from an infeasible kernel density estimator which uses the true values of α.[7] Not surprisingly, this infeasible estimator clearly dominates all estimators that use only the information available in the observed data for y and x. Using this as a benchmark, the estimates obtained from any random effects specification perform quite poorly. Comparing across the feasible estimators, we see that the normal random effects model that assumes α is independent from x is clearly dominated by all of the specifications that allow for dependence between α and x. We also

[6] "Linear" SNP has three polynomical terms with $a_\alpha(j) = 1$ if $j = 1$ and $a_\alpha(j) = 0$ otherwise, $a_{x_1}(j) = 1$ if $j = 2$ and $a_{x_1}(j) = 0$ otherwise, and $a_{x_2}(j) = 1$ if $j = 3$ and $a_{x_2}(j) = 0$ otherwise. "Quadratic" SNP has nine polynomial terms with the a functions producing all first and second order terms including interactions, and "Cubic" SNP has 19 polynomial terms with the a functions producing all first, second, and third order terms including interactions.

[7] We used a normal kernel and the bandwidth given by Silverman's rule of thumb.

see that CRE performs quite well relative to other feasible estimators. This is unsurpising in the beta design where one anticipates that a large amount of data would be necessary to pick up the skewness and difference in tails between the normal and actual distribution of α given x after being smoothed against products of normal CDF's. In this case, CRE offers a parsimonious specification that should be able to capture the mean behavior of α given x and so produce sensible estimates with small amounts of data. In the mixture of normals design, the results are more mixed. With x_1 and x_2 negative, the distribution of α given x is distinctly bimodal, and CRE fares poorly relative to the SNP approximations. This is offset for positive values of x where the distribution is better approximated by a Gaussian distribution. In this case, we see that the nonparametric density estimate does pick up some of this behavior but again suspect that a very large sample would be necessary to generate reliable distributional estimates.

Results for estimates of the common parameter θ and marginal effects for the beta design and normal mixture design are given respecitively in Tables 2 and 3. We report results for estimating the average marginal effect, $\int \int_{\mathcal{A}} \theta\phi(x\theta+\alpha)q(\alpha|x)\mathrm{d}\alpha\mathrm{d}F(x)$, as well as estimates of marginal effects at specific x-values, $\int_{\mathcal{A}} \theta\phi(x\theta+\alpha)q(\alpha|x)\mathrm{d}\alpha$. For OLS and OLS-FE, we estimate all effects with the slope coefficient on x from the fixed effects regression and note that this is commonly done in practice to estimate the average marginal effect but is clearly inappropriate for effects evaluated at given x values. For the fixed effects estimates, we estimate the average marginal effect as $\frac{1}{NT}\sum_{i:y_1\neq y_2}\sum_t \widehat{\theta}\phi(\widehat{\theta}x_{it}+\widehat{\alpha}_i)$ and the average marginal effect at a specific x-value as $\frac{1}{N}\sum_{i:y_1\neq y_2}\widehat{\theta}\phi(\widehat{\theta}x+\widehat{\alpha}_i)$. Since the latter effect is implicitly averaged against the marginal distribution of α it will also be inappropriate. We use the bias-reduction in Bester and Hansen (2007) to bias-correct these estimates of the marginal effects

for FE-BC. We calculate the marginal effect at a specific x-value as $\int_{\mathcal{A}} \widehat{\theta}\phi(x\widehat{\theta}+\alpha)\widehat{q}(\alpha|x)\mathrm{d}\alpha$ and the average marginal effect as $\frac{1}{NT}\sum_i \sum_t \int_{\mathcal{A}} \widehat{\theta}\phi(x_i t\widehat{\theta}+\alpha)\widehat{q}(\alpha|x_i)\mathrm{d}\alpha$.

The results for common parameters and marginal effects are much more encouraging than the results for estimating the distribution of unobserved effects. For the common parameters, we see that the fixed effects based approaches do very poorly. This poor performance is due to the incidental parameters problem and the very short time series dimension. We also see that the bias-reduction method does little to improve the performance, again because the time series is too short. The results also clearly show that estimation methods that ignore the correlation between the unobserved individual specific heterogeneity and the observables, Probit and RE, do poorly relative to the random effects approaches that allow for correlation between α and x. For the two designs we consider, we see that all of the random effects specifications that allow for correlation between α and x do a good job in estimating the common parameters. In all cases, the biases are less than 5%, and the correlated random effects estimates clearly dominate all other estimation procedures considered in terms of root mean squared error (RMSE) or median absolute error (MAD).

The marginal effects tell a similar story. We see that the correlated random effects estimators uniformly dominate all other estimators considered, though OLS performs similarly to the correlated random effects estimators for estimating the average marginal effect in the beta design. All of the correlated random effects perform quite well for estimating the average marginal effect. They appear to be approximately unbiased for this parameter and have dispersion comparable to OLS. For estimating the marginal effects at specific x-values, it appears that there may be a slight gain to using the SNP-based methods relative to CRE. For these parameters, the shape of the density of α given x at the x value of interest is

relatively more important than for average marginal effects and the more flexible models appear to do a better job at capturing these shapes, though in no case do they substantially outperform CRE in the data generating processes and sample sizes we have considered. We conclude by noting that the correlated random effects models appear to do an excellent job in estimating the common parameter of the model and marginal effects despite the poor performance of the correlated RE models in estimating $q(\alpha|x)$. We believe that this behavior warrants further study.

6. Conclusion

In this paper, we have considered identification in a flexible correlated random effects estimator for likelihood-based panel models with unobserved individual specific heterogeneity. We have shown that identification of common model parameters and the density of unobserved heterogeneity conditional on observed covariates is possible under restrictions on the space of density functions for unobserved heterogeneity. The set of sufficient conditions we impose do not restrict the density of unobserved heterogeneity given observables to a parametric family provided the observed data have sufficient variation. The restrictions will however impose substantial smoothness on the density of unobserved heterogeneity in most common models employed in applications. We also provide sufficient conditions for consistency of a sieve maximum likelihood for our correlated random effects models and provide simulation evidence regarding the performance of this estimator in a panel probit with two time periods. The simulation results show that the correlated random effects performs very well and may significantly outperform fixed effects or bias-reduced fixed effects in short panels.

Appendix A. Proof of Proposition 3.

Let $\|A\| = (\max \text{eig} A^* A)^{1/2}$ denote the spectral norm of a matrix A.

Consider $\widetilde{q} \in \mathcal{H}$ and $\widetilde{\theta} \in \mathcal{N}(\theta_0)$ with $(\widetilde{\theta}, \widetilde{q}) \neq (\theta_0, q_0)$ and suppose that $\int_{\mathcal{A}} g(y|x,\alpha,\widetilde{\theta})\widetilde{q}(\alpha|x)d\alpha = \int_{\mathcal{A}} g(y|x,\alpha,\theta_0)q_0(\alpha|x)d\alpha$. It then follows that

$$
\begin{aligned}
0 &= \int_{\mathcal{A}} \left[g(y|x,\alpha,\widetilde{\theta}) - g(y|x,\alpha,\theta_0) \right] q_0(\alpha|x)d\alpha + \int_{\mathcal{A}} g(y|x,\alpha,\widetilde{\theta}) \left[\widetilde{q}(\alpha|x) - q_0(\alpha|x) \right] d\alpha \\
&= \sum_{j=1}^{d_\theta} \left[\int_{\mathcal{A}} \frac{\partial g(y|x,\alpha,\theta)}{\partial \theta_j} |_{\bar{\theta}} q_0(\alpha|x)d\alpha \right] \left(\widetilde{\theta}_j - \theta_{0j} \right) + \sum_{k \leq K_m} \left[\int_{\mathcal{A}} g(y|x,\alpha,\widetilde{\theta}) p_k(\alpha,x)d\alpha \right] (\widetilde{\gamma}_k - \gamma_{0k}) \\
&\qquad + \sum_{k > K_m} \left[\int_{\mathcal{A}} g(y|x,\alpha,\widetilde{\theta}) p_k(\alpha,x)d\alpha \right] (\widetilde{\gamma}_k - \gamma_{0k})
\end{aligned}
$$

over a set $E \subseteq \mathcal{Y} \times \mathcal{X}$ with $\Pr_{XY}(E) = 1$ where $\bar{\theta}$ is an intermediate value that will differ for each value of y and x.

By assumption, we can define a sequence of positive numbers $\{\epsilon_m\}$ with $\epsilon_m \to 0$ such that for K_m, $\sum_{k>K_m} p_k(\alpha,x)\gamma_{0k} < \epsilon_m/2$ and $\sum_{k>K_m} p_k(\alpha,x)\widetilde{\gamma}_{0k} < \epsilon_m/2$. By assumption, we may also find positive integers R_m and $J_m \leq J-1$ such that $d_\theta + K_m \leq J_m R_m$ and points $\left\{ \mathrm{y}_j^*, \mathrm{x}_r^* \right\} \in E$ with $j = 1, \ldots, J_m$ and $r = 1, \ldots, R_m$ such that the matrix $\mathcal{D}_{K_m}(\widetilde{\theta})$ defined in (3.2) in the text has full column rank.

Let b_m be a column vector defined with jr^{th} entry given by

$$
[b_m]_{jr} = \sum_{k > K_m} \left[\int_{\mathcal{A}} g(y_j|x_r,\alpha,\widetilde{\theta}) p_k(\alpha,x_r)d\alpha \right] (\widetilde{\gamma}_k - \gamma_{0k}) \,.
$$

and $\psi_m = \begin{pmatrix} \theta \\ \gamma_{K_m} \end{pmatrix}$ be a column vector. Then $0 = \int_{\mathcal{A}} \left[g(y|x,\alpha,\widetilde{\theta}) - g(y|x,\alpha,\theta_0) \right] q_0(\alpha|x)d\alpha + \int_{\mathcal{A}} g(y|x,\alpha,\widetilde{\theta}) \left[\widetilde{q}(\alpha|x) - q_0(\alpha|x) \right] d\alpha$ may be written as

$$
-b_m = \mathcal{D}_{K_m}(\widetilde{\theta}) \left(\widetilde{\psi}_m - \psi_{0m} \right)
$$

Since $\mathcal{D}_{K_m}$ is full rank, we may solve this for

$$
\widetilde{\psi}_m - \psi_{0m} = -(\mathcal{D}'_{K_m} \mathcal{D}_{K_m})^{-1} \mathcal{D}_{K_m} b_m.
$$

It then follows that

$$\begin{aligned}\|\widetilde{\psi}_m - \psi_{0m}\| &= \|(\mathcal{D}'_{K_m}\mathcal{D}_{K_m}/K_m)^{-1}\mathcal{D}_{K_m}b_m/K_m \\ &\leq \|(\mathcal{D}'_{K_m}\mathcal{D}_{K_m}/K_m)^{-1}\mathcal{D}_{K_m}/K_m^{1/2}\|\|b_m/K_m^{1/2}\| \\ &= \sqrt{\max \mathrm{eig}(\mathcal{D}'_{K_m}\mathcal{D}_{K_m}/K_m)^{-1}}\|b_m/K_m^{1/2}\| \\ &\leq C\epsilon_m/\delta_{K_m}^{1/2}. \end{aligned} \tag{A.1}$$

$\|\widetilde{\psi}_m - \psi_{0m}\| \to 0$ as $m \to \infty$ then follows from (A.1) as long as $\epsilon_m = o(\delta_{K_m}^{-1/2})$ which will be satisfied as long as $q \in \mathcal{H}$ from our definition of $\mathcal{H}$. However, $\|\widetilde{\psi}_m - \psi_{0m}\| \to 0$ contradicts $(\widetilde{\theta}, \widetilde{q}) \neq (\theta_0, q_0)$. It then follows that $\int_{\mathcal{A}} g(y|x,\alpha,\widetilde{\theta})\widetilde{q}(\alpha|x)d\alpha = \int_{\mathcal{A}} g(y|x,\alpha,\theta_0)q_0(\alpha|x)d\alpha$ implies $(\widetilde{\theta}, \widetilde{q}) = (\theta_0, q_0)$. ■

Appendix B. Proof of Proposition 4.

We establish Proposition 4 by verifying the conditions of Newey and Powell (2003) Lemma A1 (NP.A1). Assumption 1 guarantees that condition (i) of NP.A1 is satisfied, and condition (v) of Proposition 4 gives NP.A1 condition (iii). It only remains to be shown that condition (ii) of NP.A1 is satisfied which can be done by verifying the conditions of Lemma A2 of Newey and Powell (2003) (NP.A2).

Condition (i) of Proposition 4 immediately gives condition (i) of NP.A2. Condition (ii) of NP.A2 follows from a usual law of large numbers for independent data under conditions (ii) and (iii) of Proposition 4. To verify condition (iii) of NP.A2, we consider two different values of ψ, say $\psi_1 = (\theta_1, q_1)$ and $\psi_2 = (\theta_2, q_2)$. We then have the absolute difference in objective function evaluated at these two values given by

$$D(\psi_1, \psi_2) = |\frac{1}{n}\sum \ln \int g_i(\theta_1)q_{i1} - \frac{1}{n}\sum \ln \int g_i(\theta_2)q_{i2}|$$

where we write $g(y_i|x_i,\alpha,\theta) = g_i(\theta)$ and $q(\alpha|x_i) = q_i$ and note that all integrals are over the set $\mathcal{A}$ and against the argument α. Let $M_{1i} = \sup_{\psi\in\Psi} |\frac{\partial}{\partial\theta} \ln \int g_i(\theta_1)q_{i1}|$ and $M_{2i} = \sum_{\psi\in\Psi} |\int g_i(\theta_1)q_i|^{-1}$. It is then straightforward to demonstrate that

$$\begin{aligned} D(\psi_1, \psi_2) &\leq \left(\frac{1}{n}\sum M_{1i}\right)\|\theta_1 - \theta_2\| + \left(\frac{1}{n}\sum M_{2i}\right)|\int g_i(\theta_2)(q_{i1} - q_{i2})| \\ &\leq \left(\frac{1}{n}\sum M_{1i}\right)\|\theta_1 - \theta_2\| + \Delta\left(\frac{1}{n}\sum M_{2i}\right)\int |(q_{i1} - q_{i2})| \end{aligned}$$

where the second inequality is from condition (iv) of Proposition 4. Also, from conditions (ii) and (iv) of Proposition 4 and a standard law of large numbers, $\frac{1}{n}\sum M_{1i} = O_p(1)$ and $\frac{1}{n}\sum M_{2i} = O_p(1)$, and it follows immediately that $D(\psi_1, \psi_2) < C(z)\|\psi_1 - \psi_2\|$ where $C(z) = O_p(1)$. Thus, condition

(iii) of NP.A2 is satisfied from which condition (ii) of NP.A1 follows. We then conclude from NP.A1 that $\widehat{\psi} - \psi_0 \xrightarrow{p} 0$. ■

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Altonji, J. G., and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053–1102.

Arellano, M., and J. Hahn (2005): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," Invited Lecture, Econometric Society World Congress, London.

Bester, A. C., and C. Hansen (2007): "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects," forthcoming *Journal of Business and Economic Statistics*.

Butler, S. M., and T. A. Louis (1997): "Consistency of Maximum Likelihood Estimators in General Random Effects Models for Binary Data," *Annals of Statistics*, 25(1), 351–377.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225–238.

——— (1992): "Binary Response Models for Panel Data: Identification and Information," Manuscript.

Chen, X., and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

Chernozhukov, V., J. Hahn, and W. Newey (2004): "Bound Analysis in Panel Models with Correlated Random Effects," mimeo.

Gallant, A. R., and D. W. Nychka (1987): "Seminonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

Gallant, A. R., and G. Tauchen (1989): "Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications," *Econometrica*, 57, 1091–1120.

Hahn, J. (2001): "The Information Bound of a Dynamic Panel Logit Model with Fixed Effects," *Econometric Theory*, 17, 913–932.

Hahn, J., and G. Kuersteiner (2004): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," Mimeo.

Hahn, J., and G. M. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both N and T Are Large," *Econometrica*, 70(4), 1639–1657.

Hahn, J., and W. K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Hausman, J. A., and W. E. Taylor (1981): "Panel Data and Unobservable Individual Effects," *Econometrica*, 49(6), 1377–1398.

Honoré, B. E., and E. Kyriazidou (2000a): "Estimation of Tobit-Type Models with Individual Specific Effects," *Econometric Reviews*, 19(3), 341–366.

——— (2000b): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

Honoré, B. E., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," *Econometrica*, 70, 2053–2063.

Honoré, B. E., and E. Tamer (2006): "Bounds on the Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–632.

Horowitz, J. L., and S. Lee (2004): "Semiparametric Estimation of a Panel Data Proportional Hazard Model with Fixed Effects," *Journal of Econometrics*, 119(1), 155–198.

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

Lewbel, A. (2005): "Simple Endogenous Binary Choice and Selection Panel Model Estimators," mimeo.

Lindsay, B. G. (1983a): "The Geometry of Mixture Likelihoods: A General Theory," *The Annals of Statistics*, 11(1), 86–94.

——— (1983b): "The Geometry of Mixture Likelihoods, Part II: The Exponential Family," *The Annals of Statistics*, 11(3), 783–792.

Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55(2), 357–362.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Neyman, J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1–32.

Shen, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.

Woutersen, T. (2005): "Robustness against Incidental Parameters and Mixing Distributions," Mimeo.

Table 1. Simulation Results. IMSE for Estimates of
Density of Individual Specific Effects Given x_1 and x_2

Estimator	$x_1 = 0, x_2 = 0$	$x_1 = 1, x_2 = 1$	$x_1 = .5, x_2 = .5$	$x_1 = -1, x_2 = -1$	$x_1 = -.5, x_2 = -.5$
		A. Mean IMSE for Beta Design			
Kernel	0.021	0.022	0.019	0.020	0.020
RE	0.021	0.535	0.362	0.535	0.362
CRE	0.020	0.079	0.071	0.077	0.072
SNP-linear	0.130	0.160	0.119	0.171	0.122
SNP-quadratic	0.169	0.157	0.143	0.162	0.125
SNP-cubic	0.142	0.191	0.140	0.196	0.127
		B. Median IMSE for Beta Design			
Kernel	0.019	0.018	0.016	0.015	0.017
RE	0.014	0.520	0.346	0.526	0.351
CRE	0.009	0.069	0.060	0.067	0.060
SNP-linear	0.117	0.131	0.086	0.141	0.088
SNP-quadratic	0.222	0.100	0.110	0.111	0.093
SNP-cubic	0.087	0.111	0.083	0.118	0.083
		C. Mean IMSE for Normal Mixture Design			
Kernel	0.030	0.054	0.040	0.113	0.035
RE	0.086	0.446	0.230	0.345	0.128
CRE	0.086	0.298	0.202	0.337	0.090
SNP-linear	0.183	0.326	0.329	0.198	0.061
SNP-quadratic	0.208	0.357	0.331	0.222	0.068
SNP-cubic	0.208	0.364	0.331	0.309	0.103
		D. Median IMSE for Normal Mixture Design			
Kernel	0.029	0.052	0.039	0.108	0.033
RE	0.083	0.439	0.224	0.346	0.128
CRE	0.083	0.291	0.196	0.338	0.090
SNP-linear	0.187	0.321	0.350	0.180	0.046
SNP-quadratic	0.200	0.330	0.354	0.188	0.044
SNP-cubic	0.203	0.344	0.340	0.235	0.053

Table 2. Simulation Results for Beta Design
Common Parameter and Marginal Effects

Estimator	Mean Bias	Median Bias	RMSE	MAD
A. Common Parameter (θ)				
Probit	0.242	0.240	0.247	0.240
FE	1.066	1.051	1.088	1.051
FE-BC	1.019	1.006	1.041	1.006
RE	0.431	0.423	0.439	0.423
CRE	0.012	0.008	0.085	0.056
SNP-linear	0.005	0.000	0.079	0.052
SNP-quadratic	0.015	0.011	0.084	0.055
SNP-cubic	0.059	0.048	0.110	0.062
B. Average Marginal Effect				
OLS	-0.003	-0.004	0.011	0.008
OLS-FE	0.229	0.228	0.229	0.228
Probit	0.117	0.117	0.117	0.117
FE	-0.075	-0.075	0.075	0.075
FE-BC	-0.103	-0.098	0.106	0.098
RE	0.109	0.109	0.109	0.109
CRE	-0.001	-0.001	0.011	0.007
SNP-linear	-0.003	-0.003	0.011	0.007
SNP-quadratic	-0.005	-0.005	0.012	0.008
SNP-cubic	-0.004	-0.004	0.012	0.008
C. Marginal Effect: x_1=0.0, x_2=0.0				
OLS	-0.157	-0.158	0.158	0.158
OLS-FE	0.075	0.074	0.076	0.074
Probit	0.147	0.147	0.149	0.147
FE	0.382	0.381	0.387	0.381
FE-BC	0.368	0.368	0.373	0.368
RE	0.117	0.117	0.119	0.117
CRE	-0.002	-0.003	0.022	0.015
SNP-linear	-0.013	-0.012	0.030	0.020
SNP-quadratic	-0.014	-0.015	0.035	0.022
SNP-cubic	-0.005	-0.002	0.042	0.025
D. Marginal Effect: x_1=0.5, x_2=0.5				
OLS	0.004	0.003	0.011	0.007
OLS-FE	0.236	0.235	0.236	0.235
Probit	0.221	0.221	0.222	0.221
FE	0.514	0.509	0.517	0.509
FE-BC	0.502	0.498	0.505	0.498
RE	0.206	0.205	0.206	0.205
CRE	0.032	0.032	0.036	0.032
SNP-linear	0.016	0.016	0.026	0.017
SNP-quadratic	0.005	0.007	0.040	0.020
SNP-cubic	0.008	0.010	0.047	0.028

Table 3. Simulation Results for Normal Mixture Design
Common Parameter and Marginal Effects

Estimator	Mean Bias	Median Bias	RMSE	MAD
A. Common Parameter (θ)				
Probit	-0.086	-0.087	0.095	0.087
FE	1.038	1.011	1.060	1.011
FE-BC	0.991	0.965	1.014	0.965
RE	0.232	0.232	0.249	0.232
CRE	-0.027	-0.032	0.089	0.061
SNP-linear	0.037	0.033	0.100	0.061
SNP-quadratic	0.047	0.040	0.109	0.063
SNP-cubic	0.058	0.055	0.120	0.074
B. Average Marginal Effect				
OLS	-0.013	-0.013	0.017	0.014
OLS-FE	0.242	0.241	0.243	0.241
Probit	0.078	0.078	0.078	0.078
FE	-0.069	-0.069	0.069	0.069
FE-BC	-0.087	-0.084	0.089	0.084
RE	0.063	0.063	0.064	0.063
CRE	-0.007	-0.008	0.013	0.009
SNP-linear	-0.002	-0.003	0.012	0.007
SNP-quadratic	-0.001	-0.002	0.013	0.009
SNP-cubic	-0.001	-0.002	0.014	0.009
C. Marginal Effect: x_1=0.0, x_2=0.0				
OLS	-0.114	-0.115	0.115	0.115
OLS-FE	0.141	0.140	0.142	0.140
Probit	0.073	0.072	0.075	0.072
FE	0.416	0.412	0.420	0.412
FE-BC	0.403	0.399	0.407	0.399
RE	0.038	0.037	0.042	0.037
CRE	-0.020	-0.020	0.027	0.022
SNP-linear	-0.003	-0.003	0.022	0.015
SNP-quadratic	0.004	0.002	0.041	0.025
SNP-cubic	0.004	0.009	0.055	0.035
D. Marginal Effect: x_1=0.5, x_2=0.5				
OLS	-0.013	-0.014	0.017	0.014
OLS-FE	0.241	0.241	0.242	0.241
Probit	0.137	0.136	0.137	0.136
FE	0.502	0.496	0.505	0.496
FE-BC	0.490	0.484	0.493	0.484
RE	0.110	0.109	0.111	0.109
CRE	0.024	0.024	0.029	0.024
SNP-linear	0.023	0.024	0.035	0.025
SNP-quadratic	0.014	0.012	0.040	0.025
SNP-cubic	0.014	0.015	0.054	0.037