A Simple Nonparametric Approach to Estimating the Distribution of Random Coefficients in Structural Models

Jeremy T. Fox
University of Michigan & NBER

Kyoo il Kim
University of Minnesota

July 2011

Abstract

We explore a nonparametric mixtures estimator for recovering the joint distribution of random coefficients in economic models. The estimator is based on linear regression subject to linear inequality constraints and is computationally attractive compared to alternative, nonparametric estimators. We provide conditions under which the estimated distribution function converges to the true distribution in the weak topology on the space of distributions. We verify the consistency conditions for discrete choice, continuous outcome and selection models.

1 Introduction

Economic researchers often work with models where the parameters are heterogeneous across the population. A classic example is that consumers may have heterogeneous preferences over a set of product characteristics in an industry with differentiated products. These heterogeneous parameters are often known as random coefficients. When working with cross sectional data, the goal is often to estimate the distribution of random coefficients. This distribution captures the essential heterogeneity that is key to explaining the economic phenomena under study.

This paper studies estimating the distribution $F(\beta)$ in the model

$$P_A(x) = \int g_A(x, \beta)\, dF(\beta), \tag{1}$$

where A is some subset of the observable outcomes y, x is a vector of covariates, β is the vector of random coefficients, and $g_A(x, \beta)$ is the probability that an outcome in A occurs for an observation with random coefficients β and covariates x. Given this structure, $P_A(x)$ is the cross sectional probability of observing an outcome in the set A when the covariates are observed to be x. The researcher picks $g_A(x, \beta)$ as the underlying model, has an i.i.d. sample of observations (y_i, x_i), and wishes to estimate $F(\beta)$.

Previous so-called nonparametric or flexible estimators for $F(\beta)$ include the EM algorithm, Markov Chain Monte Carlo (MCMC), simulated maximum likelihood, simulated method of moments, and minimum distance. As typically implemented, these estimators suffer from a key flaw: they are computationally challenging. The researcher must code some iterative search or simulation procedure. Often, convergence may not be to the preferred, global solution. Convergence may be hard to ensure and to detect.

Our insight is to notice that the unknown distribution $F(\beta)$ enters (1) linearly. Thus, we can exploit linearity and achieve a computationally simpler estimator than the alternatives. Fox, Kim, Ryan and Bajari (2011), henceforth FKRB, propose dividing the support of β into a finite and known grid of points $\beta^1, \ldots, \beta^R$. Let y_A equal 1 when an outcome is in A, and 0 otherwise. The researcher then estimates the weights $\theta^1, \ldots, \theta^R$ on the R grid points as the linear probability model regression of y_A on the R predicted probabilities $g(x, \beta^r)$. We also impose the constraints that each $\theta^r \geq 0$ and that $\sum_{r=1}^{R} \theta^r = 1$. Thus, the estimator of the distribution $F(\beta)$ with N observations and R grid points becomes

$$\hat{F}_N(\beta) = \sum\nolimits_{r=1}^{R} \hat{\theta}^r 1\left[\beta^r \leq \beta\right]$$

where $\hat{\theta}^r$'s denote estimated weights. The computational advantages of the procedure are immediate. Computationally, the estimator is linear regression (least squares) subject to linear inequality and equality constraints. This optimization problem is globally convex and specialized routines (such as LSSOL, built into MATLAB as the command lsqlin) are guaranteed to converge to the global optimum.

FKRB highlight the practical usefulness of the estimator by showing how it can be used in a series of examples of random coefficient models employed in consumer choice and other settings, including applications with endogenous covariates handled with instruments and applications with aggregate data. The estimator also has computational savings for complex structural models where economic models must be solved as part of the estimation procedure. In a dynamic programming application such as adding random coefficients to Rust (1987), the dynamic programming model must be solved only R times, once for each random coefficient β^r. These solutions occur before optimization commences, and are not nested inside an iterative search or simulation procedure. This contrasts with competing approaches, where multiple dynamic programs must be solved for every change in the estimation algorithm's guess of $F(\beta)$. In this respect, our estimator shares some computational advantages with the parametric approach in Ackerberg (2009).

A serious limitation is that the analysis in FKRB assumes that the R grid points used in a finite sample are indeed the true grid points that take on nonnegative support in the true $F_0(\beta)$. Thus, the true distribution $F_0(\beta)$ is assumed to be known up to a finite number of weights $\theta^1, \ldots, \theta^R$. This assumption is convenient as the estimator is consistent under standard conditions for the consistency of least squares under inequality and equality constraints (Andrews 2002). As economists often lack convincing economic rationales to pick one set of grid points over another, assuming that the researcher knows the true distribution up to finite weights is unrealistic.

This paper seeks to place this appealing, computationally simple estimator on firmer theoretical ground. Instead of assuming that the distribution is known up to weights $\theta^1, \ldots, \theta^R$, we require the true distribution $F_0(\beta)$ to satisfy much weaker restrictions. In particular, the true $F_0(\beta)$ can have any of continuous, discrete and mixed continuous and discrete supports. The prior approach in FKRB is parametric as the true weights $\theta^1, \ldots, \theta^R$ lie in a finite-dimensional subset of a real space. Here, the approach is nonparametric as the true $F_0(\beta)$ is known to lie only in the infinite-dimensional space of multivariate distributions on the space of random coefficients β.

In a finite sample of N observations, our estimator is still implemented by choosing a grid of points $\theta^1, \ldots, \theta^R$, ideally to trade off bias and variance in the estimate $\hat{F}_N(\beta)$. We, however, recognize that as the sample increases, R and thus the fineness of the grid of points should also increase in order to reduce the bias in the approximation of $F(\beta)$. We write $R(N)$ to emphasize that the number of grid points (and implicitly the grid of points itself) is now a function of the sample size. The main theorem in our paper is that, under restrictions on the economic model and an appropriate choice of $R(N)$, the estimator $\hat{F}_N(\beta)$ converges to the true $F_0(\beta)$ as $N \to \infty$, in a function space. The topology on our function space is induced by the Lévy-Prokhorov metric, a common metrization of the weak topology on the space of multivariate distributions.

We recognize that the nonparametric version of our estimator is a special case of a sieve estimator (Chen 2007). Sieve estimators estimate functions by increasing the flexibility of the approximating class used for estimation as the sample size increases. A sieve estimator for a smooth function might use an approximating class defined by a Fourier series, for example. As we are motivated by practical considerations in empirical work, our estimator's choice of basis, a discrete grid points, is justified by the estimator's computational simplicity. Further and unlike a typical sieve estimator, we need to constrain our estimated functions to be valid distribution functions. Our constrained linear regression approach is both computationally simple and ensures that the estimated CDF satisfies the theoretical properties of a valid CDF.[1]

Because our estimator is a sieve estimator, we prove its consistency by satisfying high-level conditions for the consistency of a sieve extremum estimator, as given in an appendix lemma in Chen and Pouzo (2009). We repeat this lemma and its proof in our paper so our consistency proof is self-contained. Our estimator is not a special case of the two-step sieve estimators explored using lower-level conditions in the main text of Chen and Pouzo. Several issues arise in proving consistency. Most interestingly, under the Lévy-Prokhorov metric on the space of multivariate distributions, the problem of optimizing the population objective function over the space of distributions turns out to be well posed under the definition of Chen (2007). Thus, our method does not rely on a sieve space to regularize the estimation problem to address the ill-posed inverse problem, as much of the sieve literature focuses on.

Our approach is a general, nonparametric mixtures estimator. The most common frequentist, nonparametric estimator is nonparametric maximum likelihood or NPMLE (Laird 1978, Böhning 1982, Lindsay 1983, Heckman and Singer 1984). Often the EM algorithm is used for computation (Dempster, Laird and Rubin 1977), but this approach is not guaranteed to find the global maximum. The literature worries about the strong dependence of the output of the EM algorithm on initial starting values and well as the difficulty in diagnosing convergence (Seidel, Mosler and Alker 2000, Verbeek, Vlassis and Kröse 2002, Biernacki, Celeux and Govaert 2003, Karlis and Xekalaki 2003).[2] Further, the EM algorithm has a slow rate of convergence even when it does converge to a global solution (Pilla and Lindsay 2001). Li and Barron (2000) introduce another alternative, but again our approach is computationally simpler. Our estimator is also computationally simpler than the minimum distance estimator of Beran and Millar (1994), which in our experience often has an objective function with an intractably large number of local minima. The discrete-grid idea (called the "histogram" approach) is found outside of economics in Kamakura (1991), who uses a discrete grid to estimate an ideal-point model. He does not discuss the nonparametric statistical properties of his approach. Of course, mixtures themselves have a long history in economics, such as Quandt and Ramsey (1978).

We prove the consistency of our estimator for the distribution of random parameters, in function space under the weak topology. To our knowledge, many of the alternative estimators discussed above do not have general, nonparametric consistency theorems for the estimator for the distribution of random parameters.[3] Our consistency theorem is not specific to the economic model being estimated.[4]

[1] FKRB also discuss the cases where β has continuous support and the researcher approximates the density with a mixture of normals.

[2] Another drawback of NPMLE that is specific to mixtures of normal distributions, a common approximating choice, is that the likelihood is unbounded and hence maximizing the likelihood does not produce a consistent estimator. There is a consistent root but it is not the global maximum of the likelihood function (McLachlan and Peel 2000).

[3] Beran (1995, Proposition 3) provides a proof for the consistency of the minimum distance estimator of the distribution of random coefficients in the linear regression model. However, the proof itself does not rely on properties of the linear regression model, other than its identification.

[4] In an earlier version of this paper, we also provide the rate of convergence of our estimator. Many of the competing

Despite the presence of nonparametric estimators in the literature, they are not commonly used by applied practitioners estimating distributions of random coefficients. Also, the theoretical results on consistency for these other estimators are not always of the generality needed for many economic models used in structural empirical work. By proving the nonparametric consistency of a computationally simple estimator for general economic choice models, we hope that nonparametric methods will be increasingly adopted by practitioners in industrial organization, marketing and other applied fields.

The outline of our paper is as follows. Section 2 reviews the general notation for the economic model and presents five examples of mixture models. Section 3 introduces the estimation procedure. Section 4 demonstrates consistency of our estimator in the space of multivariate distributions. Section 5 argues that our estimation problem is well-posed using the definition of Chen (2007). Section 6 extends our consistency results to models with both random coefficients and homogeneous parameters. Section 7 verifies the primitive conditions for consistency established in Section 4 using the five examples of mixture models in Section 2.

2 True Model and Examples

The econometrician observes a real valued vector of covariates x. The dependent variable in our model is denoted y, which indicates an underlying random variable y^* that takes values in the range of y^*, $\mathcal{Y}^*$. Note that y^* is not a latent variable. Some of our examples will focus primarily on the case where the range of y^* is a finite number of integer values, as is customary in discrete choice models. However, much of our analysis extends to the case where y^* is real valued.

Let A denote a (measurable) set in the range of y^*, $\mathcal{Y}^*$. We let $P_A(x)$ denote the probability that $y^* \in A$ when the decision problem has characteristics x. Let β denote a random coefficient that we assume is distributed independently of x. In our framework, this is a finite-dimensional, real-valued vector. We let $g_A(x, \beta)$ be the probability of A conditional on the random coefficients β and characteristics x. The CDF of the random coefficients is denoted by $F(\beta)$. The function g_A is specified as a modeling primitive. Given these definitions it follows that

$$P_A(x) = \int g_A(x, \beta)\, dF(\beta). \tag{2}$$

On the right hand side of the above equation, $g_A(x, \beta)$ gives the probability of A conditional on x and β. We average over the distribution of β using the CDF $F(\beta)$ to arrive at $P_A(x)$, the population probability of the event A conditional on x.

In our framework, the object the econometrician wishes to estimate is $F(\beta)$, the distribution of random coefficients. One definition of identification means that a unique $F(\beta)$ solves (2) for all x and all A. This is the definition used in certain relevant papers on identification in the statistics literature, for example Teicher (1963).

nonparametric estimators lack results on the rate of convergence. For example, Horowitz (1999, footnote 5) writes that “The rates of convergence of the Heckman-Singer estimators ... are unknown ...”

2.1 Examples of Mixture Models

We will return to these five examples later in the paper. Each example considers economic models with random coefficients that play a large role in empirical work. Some of the example models are nested in others, but verification of the conditions for consistency in Section 7 will use additional restrictions on the supports of x and β that are non-nested across models.

Example 1. (logit) Let there be a multinomial choice model such that y is one of $J+1$ unordered choices, such as types of cars for sale. The utility of choice j to consumer i is $u_{i,j} = x_{i,j}'\beta_i + \epsilon_{i,j}$, where $x_{i,j}$ is a vector of observable product characteristics of choice j and the demographics of consumer i, β_i is a vector of random coefficients giving the marginal utility of each car's characteristics to consumer i, and $\epsilon_{i,j}$ is an additive, consumer- and choice-specific error. There is an outside good 0 with utility $u_{i,0} = \varepsilon_{i,0}$. The consumer picks choice j when $u_{i,j} > u_{i,h} \, \forall \, h \neq j$. The random coefficients logit model occurs when $\epsilon_{i,j}$ is known to have the type I extreme value distribution. In this example, (1) becomes for $A = \{j\}$,

$$P_j(x) = \int g_j(x, \beta) \, dF(\beta) = \int \frac{\exp\left(x_j'\beta\right)}{1 + \sum_{h=1}^{J} \exp\left(x_h'\beta\right)} dF(\beta),$$

where $x = (x_1, \ldots, x_J)$. A similar expression occurs for other choices $h \neq j$. Compared to prior empirical work using the random coefficients logit, our goal is to estimate $F(\beta)$ nonparametrically.

Example 2. (binary choice) Let $J = 1$ in the previous example, so that there is one inside good and one outside good. Thus, the utility of the inside good 1 is $u_{i,1} = \epsilon_i + x_i'\beta_{1,i}$, where $\beta_i = (\epsilon_i, \beta_{1,i})$ is seen as one long vector and ϵ_i supplants the logit errors in Example 1 and plays the role of a random intercept. The outside good has utility $u_{i,0} = 0$. In this example, (1) becomes for $A = \{1\}$,

$$P_1(x) = \int g_1(x, \beta) \, dF(\beta) = \int 1\left[\epsilon + x'\beta_1 \geq 0\right] dF(\beta),$$

where $1[\cdot]$ is the indicator function equal to 1 if the inequality in the brackets is true. Without logit errors, the distribution of all unknowns in the model is estimated nonparametrically. In this example and others below, we allow the case where $g_A(x, \beta)$ in (1) is discontinuous in β.

Example 3. (multinomial choice without logit errors) Consider a multinomial choice model where the distribution of the previously logit errors is also estimated nonparametrically. In this case, the utility to choice j is $u_{i,j} = x_{i,j}'\tilde{\beta}_i + \epsilon_{i,j}$ and the utility of the outside good 0 is $u_{i,0} = 0$. The notation $\tilde{\beta}_i$ is used because the full random coefficient vector is now $\beta_i = \left(\tilde{\beta}_i, \epsilon_{i,1}, \ldots, \epsilon_{i,J}\right)$, which is seen as one long vector. We will not assume that the additive errors $\epsilon_{i,j}$ are distributed independently of β_i.

In this example, (1) becomes for $A = \{j\}$,

$$P_j\left(x\right) = \int g_j\left(x, \beta\right) dF\left(\beta\right) = \int 1\left[x_j'\tilde{\beta} + \epsilon_j \geq \max\left\{0, x_h'\tilde{\beta} + \epsilon_h\right\} \forall\, h \neq j\right] dF\left(\beta\right).$$

Example 4. (Cobb-Douglas production function / linear regression) Consider now a continuous outcome y_i from a Cobb-Douglas production function in logs, $y_i = \beta_i^0 + \beta_i^1 x_{i,l} + \beta_i^2 x_{i,k}$, where i is a manufacturing plant, y_i is the log of value added output, β_i^0 is the log of total factor productivity, β_i^1 is the input elasticity on labor (logged) $x_{i,l}$, and β_i^2 is the input elasticity on capital (logged) $x_{i,k}$. Let $x_i = \left(x_{i,l}, x_{i,k}\right)$ and $\beta_i = \left(\beta_i^0, \beta_i^1, \beta_i^2\right)$. Let $A = [a_1, a_2)$ be an interval in the range of y_i. In this example, (1) becomes

$$P_{[a_1,a_2)}\left(x\right) = \int g_{[a_1,a_2)}\left(x, \beta\right) dF\left(\beta\right) = \int 1\left[a_1 \leq \beta_i^0 + \beta_i^1 x_{i,l} + \beta_i^2 x_{i,k} < a_2\right] dF\left(\beta\right).$$

The goal is to estimate the joint distribution of total factor productivity and the input elasticities of labor and capital.

Example 5. (joint continuous and discrete demand) Consider now a model where a consumer picks a discrete choice j and, conditional on that choice j, the researcher observes a measure of usage, y_j. For example, j could be a brand of air conditioner, some of which use less energy than others. In this case, y_j might be the observed energy consumption from the chosen air conditioner. This is a variant of the generalized Roy selection model of Heckman (1990), as energy consumption y_j is observed only for the chosen air conditioner. Let the utility of choice j be $u_{i,j} = \tilde{x}_{i,j}'\tilde{\beta}_i$ and the utility of choice 0 be $u_{i,0} = 0$. Let the outcome for choice j be $y_{i,j} = \bar{x}_{i,j}'\bar{\beta}_i + \epsilon_{i,j}$, where $\bar{x}_{i,j}$ is a vector of characteristics that enters energy usage. Let $x_i = \left(\tilde{x}_{i,j}, \bar{x}_{i,j}\right)$, eliminating overlap, and let $\beta_i = \left(\tilde{\beta}_i, \bar{\beta}_i, \epsilon_{i,1}, \ldots, \epsilon_{i,J}\right)$. Let $A = \left([a_1, a_2), \{j\}\right)$. In this example, (1) becomes

$$P_{([a_1,a_2),\{j\})}\left(x\right) = \int g_{([a_1,a_2),\{j\})}\left(x, \beta\right) dF\left(\beta\right) = \int 1\left[a_1 \leq \bar{x}_j'\bar{\beta} + \epsilon_j < a_2\right] 1\left[\tilde{x}_j'\tilde{\beta} \geq \max\left\{0, \tilde{x}_h'\tilde{\beta}\right\} \forall\, h \neq j\right] dF\left(\beta\right).$$

In other words, we compute the probability of the discrete choice being j and the continuous outcome lying in the interval $[a_1, a_2)$. Compared to Heckman (1990), this model rules out an additive error in discrete choice utility (which leads to identification at infinity if included) but allows random coefficients in both the discrete choice and continuous outcomes portions of the model, which are not allowed in Heckman.

3 Estimator

The researcher picks $g_A(x,\beta)$ as the model of interest and seeks to estimate $F(\beta)$, the distribution of heterogeneity. We assume the researcher has access to N observations on (y_i, x_i). Given this, we divide the range $\mathcal{Y}$ of y into mutually exclusive sets $A_1, \ldots, A_J$. Let $y_{i,j} = 1$ if $y_i \in A_j$ and 0 otherwise.

Let $A = A_j$ and use j for A_j. Start with the model (2) and add $y_{i,j}$ to both sides while moving $P_j(x)$ to the right side. For the statistical observation i, this gives

$$y_{i,j} = \int g_j(x_i, \beta)\, dF(\beta) + (y_{i,j} - P_j(x_i)). \tag{3}$$

By the definition of $P_j(x)$, the expectation of the composite error term $y_{i,j} - P_j(x)$, conditional on x, is 0. This is a linear probability model with an infinite-dimensional parameter, the distribution $F(\beta)$. We could work directly with this equation if it was computationally simple to estimate this infinite-dimensional parameter while constraining it to be a valid CDF.

Instead, we work with a finite-dimensional sieve space approximation to F. In particular, we let $R(N)$ be the number of grid points in the grid $\mathcal{B}_{R(N)} = \left(\beta^1, \ldots, \beta^{R(N)}\right)$. A grid point is a vector if β is a vector, so $R(N)$ is the total number of points in all dimensions. The researcher chooses $\mathcal{B}_{R(N)}$. Given the choice of $\mathcal{B}_{R(N)}$, the researcher estimates $\theta = \left(\theta^1, \ldots, \theta^{R(N)}\right)$, the weights on each of the grid points. With this approximation, (3) becomes

$$y_{i,j} \approx \sum_{r=1}^{R(N)} \theta^r g_j(x_i, \beta^r) + (y_{i,j} - P_j(x)).$$

We use the $\approx$ symbol to emphasize that this uses a sieve approximation to the distribution function $F(\beta)$. Because each θ^r enters $y_{i,j}$ linearly, we estimate $\left(\theta^1, \ldots, \theta^{R(N)}\right)$ using the linear probability model regression of $y_{i,j}$ on the R "regressors" $z_{i,j}^r = g_j(x_i, \beta^r)$.

To be a valid CDF, $\theta^r \geq 0 \, \forall \, r$ and $\sum_{r=1}^{R(N)} \theta^r = 1$. Therefore, the estimator is

$$\begin{aligned} \widehat{\theta} &= \arg\min_{\theta} \frac{1}{NJ} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R(N)} \theta^r z_{i,j}^r \right)^2 \\ & \text{subject to } \theta^r \geq 0 \, \forall \, r = 1, \ldots, R(N) \text{ and } \sum_{r=1}^{R(N)} \theta^r = 1. \end{aligned} \tag{4}$$

There are J "regression observations" for each statistical observation (y_i, x_i). This minimization problem is a quadratic programming problem subject to linear inequality constraints. The minimization problem is convex and routines like MATLAB's lsqlin guarantee finding a global optimum. One can construct the estimated cumulative distribution function for the random coefficients as

$$\hat{F}_N(\beta) = \sum\nolimits_{r=1}^{R(N)} \hat{\theta}^r 1[\beta^r \leq \beta],$$

where $1[\beta^r \leq \beta]$ is equal to 1 when $\beta^r \leq \beta$. Thus, we have a structural estimator for a distribution of random parameters in addition to a flexible method for approximating choice probabilities.

The approach just presented has two main advantages over other approaches to estimating distributions of random coefficients. First, the approach is computationally simple: we can always find a global optimum and, by solving for $z_{i,j}^r = g_j(x_i, \beta^r)$ before optimization commences, we avoid many evaluations of complex structural models such as dynamic programming problems. Second, the approach is nonparametric. In the next section, we show that if the grid of points is made finer as the sample size N increases, the estimator $\hat{F}_N(\beta)$ converges to the true distribution F_0. We do not need to impose that F_0 lies in known parametric family.

On the other hand, a disadvantage is that the estimates may be sensitive to the choice of tuning parameters. While most nonparametric approaches require choices of tuning parameters, here the choice of a grid of points is a particularly high-dimensional tuning parameter. FKRB propose cross-validation methods to pick these tuning parameters, including the number of grid points, the support of the points, and the grid points within the support.

Example. 1 (logit) For the logit example, $\frac{\exp\left(x_{i,j}'\beta^r\right)}{1+\sum_{h=1}^{J}\exp\left(x_{i,h}'\beta^r\right)} = g_j(x_i, \beta^r)$. Therefore, for each statistical observation i, the researcher computes $R \cdot J$ linear probability model regressors $z_{i,j}^r = \frac{\exp\left(x_{i,j}'\beta^r\right)}{1+\sum_{h=1}^{J}\exp\left(x_{i,h}'\beta^r\right)}$. This computation is done before optimization commences. The outcome for choosing the outside good 0 does not need to included in the objective function, as $\sum_{j=0}^{J} g_j(x_i, \beta^r) = 1$.

4 Consistency in Function Space

Assume that the true distribution function F_0 lies in the space $\mathcal{F}$ of distribution functions on the support $\mathcal{B}$ of the parameters β. We wish to show that the estimated distribution function $\hat{F}_N(\beta) = \sum_{r=1}^{R(N)} \hat{\theta}^r 1[\beta^r \leq \beta]$ converges to the true $F_0 \in \mathcal{F}$ as the sample size N grows large. Most of the competing nonparametric estimators discussed in the introduction are not only computationally more challenging, but lack nonparametric consistency theorems for as general a class of economic models.

To prove consistency, we use the recent results for sieve estimators developed by Chen and Pouzo (2009), hereafter referred to as CP. We define a sieve space to approximate $\mathcal{F}$ as

$$\mathcal{F}_R = \left\{ F \mid F(\beta) = \sum\nolimits_{r=1}^{R} \theta^r 1[\beta^r \leq \beta], \theta \in \Delta_R \equiv \left\{ \left(\theta^1, \ldots, \theta^R\right)' \mid \theta^r \geq 0, \sum\nolimits_{r=1}^{R} \theta^r = 1 \right\} \right\},$$

for a choice of grid $\mathcal{B}_R = \left\{\beta^1, \ldots, \beta^R\right\}$ that becomes finer as R increases. We require $\mathcal{F}_R \subseteq \mathcal{F}_S \subset \mathcal{F}$ for $S > R$, or that large sieve spaces encompass smaller sieve spaces. The choice of the grid and $R(N)$ are up to the researcher; however consistency will require conditions on these choices.

Based on CP, we prove that the estimator $\hat{F}_N$ converges to the true F_0. In their main text, CP study sieve minimum distance estimators that involve a two-stage procedure. Our estimator is a one-stage sieve least squares estimator (Chen, 2007) and so we cannot proceed by verifying the conditions in the theorems in the main text of CP. Instead, we show its consistency based on CP's general consistency theorem in their appendix, their Lemma B.1, which we quote in the proof of our consistency theorem for completeness. As a consequence, our consistency proof verifies CP's high-level conditions for the consistency of a sieve extremum estimator.

Let y_i denote the $J \times 1$ finite vector of binary outcomes $(y_{i,1}, \ldots, y_{i,J})$ and let $g(x_i, \beta)$ denote the corresponding $J \times 1$ vector of choice probabilities $(g_1(x_i, \beta), \ldots, g_J(x_i, \beta))$ given x_i and the random coefficient β. Then we can define our sample criterion function as

$$\hat{Q}_N(F) \equiv \frac{1}{NJ} \sum\nolimits_{i=1}^{N} \left\| y_i - \int g(x_i, \beta) dF(\beta) \right\|_E^2 = \frac{1}{NJ} \sum\nolimits_{i=1}^{N} \left\| y_i - \sum\nolimits_{r=1}^{R} \theta^r g(x_i, \beta^r) \right\|_E^2 \tag{5}$$

for $F \in \mathcal{F}_{R(N)}$, where $|| \cdot ||_E$ denotes the Euclidean norm. We can rewrite our estimator as

$$\hat{F}_N = \operatorname{argmin}_{F \in \mathcal{F}_{R(N)}} \hat{Q}_N(F) + C \cdot \nu_N \tag{6}$$

where we can allow for some tolerance (slackness) of minimization, $C \cdot \nu_N$, that is a positive sequence tending to zero as N gets larger.

Also let

$$Q(F) \equiv E\left[\left\| y - \int g(x, \beta) \, dF(\beta) \right\|_E^2 / J \right]$$

be the population objective function.

We state assumptions on the model first. We write $P(x, F) = \int g(x, \beta) dF(\beta)$. As a distance measure for distributions, we use the Lévy-Prokhorov metric, denoted by $d_{\mathrm{LP}}(\cdot)$, which is a metrization of the weak topology for the space of multivariate distributions $\mathcal{F}$. The Lévy-Prokhorov metric in the space of $\mathcal{F}$ is defined on a metric space $(\mathcal{B}, d)$. We use notation $d_{\mathrm{LP}}(F_1, F_2)$ where the measures are implicit. This denotes the Lévy-Prokhorov metric $d_{\mathrm{LP}}(\mu_1, \mu_2)$, where μ_1 and μ_2 are probability measures corresponding to F_1 and F_2. The Lévy-Prokhorov metric is defined as

$$d_{\mathrm{LP}}(\mu_1, \mu_2) = \inf \left\{ \epsilon > 0 \mid \mu_1(C) \leq \mu_2(C^{\epsilon}) + \epsilon \, \mathrm{and} \, \mu_2(C) \leq \mu_1(C^{\epsilon}) + \epsilon \, \mathrm{for\,all\,Borel\,measurable} \, C \in \mathcal{B} \right\},$$

where C is some set of random coefficients and $C^{\epsilon} = \{ b \in \mathcal{B} \mid \exists \, a \in C, \, d(a, b) < \epsilon \}$. The Lévy-Prokhorov metric is a metric, so that $d_{\mathrm{LP}}(\mu_1, \mu_2) = 0$ only when $\mu_1 = \mu_2$. See Huber (1981, 2004).

The following assumptions are on the economic model and data generating process.

Assumption 1.

1. *Let $\mathcal{F}$ be a space of distribution functions on a finite-dimensional real space $\mathcal{B}$, where $\mathcal{B}$ is compact. $\mathcal{F}$ contains F_0.*

2. *Let $\{(y_i, x_i)\}_{i=1}^{N}$ be i.i.d.*

3. *Let β be independently distributed from x.*

4. *Assume the model $g(x,\beta)$ is identified, meaning that for any $F_1 \neq F_0$, $F_1 \in \mathcal{F}$, we have $P(x,F_0) \neq P(x,F_1)$ for almost all $x \in \tilde{\mathcal{X}}$, where $\tilde{\mathcal{X}}$ is a subset of $\mathcal{X}$, the support of x, with positive probability.*[5]

5. *$Q(F)$ is continuous on $\mathcal{F}$ in the weak topology.*

Assumptions 1.1, 1.2, and 1.3 are standard for nonparametric mixtures models with cross-sectional data. Assumption 1.4 requires that the model be identified at a set of values of x_i that occurs with positive probability. The assumption rules out so-called fragile identification that could occur at values of x with measure zero (e.g. identification at infinity). Assumptions 1.4 and 1.5 need to be verified for each economic model $g_A(x,\beta)$. We will do so for our five examples below.[6]

Remark 1. Assumption 1.5 is satisfied when $g(x,\beta)$ is continuous in β for all x because in this case $P(x,F)$ is also continuous on $\mathcal{F}$ for all x in the Lévy-Prokhorov metric. Then by the dominated convergence theorem, the continuity of $Q(F)$ in the weak topology follows from the continuity of $P(x,F)$ on $\mathcal{F}$ for all x and $P(x,F) \leq 1$ (uniformly bounded). Here the continuity of $P(x,F)$ on $\mathcal{F}$ means for any $F_1, F_2 \in \mathcal{F}$ such that $d_{LP}(F_1,F_2) \to 0$ it must follow that $\left|\int g_j(x,\beta)dF_1(\beta) - \int g_j(x,\beta)dF_2(\beta)\right| \to 0$ for all j. This holds by the definition of weak convergence when $g(x,\beta)$ is continuous and bounded and because the Lévy-Prokhorov metric is a metrization of the weak topology.

Remark 2. If the support $\mathcal{B}$ is a finite set (i.e. types are discrete with known support), the continuity in the weak topology holds even when $g_j(x,\beta)$ is discontinuous as long as it is bounded because in this case the Lévy-Prokhorov metric becomes equivalent to the total variation metric (see Huber 1981, p.34). This implies $\left|\int g_j(x,\beta)dF_1(\beta) - \int g_j(x,\beta)dF_2(\beta)\right| \to 0$ as long as $g_j(x,\beta)$ is bounded.

In addition to Assumption 1, we also require that the grid of points be chosen so that the grid $\mathcal{B}_R$ becomes dense in $\mathcal{B}$ in the usual topology on the reals.

Condition 1. Let the choice of grids satisfy the following properties:

1. Let $\mathcal{B}_R$ become dense in $\mathcal{B}$ as $R \to \infty$.

2. $\mathcal{F}_R \subseteq \mathcal{F}_{R+1} \subseteq \mathcal{F}$ for all $R \geq 1$.

3. $R(N) \to \infty$ as $N \to \infty$ and it satisfies $\frac{R(N)\log R(N)}{N} \to 0$ as $N \to \infty$.

The first two parts of this condition have previously been mentioned, and ensure that the sieve spaces give increasingly better approximations to the space of multivariate distributions. Condition 1.3 specifies a rate condition so that the convergence of the sample criterion function $\hat{Q}_N(F)$ to the population criterion function $Q(F)$ is uniform over $\mathcal{F}_R$. Uniform convergence of the criterion function and identification are both key conditions for consistency.

[5] This is with respect to the probability measure of the underlying probability space. This probability is well defined whether x is continuous, discrete or some elements of x_i are functions of other elements (e.g. polynomials or interactions).

[6] The continuity condition in Assumption 1.5 can be relaxed to lower semicontinuity. The examples we consider in this paper satisfy the continuity condition.

Theorem 1. *Suppose Assumption 1 and Condition 1 hold. Then,* $d_{\text{LP}}\left(\hat{F}_N, F_0\right) \xrightarrow{p} 0$.

See Appendix A for the proof.

Remark 3. The literature on sieve estimation has not established general results on the asymptotic distribution of sieve estimators, in function space. However, for rich classes of approximating basis functions that do not include our approximation problem, the literature has shown conditions under which finite dimensional functionals of sieve estimators have asymptotically normal distributions. In the case of nonparametric random coefficients, we might be interested in inference in the mean or median of β. For the demand estimation, say Example 3, we might be interested in average responses (or elasticities) of choice probabilities with respect to changes in particular product characteristics. Let $\Pi_N F_0$ be a sieve approximation to F_0 in our sieve space $\mathcal{F}_{R(N)}$. If we could obtain an error bound for $d_{\text{LP}}\left(\Pi_N F_0, F_0\right)$, we could also derive the convergence rate in the Lévy-Prokhorov metric (Chen 2007). If the error bound shrinks fast enough as $R(N)$ increases, we conjecture that we could also prove that plug-in estimators for functionals of F_0 are asymptotically normal (Chen, Linton, and van Keilegom 2003).[7] Error bounds for discrete approximations are available in the literature for a class of parametric distributions F, but we are not aware of results for the unrestricted class of multivariate distributions. In an earlier version of the paper, we derived the convergence rate in the L_1 metric instead.

5 Well-Posedness

Chen (2007) and Carrasco, Florens and Renault (2007) distinguish between functional (here the distribution) optimization and identification problems that are well-posed and problems that are ill-posed. Using Chen's definition, the optimization problem of maximizing the population criterion function $Q(F)$ with respect to the distribution function F will be well-posed if $d_{\text{LP}}(F_n, F_0) \to 0$ for all sequences $\{F_n\}$ in $\mathcal{F}$ such that $Q(F_n) - Q(F_0) \to 0$. The problem will be ill-posed if there exists a sequence $\{F_n\}$ in $\mathcal{F}$ such that $Q(F_n) - Q(F_0) \to 0$ but $d_{\text{LP}}(F_n, F_0) \nrightarrow 0$.[8] We now argue that our problem is well-posed.

The space $\mathcal{F}$ of distributions on $\mathcal{B}$ is compact in the weak topology if $\mathcal{B}$ itself is compact (Assumption 1.1) in Euclidean space (Parthasarathy 1967, Theorem 6.4). Also, $Q(F)$ is continuous on $\mathcal{F}$ by Assumption 1.5. It follows that with our choice of the criterion function and metric, our optimization problem is well posed in the sense of Chen (2007) because for every $\epsilon > 0$ we have

$$\inf_{F \in \mathcal{F}_{R(N)}: d_{\text{LP}}(F, F_0) \geq \epsilon} \left(Q(F) - Q(F_0)\right) \geq \inf_{F \in \mathcal{F}: d_{\text{LP}}(F, F_0) \geq \epsilon} \left(Q(F) - Q(F_0)\right) > 0, \tag{7}$$

[7] We conjecture that we could prove an analog to Theorem 2 in Chen et al (2003) if we could verify analogs to conditions (2.4)–(2.6) in that paper for our sieve space.

[8] Whether the problem is well-posed or ill-posed also depends on the choice of the metric. For example, if one uses the total variation distance metric instead of the the Lévy-Prokhorov metric, the problem will be ill-posed because the distance between a continuous distribution and any discrete distribution will always be equal to one in the total variation metric.

where the first inequality holds because $\mathcal{F}_{R(N)} \subset \mathcal{F}$ by construction and the second, strict inequality holds as the minimum is attained by continuity and compactness and because the model is identified (Assumption 1.4), as we argue in the proof of Theorem 1. Therefore, our optimization problem satisfies Chen's definition of well-posedness.

6 Consistency for Models with Homogenous Parameters

In many empirical applications, it is common to have both random coefficients β and finite-dimensional parameters $\alpha \in \mathcal{A} \subseteq \mathbb{R}^{\dim(\alpha)}$. We write the model choice probabilities as $g(x, \beta, \alpha)$ and the aggregate choice probabilities as $P(x, F, \alpha)$. Here we consider the consistency of estimators for models with both homogenous parameters and random coefficients.

Remark 4. Estimating a model allowing a parameter to be a random coefficient when in truth the parameter is homogeneous will not affect consistency if the model with random coefficients is identified.

Remark 5. Searching over α as a homogeneous parameter requires nonlinear least squares. The optimization problem may also not be globally convex. The objective function may not be differentiable for our examples where $g(x, \beta, \alpha)$ involves an indicator function.

Our estimator for models with homogeneous parameters is defined as (similarly to (6))

$$(\hat{\alpha}_N, \hat{F}_N) = \text{argmin}_{(\alpha,F)\in\mathcal{A}\times\mathcal{F}_{R(N)}} \hat{Q}_N(\alpha, F) + C \cdot \nu_N,$$

where $\hat{Q}_N(\alpha, F)$ denotes the corresponding sample criterion function. $Q(\alpha, F)$ is the population criterion function based on the model $g(x, \beta, \alpha)$. An alternative computational strategy is that the estimator can be profiled as

$$\hat{F}_N(\alpha) = \text{argmin}_{F\in\mathcal{F}_{R(N)}} \hat{Q}_N(\alpha, F) + C \cdot \nu_N \text{ for all } \alpha \in \mathcal{A}.$$

Profiling gives us

$$\hat{\alpha}_N = \text{argmin}_{a\in\mathcal{A}} \hat{Q}_N\left(\alpha, \hat{F}_N(\alpha)\right) + C \cdot \nu_N, \tag{8}$$

and therefore $\hat{F}_N = \hat{F}_N(\hat{\alpha}_N)$.

Using Theorem 1 of Chen, Linton, and van Keilegom (2003), below we show $\hat{\alpha}_N$ is consistent (so $\hat{F}_N$ is as well when combined with Theorem 1). We replace Assumptions 1.4 and 1.5 with Assumptions 2.2 and 2.3 below and add one restriction on $g(x, \beta, \alpha)$.

Assumption 2.

1. *Let $\mathcal{A}$ be the parameter space of α, which is a compact subset of $\mathbb{R}^{\dim(\alpha)}$.*

2. *Assume the model $g(x, \beta, \alpha)$ is identified, meaning that for any $(\alpha_1, F_1) \neq (\alpha_0, F_0)$, $(\alpha_1, F_1) \in \mathcal{A} \times \mathcal{F}$, we have $P(x, F_0, \alpha_0) \neq P(x, F_1, \alpha_1)$ for almost all $x \in \tilde{\mathcal{X}}$, where $\tilde{\mathcal{X}}$ is a subset of $\mathcal{X}$, the support of x, with positive probability.*

3. *$Q(\alpha, F)$ is continuous on $\mathcal{A}$ and is continuous on $\mathcal{F}$ in the weak topology.*

4. *Either (i) $g(x, \beta, \alpha)$ is Lipschitz continuous in α, (ii) α enters $g(x, \beta, \alpha)$ only through indicator functions or (iii) $g(x, \beta, \alpha^1, \alpha^2)$ with $\alpha = (\alpha^1, \alpha^2)$ is Lipschitz continuous in α^1 and α^2 enters $g(x, \beta, \alpha^1, \alpha^2)$ only through indicator functions.*

If homogeneous parameters were added, Assumption 2.4.i would hold for Example 1, the logit model with random coefficients. Assumption 2.4.ii would hold for Examples 2–5. Finally, Assumption 2.4.iii would hold for a joint discrete and continuous demand model with logit errors in the discrete choice utility and for a discrete choice demand model with logit errors and price endogeneity, both of which are described in FKRB, Section 5.

We present the consistency theorem for the estimator of the homogenous parameters.

Corollary 1. *Suppose Assumptions 1.1 through 1.3, Assumption 2 and Condition 1 hold. Then, $\hat{\alpha}_N \xrightarrow{p} \alpha_0$.*

7 Identification and Continuity for Examples

We return to the examples we introduced in Section 2.1. We verify the two key conditions for each model $g_A(x, \beta)$: Assumption 1.4, identification of $F(\beta)$, and Assumption 1.5, continuity of the population objective function under the Lévy-Prokhorov metric. Throughout this section, we assume Assumptions 1.1–1.3 hold. Note that Matzkin (2007) is an excellent survey of older results on the identification of models with heterogeneity.

Example. 1 (logit) The identification of $F(\beta)$ in the random coefficients logit model is the main content of Bajari, Fox, Kim and Ryan (2010, Theorem 13).[9] Assumption 12 in Bajari et al states that "The support of x, $\mathcal{X}$ contains $x = 0$, but not necessarily an open set surrounding it. Further, the support contains a nonempty open set of points (open in $\mathbb{R}^{\dim(x_j)}$) of the form $\left(x_1', \ldots, x_{j-1}', x_j', x_{j+1}', \ldots, x_J'\right) = \left(0', \ldots, 0', x_j', 0', \ldots, 0'\right)$."[10] Bajari et al also require the support of $\mathcal{X}$ to be a product space, which rules out including polynomial terms in an element of x_j or including interactions of two elements of x_j. Given this assumption, Assumption 1.4 holds. Assumption 1.5 holds by Remark 1 in the current paper.

[9] Theorem 13 of Bajari et al also allows homogeneous, product-specific intercepts.

[10] Bajari et al discusses what $x = 0$ means when the means of product characteristics can be shifted.

Example. 2 (binary choice) Ichimura and Thompson (1998, Theorem 1) establish the identification of $F(\beta)$ under the conditions that i) the coefficient on one of the the non-intercept regressors in x is known to always be positive or negative and ii) there are large and product supports on each of the regressors other than the intercept. This rules out polynomial terms and interactions. For i), we formally make this "special regressor" assumption as follows. Let x_{i,k^*} be an element in x_i and $x_{i,-k^*}$ be the subvector of x_i that removes x_{i,k^*}.

Assumption 3. ***(binary choice)*** *(i)* The conditional CDF of x_{i,k^*} given $x_{i,-k^*}$, labeled as $G_{x_{i,k^*}|x_{i,-k^*}}$, is continuous in x_{i,k^*} for almost every value of $x_{i,-k^*}$. (ii) Let the support of the coefficient β_{k^*} be known to be strictly positive or negative.

If we impose the scale normalization $\beta_{k^*} = \pm 1$, Assumption 1.4 holds if we add large support conditions on each regressor in x. It turns out that Assumption 3, without needing large support, also ensures the continuity of $Q(F)$ on $\mathcal{F}$, in the weak topology. The proof is in the appendix.

Lemma 1. ***(binary choice)*** *Suppose Assumptions 1.1-1.3 and 3 hold. Then $Q(F)$ is continuous on $\mathcal{F}$ in the weak topology and Assumption 1.5 holds.*

An advantage of our estimator is the ease of imposing sign restrictions if necessary. Because the researcher picks $\mathcal{B}_{R(N)} = \left(\beta^1, \ldots, \beta^{R(N)}\right)$, the researcher can choose the grid so that the first element of each vector β^r is always positive, for example. Note that binary choice is a special case of multinomial choice, so the non-nested identification conditions in example 3, below, can replace these used here.

Example. 3 (multinomial choice without logit errors) Fox and Gandhi (2010, Theorem 7.6) study the identification of the multinomial choice model without logit errors. Our linear specification of the utility function for each choice is a special case of what they allow. Fox and Gandhi require a choice-j-specific special regressor along the lines of Assumption 3. On other hand, Fox and Gandhi allow polynomial terms and interactions for x's other than the choice-j-specific special regressors, unlike examples 1 and 2. They do not require large support for the x's that are not special regressors. The most important additional assumption for identification is that Fox and Gandhi require that $F(\beta)$ takes on at most a finite number T of support points, although the number T and support point identities $\beta^1, \ldots, \beta^T$ are learned in identification. The number T in the true F^0 is not related in any way to the finite-sample $R(N)$ used for estimation in this paper. So Assumption 1.4 holds under this restriction on $\mathcal{F}$. For the continuity Assumption 1.5, the equivalent of Assumption 3 also implies the equivalent of Lemma 1. We omit the formal statement and proof for conciseness.

Example. 4 (Cobb-Douglas production function / linear regression) Beran and Millar (1994, Proposition 2.2) establish identification of $F(\beta)$, in a model with one regressor and one slope coefficient. However, they use individual realizations of the continuous outcome y and not intervals such as $[a_1, a_2)$.

To implement our estimator, we discretize the outcome space $\mathcal{Y}$ into J intervals $[a_{j-1}, a_j)$, $j = 1, \ldots, J$. The following lemma uses techniques related to Fox and Gandhi (2010) to prove that the distribution of random coefficients is identified in an interval-censored linear regression model with any number of regressors.

Lemma 2. ***(linear regression with interval censoring)*** *Let the underlying regression model be $y = x'\beta$ where the first covariate, x_1, can take on any value on $\mathbb{R}^+$, conditional on the other components of x, x_{-k}. Assume the researcher uses only data on the dependent variable $y_{i,j} = 1\left[y_i \in [a_{j-1}, a_j)\right]$ for J intervals. Then $F(\beta)$ is identified if the true distribution takes on at most an unknown, finite number of support points. Therefore Assumption 1.4 holds.*

The proof is in an appendix. As with example 4, polynomial terms and interactions are not ruled out, except for those on x_1. Note that there is no common sign restriction on the coefficient on the regressor with large support. For the continuity Assumption 1.5, the equivalent of Assumption 3 also implies the equivalent of Lemma 1. This assumption is almost without loss of generality in the production function example, as labor and capital have continuous supports and positive input elasticities.

Example. 5 (joint continuous and discrete demand) Fox and Gandhi (2011) explore the identification of this selection model with random coefficients. Like in Example 3 above, a choice-j-specific special regressor must enter the utility for each discrete choice. Also like in Example 3, $F(\beta)$ takes on at most a finite number T of support points, although to emphasize, T is learned in identification, as are the support points. So Assumption 1.4 might hold under this restriction on $\mathcal{F}$. The main new issue is the same as Example 4: the discretization of the continuous dependent variable in our linear probability model estimation procedure. The Fox and Gandhi result uses the non-discretized continuous outcomes, and like in Example 4, to implement our estimator, we discretize the continuous portion of the outcome space into intervals $\left[a^j_{\iota-1}, a^j_\iota\right)$ for each choice j. This requires a large-support regressor in the outcome equations as well. Combining the arguments in Fox and Gandhi (2011) with similar arguments to those in the proof of Lemma 2 shows identification of the selection model in the case of interval censoring. For the continuity Assumption 1.5, the equivalent of Assumption 3 also implies the equivalent of Lemma 1.

8 Conclusion

We analyze a nonparametric method for estimating general mixtures models. Our method allows the researcher to drop standard parametric assumptions, such as independent normal random coefficients, that are commonly used in applied work. Convergence of an optimization routine to the global optimum is guaranteed under linear regression with linear constraints, something that cannot be said for other statistical objective functions. Also, our estimator is easier to program and to use than alternatives such as the EM algorithm. The estimator is also useful for reducing the number of times complex structural models, such as dynamic programs, need to be evaluated in estimation.

We explore the asymptotic properties of the nonparametric distribution estimator. We show consistency in the function space of all distributions under the weak topology by viewing our estimator as a sieve estimator and verifying high-level conditions in Chen and Pouzo (2009). Many alternative mixtures estimators lack consistency results in such generality. We verify the conditions for consistency for five example models, each of which is widely used in empirical work.

A Proof of Consistency: Theorem 1

We verify the conditions of CP's Lemma B.1 (also see Theorem 3.1 in Chen (2007)) in our consistency proof. To provide completeness, we first present our simplified version of CP's Lemma B.1, which does not incorporate a penalty function. Define $\hat{Q}_N(F) = \frac{1}{NJ}\sum_{i=1}^N \left\|y_i - \int g(x_i,\beta)\,dF\right\|_E^2$ and $Q(F) \equiv E\left[\left\|y - \int g(x,\beta)\,dF\right\|_E^2 / J\right]$.

Lemma 3. ***Lemma B.1 of CP:*** *Let $\hat{F}_N$ be such that $\hat{Q}_N(\hat{F}_N) \le \inf_{F\in\mathcal{F}_{R(N)}} \hat{Q}_N(F) + O_p(\nu_N)$ with $\nu_N \to 0$.* Suppose the following conditions (A.3.1)-(A.3.4) hold:

- (A.3.1) (i) $Q(F_0) < \infty$; (ii) there is a positive function $\delta(N, R(N), \varepsilon)$ such that for each $N \ge 1$, $R \ge 1$, and $\varepsilon > 0$, $\inf_{F\in\mathcal{F}_{R(N)}:d_{\mathrm{LP}}(F,F_0)\ge\varepsilon} Q(F) - Q(F_0) \ge \delta(N, R(N), \varepsilon)$ and $\liminf_{N\to\infty} \delta(N, R(N), \varepsilon) \ge 0$ for all $\varepsilon > 0$.

- (A.3.2) (i) $(\mathcal{F}, d_{\mathrm{LP}}(\cdot))$ is a metric space; (ii) $\mathcal{F}_R \subseteq \mathcal{F}_{R+1} \subseteq \mathcal{F}$ for all $R \ge 1$, and there exists a sequence $\Pi_N F_0 \in \mathcal{F}_{R(N)}$ such that $d_{\mathrm{LP}}(\Pi_N F_0, F_0) = O(\varsigma_N)$ and $\varsigma_N \to 0$ as $N \to \infty$.

- (A.3.3) (i) $\hat{Q}_N(F)$ is a measurable function of the data $\{(y_i, x_i)\}_{i=1}^N$ for all $F \in \mathcal{F}_{R(N)}$; (ii) $\hat{F}_N$ is well-defined and measurable with respect to the Borel σ-field generated by the weak topology.

- (A.3.4) (i) Let $\hat{c}^Q(R(N)) = \sup_{F\in\mathcal{F}_{R(N)}} \left|\hat{Q}_N(F) - Q(F)\right| \xrightarrow{p} 0$;
  (ii) $\max\left\{\hat{c}^Q(R(N)), \nu_N, |Q(\Pi_N F_0) - Q(F_0)|\right\} / \delta(N, R(N), \varepsilon) \xrightarrow{p} 0$ for all $\varepsilon > 0$.

Then $d_{\mathrm{LP}}(\hat{F}_N, F_0) \xrightarrow{p} 0$.

Proof. Under condition (A.3.3) (ii), $\hat{F}_N$ is well-defined and measurable. Then for any $\varepsilon > 0$,

$$\begin{aligned}
&\Pr(d_{\mathrm{LP}}(\hat{F}_N, F_0) \ge \varepsilon) \le \Pr(\inf_{F\in\mathcal{F}_{R(N)}:d_{\mathrm{LP}}(F,F_0)\ge\varepsilon} \hat{Q}_N(F) \le \hat{Q}_N(\Pi_N F_0) + O(\nu_N)) \\
\le\;& \Pr(\inf_{F\in\mathcal{F}_{R(N)}:d_{\mathrm{LP}}(F,F_0)\ge\varepsilon} \{Q(F) + (\hat{Q}_N(F) - Q(F))\} \le Q(\Pi_N F_0) + (\hat{Q}_N(\Pi_N F_0) - Q(\Pi_N F_0)) + O(\nu_N)) \\
\le\;& \Pr(\inf_{F\in\mathcal{F}_{R(N)}:d_{\mathrm{LP}}(F,F_0)\ge\varepsilon} Q(F) \le 2\hat{c}^Q(R(N)) + Q(\Pi_N F_0) + O(\nu_N)) \\
\le\;& \Pr(\inf_{F\in\mathcal{F}_{R(N)}:d_{\mathrm{LP}}(F,F_0)\ge\varepsilon} Q(F) - Q(F_0) \le 2\hat{c}^Q(R(N)) + Q(\Pi_N F_0) - Q(F_0) + O(\nu_N)) \\
\le\;& \Pr(\delta(N, R(N), \varepsilon) \le 2\hat{c}^Q(R(N)) + |Q(\Pi_N F_0) - Q(F_0)| + O(\nu_N))
\end{aligned}$$

which goes to zero by condition (A.3.4). □

Now we provide our consistency proof for the baseline estimator. Because our estimator is an extremum estimator, we can take ν_N to be arbitrary small. We start with the condition (A.3.1). The condition $Q(F_0) < \infty$ holds because $Q(F) \le 1$ for all $F \in \mathcal{F}$. Next we will verify the condition

$$\inf_{F\in\mathcal{F}_{R(N)}:d_{\mathrm{LP}}(F,F_0)\ge\varepsilon} Q(F) - Q(F_0) \quad\ge\quad \delta(N, R(N), \varepsilon) > 0 \tag{9}$$

for each $N \geq 1$, $R(N) \geq 1$, $\varepsilon > 0$, and some positive function $\delta(N, R(N), \varepsilon)$ to be defined below. We will use our assumption of identification (Assumption 1.4). Let $m(x, F) = P(x, F_0) - P(x, F)$. Note that we have

$$Q(F) \;=\; E\left[||y - P(x, F_0) + m(x, F)||_E^2/J\right] = E\left[||y - P(x)||_E^2/J\right] + E\left[||m(x, F)||_E^2/J\right]$$

because $E[(y - P(x))'m(x, F)] = 0$ by the law of iterated expectation and $E[y - P(x)|x] = 0$. Therefore, for each $F \in \mathcal{F}$, we have

$$Q(F) - Q(F_0) \;=\; E\left[||m(x, F)||_E^2/J\right] - E\left[||m(x, F_0)||_E^2/J\right] = E\left[||m(x, F)||_E^2/J\right] \qquad (10)$$

because $m(x, F_0) = 0$ and the condition (9) holds due to our assumption of identification as the following argument shows.

Consider $E[||m(x, F)||_E^2]$, with $m(x, F)$ defined above, as a map from $\mathcal{F}$ to $\mathbb{R}^+ \cup \{0\}$. For any $F \neq F_0$, $E[||m(x, F)||_E^2]$ takes on positive values for each $F \in \mathcal{F}$, because the model is identified on a set $\tilde{\mathcal{X}}$ with positive probability. Then note that $E[||m(x, F)||_E^2]$ is continuous in F and also note that $\mathcal{F}_{R(N)}$ is compact. Therefore $E[||m(x, F)||_E^2]$ attains some strictly positive minimum on $\{F \in \mathcal{F}_{R(N)} : d_{\mathrm{LP}}(F, F_0) \geq \varepsilon\}$. Then we can take $\delta(N, R(N), \varepsilon) = \inf_{F \in \mathcal{F}_{R(N)}: d_{\mathrm{LP}}(F, F_0) \geq \varepsilon} E\left[||m(x, F)||_E^2/J\right] > 0$ for all $R(N) \geq 1$ with $\varepsilon > 0$.

We further claim $\liminf_{N \to \infty} \delta(N, R(N), \varepsilon) > 0$ because

$$\begin{aligned} \delta(N, R(N), \varepsilon) &= \inf_{F \in \mathcal{F}_{R(N)}: d_{\mathrm{LP}}(F, F_0) \geq \varepsilon} (Q(F) - Q(F_0)) \geq \inf_{F \in \mathcal{F}: d_{\mathrm{LP}}(F, F_0) \geq \varepsilon} (Q(F) - Q(F_0)) \\ &= \inf_{F \in \mathcal{F}: d_{\mathrm{LP}}(F, F_0) \geq \varepsilon} E[||m(x, F)||_E^2/J] > 0, \end{aligned}$$

where the first inequality holds because $\mathcal{F}_{R(N)} \subseteq \mathcal{F}$ by construction and the second, strict inequality holds because the model is identified (Assumption 1.4).[11]

Next we consider (A.3.2). First note that $(\mathcal{F}, d_{\mathrm{LP}})$ is a metric space and we have $\mathcal{F}_R \subseteq \mathcal{F}_{R+1} \subseteq \mathcal{F}$ for all $R \geq 1$ by construction of our sieve space. Then we claim that there exists a sequence of functions $\Pi_N F_0 \in \mathcal{F}_{R(N)}$ such that $d_{\mathrm{LP}}(\Pi_N F_0, F_0) \to 0$ as $N \to \infty$. First, $\mathcal{B}_{R(N)}$ becomes dense in $\mathcal{B}$ by assumption. Second, $\mathcal{F}_{R(N)}$ becomes dense in $\mathcal{F}$ because the set of distributions on a dense subset $\mathcal{B}_{R(N)} \subset \mathcal{B}$ is itself dense. To see this, remember that the class of all distributions with finite support is dense in the class of all distributions (Aliprantis and Border 2006, Theorem 15.10). Any distribution with finite support can be approximated using a finite support in a dense subset $\mathcal{B}_{R(N)}$ (Huber 2004).

Next, to show (A.3.3) holds, we use Remark B.1.(1)(a) of CP. First note that $\mathcal{F}_R$ is a compact subset of $\mathcal{F}$ for each R because $\mathcal{B}_R$ is a compact subset of $\mathcal{B}$.[12] Second we need to show that for any

[11] As we discussed in the main text the space $\mathcal{F}$ of distributions on $\mathcal{B}$ is compact in the weak topology because we assume $\mathcal{B}$ itself is compact.

[12] Alternatively we can also see that $\mathcal{F}_R$ is compact because the simplex, $\Delta_{R(N)}$, itself is compact as we argue below. For any given R and $\mathcal{B}_R$, consider two metric spaces, $(\mathcal{F}_R, d_{\mathrm{LP}})$ and $(\Delta_R, ||\cdot||_E)$. Then we can define a continuous map $\psi : \Delta_R \to \mathcal{F}_R$ because any element in Δ_R determines an element in $\mathcal{F}_R$. The map is continuous in the sense that for

data $\{(y_i,x_i)\}_{i=1}^N$, $\hat{Q}_N(F)$ is lower semicontinuous on $\mathcal{F}_R$ for each $R \geq 1$. Since $\mathcal{F}_R$ is compact, this lower semicontinuity or continuity means our estimator is well defined as the minimum in (6). Note $\int g(x,\beta)\, dF_l = \sum_{r=1}^R \theta_l^r g(x,\beta^r)$ for $F_l \in \mathcal{F}_R, l = 1,2$. Then, for any $F_1, F_2 \in \mathcal{F}_{R(N)}$, applying the triangle inequality, we obtain

$$\begin{aligned}
|\hat{Q}_N(F_1) - \hat{Q}_N(F_2)| &\leq 2\sum\nolimits_{i=1}^N \sum\nolimits_{j=1}^{\dim(y_i)} y_{i,j} \left| \int g_j(x_i,\beta)(dF_1 - dF_2) \right| /NJ \\
&+ \sum\nolimits_{i=1}^N \sum\nolimits_{j=1}^{\dim(y_i)} \{ \int g_j(x_i,\beta)(dF_1 + dF_2) \} \left| \int g_j(x_i,\beta)(dF_1 - dF_2) \right| /NJ \\
&\leq 4\sum\nolimits_{i=1}^N \sum\nolimits_{j=1}^{\dim(y_i)} \left| \int g_j(x_i,\beta)(dF_1 - dF_2) \right| /NJ,
\end{aligned}$$

where the second inequality holds because $y_{i,j}$, $g_j(x_i,\beta)$, and $\int g_j(x_i,\beta)dF(\beta)$ are uniformly bounded by 1 for all j and x_i. Then because $g_j(x_i,\beta)$ is uniformly bounded by 1 and F_1 and F_2 are discrete distributions with the finite support $\mathcal{B}_R$, in this case weak convergence implies that almost surely $\hat{Q}_N(F)$ is continuous on $\mathcal{F}_R$, i.e. for any $F_1, F_2 \in \mathcal{F}_R$ such that $d_{LP}(F_1,F_2) \to 0$, it follows that $|\hat{Q}_N(F_1) - \hat{Q}_N(F_2)| \to 0$ almost surely.[13] Continuity is stronger than lower semicontinuity. Therefore (A.3.3) holds by Remark B.1.(1) (a) of CP.

Next there are two conditions to verify in (A.3.4). We first focus on the uniform convergence of $\hat{Q}_N(F)$ to $Q(F)$ for $F \in \mathcal{F}_{R(N)}$, $\sup_{F\in\mathcal{F}_{R(N)}} |\hat{Q}_N(F) - Q(F)| \xrightarrow{p} 0$. It is convenient to view $\hat{Q}_N(F)$ and $Q(F)$ as functions of $\theta \in \Delta_{R(N)}$ and then write them as $\hat{Q}_N(\theta)$ and $Q(\theta)$, respectively. Then the uniform convergence condition to verify becomes

$$\sup_{\theta\in\Delta_{R(N)}} |\hat{Q}_N(\theta) - Q(\theta)| \xrightarrow{p} 0. \tag{11}$$

Using measures of complexity of spaces, let $\mathbf{N}(\varepsilon, \mathcal{T}, ||\cdot||)$ denote the covering number of the set $\mathcal{T}$ with balls of radius ε with an arbitrary norm $||\cdot||$ and let $\mathbf{N}_{[]}(\varepsilon, \mathcal{T}, \|\cdot\|)$ denote the bracketing number of the set $\mathcal{T}$ with ε-brackets. Define for any R, the class of measurable functions

$$\mathcal{G}_R = \{l(y,x,\theta) = ||y - \sum_r \theta^r g(x,\beta^r)||_E^2/J : \theta \in \Delta_R\}. \tag{12}$$

any sequence $\theta_n \to \theta$ in Δ_R we have $\psi(\theta_n) \to \psi(\theta)$ in $\mathcal{F}_R$. Then it is a simple proof to show that if Δ_R is compact, then $\mathcal{F}_R = \{\psi(\theta) : \theta \in \Delta_R\}$ is also compact.

Proof. Consider an arbitrary sequence $\{F_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}_R$. Since $F_n \in \{\psi(\theta) : \theta \in \Delta_R\}$ for all $n \in \mathbb{N}$, we know that there exists $\theta_n \in \Delta_R$ with $\psi(\theta_n) = F_n$ for all $n \in \mathbb{N}$. Then $\{\theta_n\}_{n\in\mathbb{N}} \subseteq \Delta_R$. Next note that since Δ_R is compact, there exists some subsequence $\{\theta_{l_n}\}_{n\in\mathbb{N}}$ with $\theta_{l_n} \to \bar{\theta} \in \Delta_R$. Since the map ψ is continuous, it follows that $\psi(\theta_{l_n}) \to \psi(\bar{\theta})$. And because $\psi(\theta_{l_n}) = F_{l_n}$, then $F_{l_n} \to \psi(\bar{\theta}) \in \mathcal{F}_R$ because $\bar{\theta} \in \Delta_R$. Therefore, we conclude $\mathcal{F}_R$ is also compact when Δ_R is compact. □

[13] Note that $\left|\int g_j(x_i,\beta)(dF_1 - dF_2)\right| \leq \sum_{r=1}^R |\theta_1^r - \theta_2^r| \to 0$ as $d_{LP}(F_1,F_2) \to 0$ for any finite R.

Note (i) $\hat{Q}_N(\theta) = N^{-1} \sum_{i=1}^{N} l(y_i, x_i, \theta)$, (ii) $\{(y_i, x_i)\}_{i=1}^{N}$ are i.i.d., and (iii) $E[\sup_{\theta \in \Delta_{R(N)}} |l(y, x, \theta)|] \leq 1 < \infty$. Then by Pollard (1984, Theorem II.24) (also see Chen (2007, Section 3.1, page 5592) for related discussion), the uniform convergence (11) holds if and only if $\log \mathbf{N}(\varepsilon, \mathcal{G}_R, || \cdot ||_{L_{1,N}})/N \xrightarrow{p} 0$ for all $\varepsilon > 0$, where $|| \cdot ||_{L_{1,N}}$ denotes the $L_1(\mathbb{P}_N)$-norm and $\mathbb{P}_N$ denotes the empirical measure of the data $\{(y_i, x_i)\}_{i=1}^{N}$.

The term $l(y, x, \theta)$ is Lipschitz in θ, as

$$\begin{aligned} |l(y,x,\theta_1) - l(y,x,\theta_2)| &\leq \frac{1}{J} \sum_{j=1}^{\dim(y)} (2y_j \sum_r g_j(x, \beta^r)|\theta_1^r - \theta_2^r| + \sum_r g_j(x, \beta^r)(\theta_1^r + \theta_2^r) \sum_r g_j(x, \beta^r)|\theta_1^r - \theta_2^r|) \\ &\leq M(\cdot) \sum_{r=1}^{R} |\theta_1^r - \theta_2^r| \leq M(\cdot)\sqrt{R}||\theta_1 - \theta_2||_E \end{aligned}$$

with some function $E[M(\cdot)^2] < \infty$. The first inequality is obtained by the triangle inequality and the third inequality holds due to the Cauchy-Schwarz inequality. We also know Δ_R is a compact subset of $\mathbb{R}^R$. Now take $M(\cdot) = 4$, noting that y_j, $g_j(\cdot)$, and $\sum_{r=1}^{R} g_j(x, \beta^r)\theta^r$ are uniformly bounded by 1. Then from Theorem 2.7.11 of van der Vaart and Wellner (1996), we have $\mathbf{N}_{[]}\left(2\varepsilon, \mathcal{G}_R, \|\cdot\|\right) \leq \mathbf{N}\left(\frac{\varepsilon}{4\sqrt{R}}, \Delta_R, \|\cdot\|_E\right) = \left(\frac{4\sqrt{R}}{\varepsilon}\right)^R$ for any norm $\|\cdot\|$. Therefore as long as $R(N) \log R(N)/N \to 0$, the uniform convergence condition holds because $\mathbf{N}\left(\varepsilon, \mathcal{G}_R, \|\cdot\|_{L_{1,N}}\right) \leq \mathbf{N}_{[]}\left(2\varepsilon, \mathcal{G}_R, \|\cdot\|_{L_{1,N}}\right) \leq \left(\frac{4\sqrt{R}}{\varepsilon}\right)^R$ (van der Vaart and Wellner 1996, page 84).

To satisfy the second condition in (A.3.4), we need to bound all three terms in the max$\{\cdot\}$ function. We have shown the uniform convergence of the sample criterion function (this also satisfies the first condition in (A.3.4.)) and we can take ν_N to be small enough. We also have $|Q(\Pi_N F_0) - Q(F_0)| \to 0$, which is trivially satisfied by the continuity of $Q(F)$ in F and $d_{\text{LP}}\left(\Pi_N F_0, F_0\right) \to 0$. Therefore because $\liminf_{N\to\infty} \delta\left(N, R(N), \varepsilon\right) > 0$, the condition (A.3.4) is satisfied.

We have verified all the conditions in Lemma 3 (Lemma B.1 of CP) and this completes the consistency proof.

A.1 Proof of Corollary 1

We verify corresponding conditions in Theorem 1 of Chen, Linton, and van Keilegom (2003). Although their Theorem 1 is written in terms of moment based estimations, it can be easily modified to "M-estimators" that include our estimator (as noted in their footnote 3).

Condition (1.1) (extremum estimator) is satisfied by definition of our estimator as an extremum estimator in (8). Condition (1.2) (identification) is satisfied by Assumption 2.2 because

$$\inf_{\alpha \in \mathcal{A}: ||\alpha - \alpha_0||_E \geq \epsilon} (Q(\alpha, F_0(\alpha)) - Q(\alpha_0, F_0)) > 0$$

where $F_0(\alpha) = \text{argmin}_{F \in \mathcal{F}} Q(\alpha, F)$ (which is well defined and exists because $\mathcal{F}$ is compact and $Q(\alpha, F)$ is continuous on $\mathcal{F}$), $F_0 = F_0(\alpha_0)$, and again the minimum is attained because $Q(F, \alpha)$ is continuous in α and F and $\mathcal{A} \times \mathcal{F}$ is compact.

Condition (1.3) (continuity) holds by Assumption 2.3. Condition (1.4) (consistency for the nonparametric component) holds by our main Theorem 1, where we treat $F_0(\alpha)$ as the true parameter for each $\alpha \in \mathcal{A}$.

The last condition we need to verify is their Condition (1.5) (uniform convergence of the sample criterion function). We have verified the uniform convergence of the sample criterion function over $\mathcal{F}_{R(N)}$ in the proof of Theorem 1. Here we need to verify the uniform convergence over $\mathcal{A} \times \mathcal{F}_{R(N)}$ such that

$$\sup_{(\alpha,F)\in\mathcal{A}\times\mathcal{F}_{R(N)}} \left|\hat{Q}_N(\alpha,F) - Q(\alpha,F)\right| \xrightarrow{p} 0. \tag{13}$$

For this purpose define for any R, the class of measurable functions (similar to (12))

$$\mathcal{G}_R^\alpha = \{l(y,x,\theta,\alpha) = ||y - \sum_r \theta^r g(x,\beta^r,\alpha)||_E^2/J : (\alpha,\theta) \in \mathcal{A} \times \Delta_R\}$$

and note that $\hat{Q}_N(\alpha,F) = N^{-1}\sum_{i=1}^N l(y_i,x_i,\theta,\alpha)$. Then again by Pollard (1984, Theorem II.24), the uniform convergence (13) holds if and only if the entropy satisfies $\log \mathbf{N}(\varepsilon, \mathcal{G}_R^\alpha, ||\cdot||_{L_1,N}) = o_p(N)$ for all $\varepsilon > 0$. Note that the entropy measure of $\mathcal{G}_R^\alpha$ is bounded by the sum of two entropies, one associated with $\mathcal{F}_{R(N)}$ and the other one associated with $\mathcal{A}$. We have shown that the former is $o_p(N)$ in the proof of Theorem 1. We also note that the latter satisfies the entropy condition (and so is $o_p(N)$) under Assumption 2.4 by Theorem 2.7.11 of van der Vaart and Wellner (1996) (for the Lipschitz case) and because the class of indicator functions belong to the Vapnik-Červonenkis class and has a uniformly bounded entropy (Theorem 2.6.7 of van der Vaart and Wellner 1996).

B Proof of Lemma 1

To establish continuity we need to show that for any $F_1, F_2 \in \mathcal{F}$ such that $d_{LP}(F_1,F_2) \to 0$ we have $|Q(F_1) - Q(F_2)| \to 0$. Obtain for any $F_1, F_2 \in \mathcal{F}$ using a similar derivation to (10)

$$\begin{aligned} &|Q(F_1) - Q(F_2)| = |Q(F_1) - Q(F_0) - \{Q(F_2) - Q(F_0)\}| \\ &= \left|E\left[\left\{\int g_1(x_i,\beta)(dF_1 - dF_0)\right\}^2 - \left\{\int g_1(x_i,\beta)(dF_2 - dF_0)\right\}^2\right]\right| \\ &= \left|E\left[\left\{\int g_1(x_i,\beta)(dF_1 + dF_2 - 2dF_0)\right\}\int g_1(x_i,\beta)(dF_1 - dF_2)\right]\right|. \end{aligned} \tag{14}$$

By Assumption 1.3 (independence of random coefficients from x_i), applying the law of iterated expectation, we can rewrite

$$\begin{aligned} & E\left[\left\{\int g_1(x_i,\beta)(dF_1+dF_2-2dF_0)\right\}\int g_1(x_i,\beta)(dF_1-dF_2)\right] \\ = & E\left[\int\left\{\int g_1\left(x_i,\tilde{\beta}\right)(d\tilde{F_1}+d\tilde{F_2}-2d\tilde{F_0})\right\}g_1(x_i,\beta)(dF_1-dF_2)\right] \\ = & \int E\left[\left\{\int g_1\left(x_i,\tilde{\beta}\right)g_1(x_i,\beta)(d\tilde{F_1}+d\tilde{F_2}-2d\tilde{F_0})\right\}\right](dF_1-dF_2) \qquad (15) \\ \equiv & \int h(\beta)dF_1-\int h(\beta)dF_2, \qquad (16) \end{aligned}$$

where we use $\tilde{\beta}$ instead of β to emphasize that $\tilde{\beta}$ is not subject to the outer integral in (15) and we use $(\tilde{F_1},\tilde{F_2},\tilde{F_0})=(F_1,F_2,F_0)$ to emphasize that β is not subject to the inner integral inside the expectation in (15).

Now we further analyze the expectation function denoted by $h(\beta)$ as a function of β in (15). Below we will show that this function $h(\beta)$ is continuous in β and is also bounded. Therefore by weak convergence, for any $F_1,F_2\in\mathcal{F}$ such that $d_{LP}(F_1,F_2)\to 0$, we also have $|Q(F_1)-Q(F_2)|\to 0$ by (14) and (16). This will complete the proof of the continuity of $Q(F)$ on $\mathcal{F}$ in the weak topology.

By Assumption 1.3 (independence of random coefficients from x_i) and Assumption 3, applying the law of iterated expectation several times and assuming (w.l.o.g.) that the support of β_{k^*} is known to take strictly positive values (or normalize the coefficient to be 1), we can write

$$\begin{aligned} h(\beta) &= E_{x_{i,-k^*}}\left[E_{x_{i,k^*}|x_{i,-k^*}}\left[\int g_1\left(x_i,\tilde{\beta}\right)g_1(x_i,\beta)(d\tilde{F_1}+d\tilde{F_2}-2d\tilde{F_0})\right]\right] \\ &= E_{x_{i,-k^*}}\left[\int E_{x_{i,k^*}|x_{i,-k^*}}\left[g_1\left(x_i,\tilde{\beta}\right)g_1(x_i,\beta)\right](d\tilde{F_1}+d\tilde{F_2}-2d\tilde{F_0})\right] \\ &= \int_{\text{support}(x_{i,-k^*})}\int \tilde{G}_{x_{i,k^*}|x_{i,-k^*}}(d\tilde{F_1}+d\tilde{F_2}-2d\tilde{F_0})dG_{x_{i,-k^*}} \qquad (17) \end{aligned}$$

where we denote[14]

$$\begin{aligned} & \tilde{G}_{x_{i,k^*}|x_{i,-k^*}}=\Pr\left(x_i'\beta\geq 0 \text{ and } x_i'\tilde{\beta}\geq 0\middle| x_{i,-k^*}\right) \\ = & \Pr\left(x_{i,k^*}\geq\max\left\{-x_{i,-k^*}'\beta_{-k^*}/\beta_{k^*},-x_{i,-k^*}'\tilde{\beta}_{-k^*}/\tilde{\beta}_{k^*}\right\}\right) \\ = & 1-G_{x_{i,k^*}|x_{i,-k^*}}\left(\max\left\{-x_{i,-k^*}'\beta_{-k^*}/\beta_{k^*},-x_{i,-k^*}'\tilde{\beta}_{-k^*}/\tilde{\beta}_{k^*}\right\}\right) \end{aligned}$$

and $G_{x_{i,-k^*}}$ denotes the CDF of $x_{i,-k^*}$. Note that $\tilde{G}_{x_{i,k^*}|x_{i,-k^*}}$ is continuous in β for given others ($x_{i,-k^*}$ and $\tilde{\beta}$). Then note that because the function (integrand) inside the inner integral in (17) is measurable in $x_{i,-k^*}$, continuous in β, and bounded, the inner integral itself has these properties itself. Therefore,

[14] Let the support of x_{i,k^*} be support$(x_{i,k^*})=[\underline{x}_{k^*},\bar{x}_{k^*}]$ (which can also depend on $x_{i,-k^*}$). Then we let $G_{x_{i,k^*}|x_{i,-k^*}}(c)=0$ for $c<\underline{x}_{k^*}$ and $G_{x_{i,k^*}|x_{i,-k^*}}(c)=1$ for $c>\bar{x}_{k^*}$.

applying the dominated convergence theorem we conclude that the function $h(\beta)$ is continuous in β and bounded. This completes the proof.

C Proof of Lemma 2

Fox and Gandhi (2010, Theorem 3.3 and Lemma 5.1) prove that the distribution $F_0(\beta)$ is identified in the sense of Assumption 1.4 whenever the following two conditions hold (using the notation of interval censored regression):

1. For any finite set of distinct random coefficients $S = \left\{\beta^1, \ldots, \beta^{|S|}\right\}$, $|S| < \infty$, there exists a pair (j, x) such that exactly one $\beta \in S$ satisfies $x'\beta \in [a_{j-1}, a_j)$.

2. There is a neighborhood $\tilde{\mathcal{X}}$ of x such that the first condition holds.

By the continuity of the term $x'\beta$ in x, it is clear that the second condition holds whenever the first condition holds. Therefore, we focus on verifying the first condition. In the first condition, S is not the true set of types in the data generating process, but any arbitrary, finite set of types.

Let S be given and focus on the lowest interval, $[a_0, a_1)$. We start with a point $x^\star$ where at least one $\beta \in S$ satisfies $(x^\star)'\beta \in [a_0, a_1)$. Linearity implies such a point exists. If there is only one $\beta \in S$ that satisfies $(x^\star)'\beta \in [a_0, a_1)$, we are done. So consider the case where two or more $\beta \in S$ satisfy $(x^\star)'\beta \in [a_0, a_1)$.

Divide $x = (x_1, x_{-1})$ into the first x_1 and all other covariates x_{-1}. Let

$$\tilde{x}_1(a_1, x_{-1}, \beta) \equiv \frac{1}{\beta_1}\left(a_1 - x_{-1}'\beta_{-1}\right)$$

be the value of x_1, for $\beta \in S$ and x_{-1}, that makes $x'\beta = a_1$ and hence makes β take on a value outside the interval $[a_0, a_1)$. As any two $\beta \in S$ differ by definition, $\tilde{x}_1(a_1, x_{-1}, \beta)$ is a distinct, linear function of x_{-1} for every β. By the properties of linear or, more generally, multivariate real analytic functions, there exists a point $x_{-1}^{\star\star}$ in a neighborhood of $x_{-1}^\star$ where $\tilde{x}_1\left(a_1, x_{-1}^{\star\star}, \beta_s\right) \neq \tilde{x}_1\left(a_1, x_{-1}^{\star\star}, \beta_t\right)$ for all $\beta_s \neq \beta_t \in S$ (Krantz and Parks 2002). Set $\bar{\beta} = \arg\min_{\beta\in S} \tilde{x}_1\left(a_1, x_{-1}^{\star\star}, \beta\right)$ and

$$x_1^{\star\star} = \min_{\beta\in S} \tilde{x}_1\left(a_1, x_{-1}^{\star\star}, \beta\right) - \epsilon$$

for sufficiently small $\epsilon > 0$ when $\bar{\beta}_1 > 0$ and sufficiently small (in absolute value) $\epsilon < 0$ when $\bar{\beta}_1 < 0$. Then only $\bar{\beta}$ satisfies $(x^{\star\star})'\beta \in [a_0, a_1)$ for $\beta \in S$ at the point $x^{\star\star} = \left(x_1^{\star\star}, x_{-1}^{\star\star}\right)$. Thus, the first condition above is satisfied.

References

[1] Ackerberg, D. (2009), "A New Use of Importance Sampling to Reduce Computational Burden in Simulation Estimation", *Quantitative Marketing and Economics*, 7(4), 343–376.

[2] Aliprantis, C.D. and K.C. Border (2006), *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer.

[3] Andrews D.K.W. (2002), "Generalized Method of Moments Estimation When a Parameter is on a Boundary,*" Journal of Business & Economics Statistics* 20-4, 530–544.

[4] Bajari, P., J.T.. Fox, K. Kim and S. Ryan (2010), "The Random Coefficients Logit Model Is Identified", University of Minnesota working paper.

[5] Beran, R. and Millar, P.W. (1994), "Minimum Distance Estimation in Random Coefficient Regression Models", *The Annals of Statistics*, 22(4), 1976–1992.

[6] Biernacki, C., G, Celeux and G. Govaert (2003), "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models," *Computational Statistics & Data Analysis*, 41, 561–575.

[7] Böhning, D. "Convergence of Simar's Algorithm for Finding the Maximum Likelihood Estimate of a Compound Poisson Process", *The Annals of Statistics*, 10(3), 1006–1008. 1982.

[8] Carrasco, M, J.P. Florens, and E. Renault (2007), "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization", *Handbook of Econometrics*, V6B, Elsevier.

[9] Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models,*" Handbook of Econometrics* V7, Elsevier.

[10] Chen, X, O. Linton, and I. van Keilegom (2003), "Estimation of Semiparametric Models When the Criterion Function is Not Smooth", *Econometrica*, 71-5, 1591-1608.

[11] Chen, X. and D. Pouzo (2009), "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Moments", Yale working paper.

[12] Dempster, A.P., N.M. Laird and D.B. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39, 1, 1-38.

[13] Fox, J.T. and A. Gandhi (2010), "Nonparametric Identification and Estimation of Random Coefficients in Nonlinear Economic Models", University of Michigan working paper.

[14] Fox, J.T. and A. Gandhi (2011), "Using Selection Decisions to Identify the Joint Distribution of Outcomes", University of Michigan working paper.

[15] Fox J.T., K. Kim, S. Ryan, and P. Bajari (2011), "A Simple Estimator for the Distribution of Random Coefficients", *Quantitative Economics*, forthcoming.

[16] Heckman, J. (1990), "Varieties of Selection Bias", *The American Economic Review*, 1990, 80 (2), 313–318.

[17] Heckman, J. and Singer, B. “Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data*”, Econometrica* 52(2), 271-320.

[18] Horowitz, J. L. (1999), “Semiparametric Estimation of a Proportional Hazard Model with Unobserved Heterogeneity”, *Econometrica*, 67(5), 1001–1028.

[19] Huber, J. (1981, 2004) *Robust Statistics,* Wiley.

[20] Ichimura, H. and T.S. Thompson (1998), “Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution,*” Journal of Econometrics*, 86(2), 269–295.

[21] Kamakura, W.A. (1991), “Estimating flexible distributions of ideal-points with external analysis of preferences”, *Psychometrika,* 56, 3, 419-431.

[22] Karlis, D. and E. Xekalaki (2003), “Choosing initial values for the EM algorithm for finite mixtures”, *Computational Statistics & Data Analysis*, 41, 577–590.

[23] Krantz, S.G. and H.R. Parks (2002), *A Primer on Real Analytic Functions,* second edition, Birkhauser.

[24] Laird, N. (1978), “Nonparametric Maximum Likelihood Estimation of a Mixing Distribution*”, Journal of the American Statistical Association*, Vol. 73, No. 364, pp. 805–811.

[25] Li, J.Q. and A.R. Barron (2000), “Mixture density estimation”, *Advances in Neural Information Processing Systems*, Vol. 12, pp. 279–285.

[26] Lindsay, B.G. (1983) “The Geometry of Mixture Likelihoods: A General Theory”, *The Annals of Statistics*, 11(1), 86–94.

[27] Matzkin, R.L. (2007) “Heterogeneous Choice”, *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress of the Econometric Society.* Cambridge.

[28] McLachlan, G.J. and D. Peel (2000), *Finite Mixture Models.* Wiley.

[29] Parthasarathy, K.R. (1967), *Probability Measures on Metric Spaces,* Academic Press.

[30] Pilla, R.S. and B.G. Lindsay (2001), “Alternative EM methods for nonparametric finite mixture models”, *Biometrika*, 88, 2, 535–550.

[31] Pollard, D. (1984), *Convergence of Statistical Processes*, Springer-Verlag, New York.

[32] Quandt, R.E. and Ramsey, J.B. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, 73, 364, 730-738.

[33] Rust, J. (1987), “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher*”, Econometrica*, 55(5): 999–1033.

[34] Seidel, W., K. Mosler and M. Alker (2000), “A Cautionary Note on Likelihood Ratio Tests in Mixture Models”, *Annals of the Institute of Statistical Mathematics*, 52, 3, 418-487,

[35] Teicher, H. (1963), "Identifiability of Finite Mixtures*", Annals of Mathematical Statistics*, 34, 1265-1269.

[36] Van der Vaart, W. and J.A. Wellner (1996), *Weak Convergence and Empirical Processes*, Springer Series in Statistics, New York.

[37] Verbeek, J.J., N. Vlassis and B. Kröse (2003), "Efficient Greedy Learning of Gaussian Mixture Models*", Neural Computation*, Vol. 15, pp 469–485.