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

Jeremy T. Fox
Rice University & NBER

Kyoo il Kim*
Michigan State University

Chenyu Yang
University of Rochester

May 2016

Abstract

We explore least squares and likelihood nonparametric mixtures estimators of the joint distribution of random coefficients in structural models. The estimators fix a grid of heterogenous parameters and estimate only the weights on the grid points, an approach that 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 most of the consistency conditions for three discrete choice models. We also derive the convergence rates of the least squares nonparametric mixtures estimator under additional restrictions. We perform a Monte Carlo study on a dynamic programming model.

Keywords: Random coefficients, mixtures, discrete choices, dynamic programming, sieve estimation

*Corresponding authors: Jeremy Fox at jeremyfox@gmail.com and Kyoo il Kim at kyookim@msu.edu.

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 heterogeneous parameters. Our paper establishes the consistency and rates of convergence of "fixed grid" nonparametric estimators for a distribution of heterogeneous parameters due to Bajari, Fox and Ryan (2007), Train (2008, Section 6), Fox, Kim, Ryan and Bajari (2011), and Koenker and Mizera (2014). These estimators are computationally simpler than some alternatives. We use FKRB to refer to Fox, Kim, Ryan and Bajari (2011).

We estimate the distribution of heterogeneous parameters $F(\beta)$ in the model

$$P_j(x) = \int g_j(x, \beta)\, dF(\beta), \tag{1}$$

where j is the index of the jth out of J finite values of the outcome y, x is a vector of observed explanatory variables, β is the vector of heterogeneous parameters, and $g_j(x, \beta)$ is the probability that the jth outcome occurs for an observation with heterogeneous parameters β and explanatory variables x. Given this structure, $P_j(x)$ is the cross sectional probability of observing the jth outcome when the explanatory variables are x. The researcher picks $g_j(x, \beta)$ as the underlying model, has an i.i.d. sample of N observations (y_i, x_i), and wishes to estimate $F(\beta)$. As F is only restricted to be a valid CDF, the mixture model (1) is nonparametric.

The unknown distribution $F(\beta)$ enters (1) linearly. The estimators we analyze exploit linearity and achieve a computationally simpler estimator than some alternatives. All the fixed grid estimators divide the support of the vector β into a finite and known grid of vectors $\beta^1, \ldots, \beta^R$. Computationally, the unknown parameters are the weights $\theta^1, \ldots, \theta^R$ on the R grid points. These can be estimated using a least squares or likelihood criterion with the constraints that each $\theta^r \geq 0$ and that $\sum_{r=1}^R \theta^r = 1$. 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 and $1\left[\beta^r \leq \beta\right]$ is equal to 1 when $\beta^r \leq \beta$. Computationally, the least squares and likelihood constrained optimization problems are globally convex and concave, respectively. Particular numerical algorithms are guaranteed to converge to a global optimum.

FKRB discuss the advantages of this estimator for complex structural models, like dynamic programming models with heterogeneous parameters. In this respect, fixed grid estimators share some computational advantages with the parametric approach in Ackerberg (2009). Our Monte Carlo study in an online appendix is to a discrete choice, dynamic programming model.

FKRB and other previous analyses assume that the R grid points used in a finite sample are indeed the true grid points that contain the finite support of the true $F_0(\beta)$. Thus, the true distribution $F_0(\beta)$ is assumed to be known up to a finite number of weights $\theta^1, \ldots, \theta^R$. 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.

Instead of assuming that the distribution is known up to weights $\theta^1, \ldots, \theta^R$, this paper requires 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 approaches are 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 heterogeneous parameters β.

In a finite sample of N observations, our estimators are still implemented by choosing a fixed 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 $R(N)$, our least squares and likelihood estimators $\hat{F}_N(\beta)$ converge to the true $F_0(\beta)$ as $N \to \infty$, in a function space. We use the Lévy-Prokhorov metric, a common metrization of the weak topology on the space of multivariate distributions.

We recognize that the nonparametric versions of our estimators are special cases of sieve estimators (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 estimators' choice of basis, a finite grid of points, is justified by the estimators' computational simplicity. Further and unlike a typical sieve estimator, we need to constrain our estimated functions to be valid distribution functions. Our constrained least squares and likelihood approaches are both computationally simple and ensure that the estimated CDFs satisfy the theoretical properties of a valid CDF.

Because our estimators are sieve estimators, we prove their consistency by satisfying high-level conditions for the consistency of sieve extremum estimators, as given in an appendix lemma in Chen and Pouzo (2012). We repeat this lemma and its proof in our paper so our consistency proof is self-contained. Our fixed grid estimators are not a special case of the two-step sieve estimators explored using lower-level conditions in the main text of Chen and Pouzo.[1]

We prove the consistency of our estimators for the distribution of heterogeneous parameters, in function space under the weak topology. We present separate theorems for mixtures of discrete grid points and mixtures of continuous densities with a grid of points over the parameters of each density. The theorem for the mixture of grid points requires the heterogeneous parameters to lie in a, not necessarily known, compact set. The theorem for a mixture of continuous densities allows for unbounded support of the heterogeneous parameters. Our consistency theorems are not specific to the economic model being estimated.

We provide the rate of convergences for a subset of the models handled by our consistency theorem, namely those that are differentiable in the heterogeneous parameters, which include the random

[1]Note that 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.

coefficients logit model. The convergence rates, the asymptotic estimation error bounds, consist of two terms: the bias and the variance. While obtaining the variance term is rather standard in the sieve estimation literature, deriving the bias term depends on the specific approximation methods (e.g., power series or splines). Because our use of approximating functions is new in the sieve estimation literature, deriving the bias term is not trivial. We provide the bias term, which is the smallest possible approximation error of the true function using sieves for the class of models we consider.

Our rate of convergence results highlight an important practical issue with any nonparametric estimator: there is a curse of dimensionality in the dimension of the heterogeneous parameters. Larger sample sizes will be needed if the vector of heterogeneous parameters has more elements. Further, the rate results indicate that our baseline estimator is not practical when there is a large number of heterogeneous parameters. In high dimensional settings, we suggest allowing heterogeneous parameters on only a subset of explanatory variables and estimating homogenous parameters on the remaining explanatory variables. We extend our consistency result to models where some parameters are homogeneous. However, including homogeneous parameters requires nonlinear optimization, which loses some of the computational advantages of our estimators.

We provide a Monte Carlo study in an online appendix. We estimate a dynamic programming, discrete choice model, adding heterogeneous parameters to the framework of Rust (1987). The dynamic programming problem must be solved once for each realization of the heterogeneous parameters. We present results for both the fixed grid likelihood and least squares estimators as well as, for comparison, a likelihood estimator where we estimate both the grid of points and the weights on those points. We show that our fixed grid estimators have superior speed but inferior statistical accuracy compared to the more usual approach of estimating a flexible grid.

The outline of our paper is as follows. Section 2 presents three examples of discrete choice mixture models. Section 3 introduces the estimation procedures. Section 4 demonstrates consistency of our estimators in the space of multivariate distributions. Section 5.1 extends our consistency results to models with both heterogeneous parameters and homogeneous parameters and Section 5.2 considers mixtures of smooth basis densities. Section 6 verifies most of the primitive conditions for consistency established in Section 4 using the three examples of mixture models in Section 2. Section 7 derives the convergence rates of the nonparametric estimator for a class of models. Finally, an online appendix presents the Monte Carlo study.

2 Examples of Mixture Models

In our framework, the object the econometrician wishes to estimate is $F(\beta)$, the distribution of the vector of heterogeneous parameters β. One definition of identification is that a unique $F(\beta)$ solves (1) for all x and all outcomes $j = 1, \ldots, J$. This is the definition used in certain relevant papers on identification in the statistics literature, for example Teicher (1963).

We will return to these three examples of discrete choice models later in the paper. Each example considers economic models with heterogeneous parameters 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 6 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 unordered choices, such as types of cars for sale. For $j \geq 2$, 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 1 with utility $u_{i,1} = \varepsilon_{i,1}$. 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 $j \geq 2$,

$$P_j(x) = \int g_j(x, \beta) \, dF(\beta) = \int \frac{\exp\left(x'_j \beta\right)}{1 + \sum_{h=2}^{J} \exp\left(x'_h \beta\right)} dF(\beta),$$

where $x = (x_2, \ldots, x_J)$. 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 $J = 2$ in the previous example, so that there is one inside good and one outside good. Thus, the utility of the inside good 2 is $u_{i,2} = \epsilon_i + x'_i \beta_{2,i}$, where $\beta_i = (\epsilon_i, \beta_{2,i})$ 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 1 has utility $u_{i,1} = 0$. In this example, (1) becomes, for $j = 2$,

$$P_2(x) = \int g_2(x, \beta) \, dF(\beta) = \int 1\left[\epsilon + x'\beta_2 \geq 0\right] dF(\beta),$$

where $1[\cdot]$ is the indicator function equal to 1 if the inequality in the brackets is true. Without logit errors, the joint distribution of both the intercept and the slope coefficients is estimated nonparametrically. In this example, $g_2(x, \beta)$ 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 $j \geq 2$ is $u_{i,j} = x'_{i,j}\tilde{\beta}_i + \epsilon_{i,j}$ and the utility of the outside good 1 is $u_{i,1} = 0$. The notation $\tilde{\beta}_i$ is used because the full heterogeneous parameter vector is now $\beta_i = \left(\tilde{\beta}_i, \epsilon_{i,2}, \ldots, \epsilon_{i,J}\right)$, which is seen as one long vector. We will not assume that the additive errors $\epsilon_{i,j}$ are distributed independently of β_i or of each other. In this example, (1) becomes, for $j \geq 2$,

$$P_j(x) = \int g_j(x, \beta) \, dF(\beta) = \int 1\left[x'_j\tilde{\beta} + \epsilon_j \geq \max\left\{0, x'_h\tilde{\beta} + \epsilon_h\right\} \, \forall \, h \neq j, \, h \geq 2\right] dF(\beta).$$

3 Estimator

We analyze both least squares (linear probability models) and maximum likelihood criteria. We first discuss the least squares criterion, from FKRB. Recall that $y_{i,j}$ is equal to 1 whenever the outcome y_i for the ith observation is j, and 0 otherwise. Start with the model (1) and add $y_{i,j}$ to both sides while

moving $P_j(x)$ to the right side. For the statistical observation i, this gives

$$y_{i,j} = \int g_j(x_i, \beta)\, dF(\beta) + (y_{i,j} - P_j(x_i)). \tag{2}$$

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, (2) becomes

$$y_{i,j} \approx \sum_{r=1}^{R(N)} \theta^r g_j(x_i, \beta^r) + (y_{i,j} - P_j(x)). \tag{3}$$

We use the $\approx$ symbol to emphasize that (3) 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" $z^r_{i,j} = g_j(x_i, \beta^r)$.

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

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

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 heterogeneous parameters as

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

Thus, we have a structural estimator for a distribution of heterogeneous parameters in addition to a flexible method for approximating choice probabilities.

Following Train, we can also use the log-likelihood criterion (divided by the sample size)

$$\mathscr{L} = \sum_{i=1}^{N} \log \left(\sum_{r=1}^{R(N)} \theta^r z^r_{i,y_i} \right) / N,$$

where the z^r_{i,y_i} are the probabilities computed above for the observed outcome y_i for observation i. As with the least squares criterion, we enforce the constraints $\theta^r \geq 0 \, \forall \, r = 1, \ldots, R(N)$ and $\sum_{r=1}^{R(N)} \theta^r = 1$. Computationally, one can use the EM algorithm in Train's Section 6 or a nonlinear, gradient-based search routine, which is available in most scientific packages. The performance of the gradient-based search routine will be improved if the gradient of the likelihood is provided to the solver in closed form. The sth element of that gradient is

$$\frac{\partial \mathscr{L}}{\partial \theta^s} = \sum_{i=1}^{N} \frac{z^s_{i,y_i}}{\sum_{r=1}^{R(N)} \theta^r z^r_{i,y_i}} / N.$$

The log likelihood is globally concave and any local maximum found will be the global maximum.

The fixed grid approach, whether based on a least squares or a likelihood criterion, has two main advantages over other approaches to estimating distributions of heterogeneous parameters. First, the approach is computationally simple: we can always find a global optimum and, by solving for $z^r_{i,j} = g_j(x_i, \beta^r)$ 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 estimators $\hat{F}_N(\beta)$ converge 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, $\frac{\exp\left(x'_{i,j}\beta^r\right)}{1+\sum_{h=2}^{J}\exp\left(x'_{i,h}\beta^r\right)} = g_j(x_i, \beta^r)$. To implement the least squares estimator, for each statistical observation i, the researcher computes $R \cdot J$ probabilities $z^r_{i,j} = \frac{\exp\left(x'_{i,j}\beta^r\right)}{1+\sum_{h=2}^{J}\exp\left(x'_{i,h}\beta^r\right)}$. This computation is done before optimization commences. The outcome for choosing the outside good 1 does not need to be included in the objective function, as $\sum_{j=1}^{J} g_j(x_i, \beta^r) = 1$. To implement the likelihood estimator, the researcher computes R probabilities z^r_{i,y_i} for each statistical observation i.

Remark 1. FKRB discuss the case of panel data on T periods. Let $y_i^T = (y_{i,1}, \ldots, y_{i,T})$ be the actual sequence of T outcomes for panel observation i. Similarly, let the explanatory variables be $(x_{i,1}, \ldots, x_{i,T})$. The likelihood criterion for panel data is

$$\mathscr{L} = \sum_{i=1}^{N} \log \left(\sum_{r=1}^{R(N)} \theta^r z^r_{i,y_i^T} \right) / N,$$

where $z^r_{i,y_i^T} = \prod_{t=1}^{T} g_{y_{i,t}}(x_{i,t}, \beta^r)$. We do not explore panel data in our theoretical results.

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 heterogeneous 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.

To prove consistency, we use the results for sieve estimators developed by Chen and Pouzo (2012), hereafter referred to as CP. We define a sieve space to approximate $\mathcal{F}$ as

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

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 grids 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 A.2, 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.

First, we consider the least squares criterion and then the likelihood criterion follows. 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 heterogeneous parameter β. Then we can define our sample criterion function for least squares as

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

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

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

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, if necessary.

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.

As a distance measure for distributions, we use the Lévy-Prokhorov metric, denoted by $d_{\text{LP}}(\cdot)$, 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 $(\mathcal{B}, d)$ with its Borel sigma algebra $\Sigma(\mathcal{B})$. We use the notation $d_{\text{LP}}(F_1, F_2)$, where the measures are implicit. This denotes the Lévy-Prokhorov metric $d_{\text{LP}}(\mu_1, \mu_2)$, where μ_1 and μ_2 are probability measures corresponding to F_1 and F_2.

The Lévy-Prokhorov metric is defined as

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

where $C \subseteq \mathcal{B}$ 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 $d_{\text{LP}}(\mu_1, \mu_2) = 0$ only when $\mu_1 = \mu_2$. See Huber (1981, 2004).

The following assumptions are on the economic model and data generating process. We write $P(x, F) = \int g(x, \beta) dF(\beta)$.

Assumption 1.

1. *Let $\mathcal{F}$ be a space of distribution functions on a finite-dimensional real space $\mathcal{B}$. $\mathcal{F}$ is compact in the weak topology and contains the true 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 $F_1 \neq F_0$, $F_1 \in \mathcal{F}$, the set $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ where $P(x, F_0) \neq P(x, F_1)$ has a positive measure in $\mathcal{X}$.*[2]
5. *$Q(F)$ is continuous on $\mathcal{F}$ in the $d_{\text{LP}}(\cdot, \cdot)$ metric.*

Assumption 1.1, the compactness of $\mathcal{F}$ is satisfied if $\mathcal{B}$ itself is compact in Euclidean space (Parthasarathy 1967, Theorem 6.4). Unfortunately, the compactness of $\mathcal{B}$ rules out some examples such as normal distributions of heterogeneous parameters. In part to address this, Section 5.2 provides a consistency theorem for a related estimator, which can use mixtures of normal distributions, where the support of the heterogeneous parameters is allowed to be $\mathbb{R}^K$. Assumptions 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 that occurs with positive probability. The assumption rules out so-called fragile identification that could occur at values of x with measure zero (such as identification at infinity). Assumptions 1.4 and 1.5 need to be verified for each economic model $g(x, \beta)$. We will discuss these assumptions for our three examples in Section 6.

Remark 2. 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 $d_{\text{LP}}(\cdot, \cdot)$ metric 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 $F_1 \in \mathcal{F}$ and such that $d_{\text{LP}}(F_1, F_2) \to 0$ 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.

[2] 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 are functions of other elements (e.g. polynomials or interactions).

Remark 3. If the support $\mathcal{B}$ is a finite set, the continuity of $Q(F)$ holds even when $g_j(x,\beta)$ is discontinuous, because in this case the Lévy-Prokhorov metric becomes equivalent to the total variation metric (see Huber 1981, p.34). This implies

$$\left| \int g_j(x,\beta) dF_1(\beta) - \int g_j(x,\beta) dF_2(\beta) \right| \to 0$$

for any $F_1, F_2 \in \mathcal{F}$ such that $d_{\text{LP}}(F_1, F_2) \to 0$, in part because $g_j(x,\beta)$ is bounded between 0 and 1. We do not have a counterexample to continuity when $\mathcal{B}$ is not a finite set. Note that our consistency result for a mixtures of parametric densities, presented in Section 5.2, has a continuity condition that is easier to verify for discontinuous $g_j(x,\beta)$, as Section 6 discusses.

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.

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 B.1 for the proof of the forthcoming Theorem 2, which nests Theorem 1.

Remark 4. Appendix A proves that this estimation problem is well posed under the definition of Chen (2007).

Next, we consider the distribution function estimator using the log-likelihood criterion. Define the population criterion function and its sample analog, respectively, as

$$Q(F) \equiv -Q^{\text{ML}}(F) \equiv -E\left[\sum_{j=1}^{J} y_{i,j} \log P_j(x_i, F)\right] = -E\left[\sum_{j=1}^{J} y_{i,j} \log \int g_j(x_i,\beta) dF(\beta)\right]$$

and

$$\hat{Q}_N(F) \equiv -\hat{Q}_N^{\text{ML}}(F) \equiv -\frac{1}{N} \sum\nolimits_{i=1}^{N} \sum_{j=1}^{J} y_{i,j} \log P_j(x_i, F) = -\frac{1}{N} \sum\nolimits_{i=1}^{N} \sum_{j=1}^{J} y_{i,j} \log \int g_j(x_i,\beta) dF(\beta)$$

for $F \in \mathcal{F}_{R(N)}$. Then we can obtain the ML estimator as in (6) and denote the resulting estimator by $\hat{F}_N^{\text{ML}}$. We obtain the following corollary for the consistency of this ML estimator

Corollary 1. *Suppose Assumption 1 and Condition 1 hold. Further suppose that $P_j(x, F)$ is bounded away from zero for all j and $F \in \mathcal{F}$. Then, $d_{\text{LP}}\left(\hat{F}_N^{\text{ML}}, F_0\right) \xrightarrow{p} 0$.*

See Appendix B.2 for the proof.

Remark 5. 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 heterogeneous parameters, we might be interested in inference in the mean or median of β. For 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 $d_{\text{LP}}(\Pi_N F_0, F_0)$, 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).[3] 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. For a subset of problems, including the random coefficients logit, we are able to derive approximation error bounds. We provide convergence rates for these cases in Section 7.

5 Extensions

5.1 Models with Homogenous Parameters

In many empirical applications, it is common to have both heterogeneous parameters β and finite-dimensional parameters $\gamma \in \Gamma \subseteq \mathbb{R}^{\dim(\gamma)}$. We write the model choice probabilities as $g(x, \beta, \gamma)$ and the aggregate choice probabilities as $P(x, F, \gamma)$. Here we consider the consistency of estimators for models with both homogenous parameters and heterogeneous parameters. For conciseness, we state a theorem for the least squares criterion and omit a corollary for the likelihood criterion.

Remark 6. Estimating a model allowing a parameter to be a heterogeneous parameter when in truth the parameter is homogeneous will not affect consistency if the model with heterogeneous parameters is identified.

Remark 7. Searching over γ as a homogeneous parameter for the least squares criterion requires nonlinear least squares. The optimization problem may also not be globally convex. The objective function may not be differentiable for our examples where $g(x, \beta, \gamma)$ involves an indicator function.

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

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

[3] 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.

where $\hat{Q}_N(\gamma, F)$ denotes the corresponding sample criterion function. $Q(\gamma, F)$ is the population criterion function based on the model $g(x, \beta, \gamma)$. An alternative computational strategy is profiling, as in

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

Profiling gives us

$$\hat{\gamma}_N = \text{argmin}_{\gamma \in \Gamma} \hat{Q}_N\left(\gamma, \hat{F}_N(\gamma)\right) + C \cdot \nu_N,$$

and therefore $\hat{F}_N = \hat{F}_N(\hat{\gamma}_N)$. We make the following assumptions to prove consistency for models with homogenous parameters.

Assumption 2.

1. *Let $\mathcal{A} \equiv \Gamma \times \mathcal{F}$ where $\mathcal{F}$ is compact in the weak topology and Γ is a compact subset of $\mathbb{R}^{\dim(\gamma)}$ and $\mathcal{A}$ contains* the true (γ_0, 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 $g(x, \beta, \gamma)$ is identified, meaning that for any $(\gamma_1, F_1) \neq (\gamma_0, F_0)$, $(\gamma_1, F_1) \in \Gamma \times \mathcal{F}$, the set $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ where $P(x, F_0, \gamma_0) \neq P(x, F_1, \gamma_1)$ has a positive measure in $\mathcal{X}$.*
5. *At least one of the following properties holds.*
   (a) *$g_j(x, \beta, \gamma)$ is Lipschitz continuous in γ for each outcome j.*
   (b) *(i) $\left(\hat{\gamma}_N, \hat{F}_N\right)$ is well-defined and measurable. (ii) For each outcome j, there exists a vector of known functions $h(x, \beta, \gamma) = (h_1(x, \beta, \gamma), \ldots, h_J(x, \beta, \gamma))$ such that, for each j, $g_j(x, \beta, \gamma) = 1[A_j \cdot h(x, \beta, \gamma) > 0]$ for some vector of known constants A_j. (iii) Each of the J functions $h_j(x, \beta, \gamma)$ is Lipschitz continuous in γ.*
6. *$Q(\gamma, F)$ is lower semicontinuous in γ, is continuous on $\mathcal{F}$ in the weak topology, and is continuous at (γ_0, F_0).*

If homogeneous parameters were added in the examples we consider, Assumption 2.5.a would hold for Example 1, the logit model with random coefficients. Assumption 2.5.b might hold for Examples 2 and 3. See the remark below.

We present the consistency theorem for the estimator with homogeneous parameters.

Theorem 2. *Suppose Assumption 2 and Condition 1 hold. Then, $\hat{\gamma}_N \xrightarrow{p} \gamma_0$ and $\hat{F}_N \xrightarrow{p} F_0$.*

See Appendix B.1 for the proof. Again, we omit a corollary for the likelihood case.

Remark 8. Assumption 2.5.b is designed to handle extensions of Examples 2 and 3 to include homogeneous parameters. In the extensions of these examples, the homogeneous parameters enter inside indicator functions. The non-primitive portion, Assumption 2.5.b.i, requires that the estimator $\left(\hat{\gamma}_N, \hat{F}_N\right)$ be well-defined and measurable, echoing a condition in Lemma A.2 of Chen and Pouzo (2012), which

our consistency proof relies on. Remark A.1.i.a in CP states that lower semicontinuity of the least squares sample objective function is a sufficient condition for the estimator to be well-defined and measurable. Even though an indicator function for an open set is lower semicontinuous, the least squares sample objective (or likelihood) function itself might not be lower semicontinuous in γ even if each $g_j(x, \beta, \gamma)$ is lower semicontinuous in γ. For example, multiplication by a negative number is enough to change lower semicontinuity into upper semicontinuity. Therefore, we follow the main text of CP and, for the case of indicator functions, maintain the non-primitive assumption that the estimator is well-defined and measurable. Assumption 2.5.b.i states in part that the sample objective function has a unique global optimum. Although Assumption 2.5.b.i may not actually hold if the homogeneous parameters enter inside indicator functions, the sample criterion is converging to a population criterion with a unique global optimum due to Assumptions 2.4 and 2.6. Extending the appendix lemma in Chen and Pouzo (2012) to the case where the sample objective function has a continuum of multiple global optima (just like in maximum score) is a minor extension that we do not pursue for space reasons. The continuity of the population criterion $Q(\gamma, F)$ with respect to F can be satisfied using the sufficient condition discussed in Remark 3 for the earlier Assumption 1.5. The lower semicontinuity with respect to γ may require further primitive conditions, like the conditions on the support of the explanatory variables in the results on binary choice in Ichimura and Thompson (1998, Theorem 1).

5.2 Continuous Distribution Function Estimator

A limitation of the discrete approximation estimator is that the CDF of the heterogeneous parameters will be a step function. In applied work, it is often attractive to have a smooth distribution of heterogeneous parameters. In this subsection, we describe one approach to obtain a continuous distribution or density function estimator that allows for unbounded supports. Instead of modeling the distribution of the heterogeneous parameters as a mixture of point masses, we instead model the density as a mixture of parametric densities, e.g. normal densities. Approximating a density or distribution function using a mixture of parametric densities or distributions is popular (e.g. Jacobs, Jordan, Nowlan, and Hinton 1991, Li and Barron 2000, McLachlan and Peel 2000, and Geweke and Keane 2007). Our estimator's advantage is its computational simplicity.

As a leading case, let a basis r be a normal distribution with mean the $K \times 1$ vector μ^r and standard deviation the $K \times 1$ vector σ^r. Let $\phi(\beta \mid \mu^r, \sigma^r)$ denote the joint normal density corresponding to the rth basis distribution. Under independent normal basis functions, the joint density for a given r is just the product of the marginals, or $\phi(\beta \mid \mu^r, \sigma^r) = \prod_{k=1}^{K} \phi(\beta_k \mid \mu_k^r, \sigma_k^r)$. We can also use only a location mixture with the basis functions $\phi(\beta \mid \mu^r) = \prod_{k=1}^{K} \phi(\beta_k \mid \mu_k^r)$ or use a multivariate normal mixture with the basis functions $\phi(\beta \mid \mu^r, \Sigma^r)$, where Σ^r denotes a variance-covariance matrix. We can also consider mixtures of other parametric density functions. We use the generic notation $\phi(\beta \mid \lambda^r)$ to denote the rth basis function, where λ^r is the rth distribution parameter. Let θ^r denote the probability weight given to the rth basis function, $\phi(\beta \mid \lambda^r)$. As in the discrete approximation estimator, the vector of weights θ lies in the unit simplex, Δ_R.

There are many ways to perform K-dimensional numerical integration, such as sparse grid quadrature (Heiss and Winschel 2008). We focus on simulation for simplicity. To implement our continuous density estimator for a given R, make S simulation draws from $\phi(\beta \mid \lambda^r)$ independently of r (i.e. use independent simulation draws for each λ^r). Let a particular draw s be denoted as $\beta^{r,s}$. We then create

the R "regressors"

$$z_{i,j}^r = \int g_j\left(x_i, \beta\right) \phi\left(\beta \mid \lambda^r\right) d\beta \approx \frac{1}{S} \sum_{s=1}^{S} g_j\left(x_i, \beta^{r,s}\right).$$

The $\approx$ emphasizes the error (possibly quite small) in numerical integration. This numerical integration step is done first, before any optimization. We then approximate $P_j\left(x_i\right)$ as

$$P_j\left(x_i\right) \approx \sum_{r=1}^{R} \theta^r z_{i,j}^r.$$

Here, we use the $\approx$ to emphasize both sieve and numerical integration approximations, although typically the sieve approximation will be a larger source of error. We estimate θ using the inequality constrained least squares problem as before

$$\widehat{\theta}_S = \arg\min_{\theta \in \Delta_R} \frac{1}{NJ} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{J} \left(y_{i,j} - \sum\nolimits_{r=1}^{R} \theta^r z_{i,j}^r\right)^2. \tag{7}$$

This is once again inequality-constrained linear least squares, a globally convex optimization problem with an easily-computed unique solution. The resulting density estimator is $\hat{f}_{N,S}\left(\beta\right) = \sum_{r=1}^{R} \hat{\theta}_S^r \phi\left(\beta \mid \lambda^r\right)$ and the distribution function estimator is $\hat{F}_{N,S}\left(\beta\right) = \sum_{r=1}^{R} \hat{\theta}_S^r \Phi\left(\beta \mid \lambda^r\right)$, where $d\Phi(\cdot) = \phi(\cdot)$.

5.2.1 Consistency of the Continuous Distribution Function Estimator

We show consistency for the continuous distribution function estimator after imposing additional restrictions on the data generating process. For this purpose, we restrict the class of the true distribution functions to

$$\mathcal{F}^M = \left\{F : F = \int \Phi(\beta \mid \lambda) P_\lambda(d\lambda),\ P_\lambda \in \mathcal{P}_\lambda,\ d\Phi(\beta \mid \lambda) \in G = \left\{d\Phi(\beta \mid \lambda) \mid \beta \in \mathcal{B} \subseteq \mathbb{R}^K, \lambda \in \Lambda \subset \mathbb{R}^d\right\}\right\}, \tag{8}$$

such that any distribution in $\mathcal{F}^M$ is in truth given by a mixture of parametric distributions in G. Note that we allow for $\mathcal{B} = \mathbb{R}^K$ in $\mathcal{F}^M$, thus removing the restriction that $\mathcal{B}$ is compact underlying Theorem 1. Here P_λ denotes a probability measure on Λ, the support of the distribution parameter λ. This means that we assume the true distribution is in the space of possibly continuous mixtures over some known parametric basis functions. Note, however, that Petersen (1983, as Lemma 3.4 of Zeevi and Meir (1997)) suggests that any density function can be approximated to arbitrary accuracy by an infinitely countable convex combination of basis densities, including normals, i.e. G can be dense in the space of continuous density functions. For example Zeevi and Meir (1997) show that G is dense in the space of all density functions that are bounded away from zero on compact support. Therefore we argue that the class $\mathcal{F}^M$ can be arbitrary close to the space of any continuous distribution functions with suitable choices of G. We approximate this possibly continuous mixture using a finite mixture over the same basis functions. Accordingly we construct our sieve space as

$$\mathcal{F}_R^M = \left\{F \mid F = \sum\nolimits_{r=1}^{R} \theta^r \Phi(\beta \mid \lambda^r),\ d\Phi(\beta \mid \lambda) \in G,\ \theta \in \Delta_R\right\}.$$

First consider an estimator ignoring numerical integration error in the regressors

$$\hat{F}_N(\beta) = \sum_{r=1}^{R(N)} \hat{\theta}^r \Phi(\beta|\lambda^r), \tag{9}$$

where $\widehat{\theta} = \arg\min_{\theta \in \Delta_{R(N)}} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R(N)} \theta^r z_{i,j}^r \right)^2$ and $z_{i,j}^r = \int g_j(x_i, \beta) d\Phi(\beta \mid \lambda^r)$ with a choice of grid $\Lambda_R = \left(\lambda^1, \ldots, \lambda^R\right)$ on Λ. We consider simulation error later.

Let $F_0 = \int \Phi(\beta \mid \lambda) P_{\lambda,0}(d\lambda) \in \mathcal{F}^M$. Then we can present the consistency result in terms of estimating the true $P_{\lambda,0}$ because knowing P_λ fully characterizes the true F_0 given a researcher's choice of $\Phi(\beta \mid \lambda)$ (e.g. normal distributions). Note that (9) can be also written in terms of estimating P_λ with the estimator

$$\hat{P}_{\lambda,N}(\lambda) = \sum_{r=1}^{R(N)} \hat{\theta}^r 1\left[\lambda^r \leq \lambda\right]$$

where the sieve space for P_λ is given by

$$\mathcal{P}_{\lambda,R} = \left\{ P_\lambda \mid P_\lambda(\lambda) = \sum_{r=1}^{R} \theta^r 1\left[\lambda^r \leq \lambda\right], \theta \in \Delta_R \right\}.$$

Therefore $\hat{F}_N = \int \Phi(\cdot|\lambda) d\hat{P}_{\lambda,N}(d\lambda) \xrightarrow{p} F_0 = \int \Phi(\cdot|\lambda) dP_{\lambda,0}(d\lambda)$ as long as $\hat{P}_{\lambda,N} \xrightarrow{p} P_{\lambda,0}$ because we assume $F_0 \in \mathcal{F}^M$ and because $\Phi(\beta|\lambda)$ is continuous in λ for all β and therefore $F = \int \Phi(\cdot|\lambda) dP_\lambda(d\lambda)$ is also continuous on $\mathcal{P}_\lambda$ in the Lévy-Prokhorov metric. This facilitates our analysis because we can directly apply Theorem 1 to $\hat{P}_{\lambda,N}$ under assumptions below. To present the consistency theorem we need additional notation.

Define a sample criterion function in terms of estimating $P_{\lambda,0}$ as

$$\hat{Q}_N(P_\lambda) \equiv \frac{1}{NJ} \sum_{i=1}^{N} \left\| y_i - \int \tilde{g}(x_i, \lambda) dP_\lambda(\lambda) \right\|_E^2 = \frac{1}{NJ} \sum_{i=1}^{N} \left\| y_i - \sum_{r=1}^{R} \theta^r \tilde{g}(x_i, \lambda^r) \right\|_E^2 \tag{10}$$

for $P_\lambda \in \mathcal{P}_{\lambda,R(N)}$, where $\tilde{g}(x, \lambda^r) = \int g(x, \beta) d\Phi(\beta|\lambda^r)$. Then we can rewrite the estimator $\hat{P}_{\lambda,N}$ as

$$\hat{P}_{\lambda,N} = \text{argmin}_{P_\lambda \in \mathcal{P}_{\lambda,R(N)}} \hat{Q}_N(P_\lambda) + C \cdot \nu_n \tag{11}$$

for some positive sequence ν_n tending to zero. Also define the population criterion function as

$$Q(P_\lambda) \equiv E\left[\left\| y - \int \tilde{g}(x, \lambda) dP_\lambda(\lambda) \right\|_E^2 / J \right].$$

We make the following assumptions, which are similar to those used in Theorem 1

Assumption 3.

1. *Let $\mathcal{F}^M$ be a space of distribution functions generated by P_λ in (8), where Λ is compact. $\mathcal{F}^M$ contains* the true F_0.
2. *Let $((y_i, x_i))_{i=1}^N$ be i.i.d.*
3. *Let both β and λ be independently distributed from x.*

4. *Assume the model $g(x,\beta)$ is identified, meaning that for any $P_{\lambda,1} \neq P_{\lambda,0}$, the set $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ where $P(x,F_0) \neq P(x,F_1)$ such that $F_0 = \int \Phi(\cdot \mid \lambda) P_{\lambda,0}(d\lambda)$ and $F_1 = \int \Phi(\cdot \mid \lambda) P_{\lambda,1}(d\lambda)$ has a positive measure in $\mathcal{X}$.*

5. *$Q(P_\lambda)$ is continuous on $\mathcal{P}_\lambda$ in the weak topology.*

We leave out lengthy discussion of these assumptions because most of these assumptions are made for similar reasons as those in Assumption 1. More discussion is in the beginning of Section 6. We also require that the choice of grids on Λ satisfies the following properties.

Condition 2. 1. Let $\Lambda_R = \left(\lambda^1, \ldots, \lambda^R\right)$ become dense in Λ as $R \to \infty$.

2. $\mathcal{P}_{\lambda,R} \subseteq \mathcal{P}_{\lambda,R+1} \subseteq \mathcal{P}_\lambda$ 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 $N \to \infty$.

Theorem 3. *Suppose Assumption 3 and Condition 2 hold. Then, $d_{\mathrm{LP}}\left(\hat{P}_{\lambda,N}, P_{\lambda,0}\right) \xrightarrow{p} 0$ and $d_{\mathrm{LP}}\left(\hat{F}_N, F_0\right) \xrightarrow{p} 0$.*

A brief proof is in an appendix. However, most of the steps mirror the proofs of Theorems 1 and 2, and so the appendix omits the unchanged steps for conciseness.

Next we account for the simulation error in the basis functions from approximating the integral with respect to $\Phi(\beta \mid \lambda^r)$. Denote the resulting distribution estimator with simulated basis functions by $\hat{F}_{N,S}(\beta) \equiv \sum_{r=1}^{R(N)} \hat{\theta}_S^r \Phi(\beta \mid \lambda^r)$, where

$$\widehat{\theta}_S = \arg\min_{\theta \in \Delta_{R(N)}} \frac{1}{N} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R(N)} \theta^r \frac{1}{S} \sum_{s=1}^{S} g(x_i, \beta^{r,s}) \right)^2. \tag{12}$$

Note that all simulation draws are independent and that different draws are used for each r. We obtain the mixing ratios $\hat{\theta}_S$ as in (12) using the simulated basis functions and our distribution function estimator is still $\widehat{F}_{N,S}(\beta)$, which belongs to $\mathcal{F}_R^M$. Note that we only use simulation to approximate $\Phi(\beta \mid \lambda^r)$ and to obtain $\hat{\theta}_S$ in (12). Our distribution estimator is still $\widehat{F}_{N,S}(\beta)$, not $\tilde{F}_{N,S}(\beta) \equiv \sum_{r=1}^{R(N)} \hat{\theta}_S^r \frac{1}{S} \sum_{s=1}^{S} 1[\beta^{r,s} \leq \beta]$, because the CDF $\Phi(\beta \mid \lambda^r)$ is known to researchers. In the normal distribution case, specialized software calculates the normal CDF $\Phi(\beta \mid \lambda^r)$ with much less error than typical simulation approaches. We show our estimator is consistent when the number of simulation draws tends to infinity.

Theorem 4. *Suppose Assumption 3 and Condition 2 hold. Let $S \to \infty$. Then, $d_{\mathrm{LP}}\left(\hat{F}_{N,S}, F_0\right) \xrightarrow{p} 0$.*

The proof is in the appendix.

6 Discussion of Examples

We return to the examples we introduced in Section 2. We discuss the two key conditions for each model $g(x,\beta)$: 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.

For the mixture of continuous densities, the identification of the mixture distribution, Assumption 3.4, will typically occur if the underlying distribution of heterogeneous parameters is identified, Assumption 1.4, and the basis functions are chosen appropriately. We do not need to explicitly verify Assumption 3.4 once Assumption 1.4 is verified. We also do not explicitly verify Assumption 3.5, continuity of the population objective function in the continuous mixtures cases. An important point is that even when $g(x,\beta)$ is nonsmooth as in Examples 2 and 3, $\tilde{g}(x,\lambda) = \int g(x,\beta)d\Phi(\beta \mid \lambda)$ can still be continuous and differentiable in λ because it is smoothed by integration with respect to the distribution $\Phi(\beta \mid \lambda)$. Therefore, Assumption 3.5 can be satisfied by Remark 2.

Example. 1 (logit) The identification of $F(\beta)$ in the random coefficients logit model is the main content of Fox, Kim, Ryan, and Bajari (2012, Theorem 15).[4] Assumption 14 in Fox 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 $\left(x_2', \ldots, x_{j-1}', x_j', x_{j+1}', \ldots, x_J'\right) = \left(0', \ldots, 0', x_j', 0', \ldots, 0'\right)$."[5] Fox 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 2 in the current paper.

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 explanatory variables in x is known to either always be positive or either always be negative (the sign can be identified), ii) there is some normalization such as the coefficient known to be positive or negative is always either $+1$ or -1 (more generally the random coefficients lie on an unknown hemisphere), iii) there are large and product supports on each of the explanatory variables other than the intercept. This rules out polynomial terms and interactions. If we impose the scale normalization $\beta_{k^*} = \pm 1$, Assumption 1.4 holds if we add large and product support conditions on each explanatory variable in x. 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. Next, note that the continuity condition (Assumption 1.5) holds by Remark 3 when the support $\mathcal{B}$ is a finite set. We have not shown that continuity holds under only the identification assumptions of Ichimura and Thompson (1998, Theorem 1), although we know of no counterexamples.

Example. 3 (multinomial choice without logit errors) Fox and Gandhi (2015, Theorem 2) 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 explanatory variable with large and product support. On other hand, Fox and Gandhi allow

[4] Theorem 15 of Fox et al also allows homogeneous, product-specific intercepts.

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

polynomial terms and interactions for x's other than the choice-j-specific large support explanatory variables, unlike examples 1 and 2. They do not require large support for the x's that are not the large support, choice-specific explanatory variables. 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 $\mathcal{F}$. Next, the continuity Assumption 1.5 holds also by Remark 3 when the support $\mathcal{B}$ is a finite set.

7 Estimation Error Bounds

We next derive convergence rates for the approximation of the underlying distribution of heterogeneous parameters $F(\beta)$ for a subset of the models allowed in the consistency theorems in Sections 3 and 5.2. These results highlight, among other issues, the curse of dimensionality in the dimension of the heterogeneous parameters. Larger sample sizes are needed if the dimension K of the heterogeneous parameters is larger. We present results for the least squares criterion and do not include corollaries for the likelihood criterion.

7.1 Discrete Approximation Estimator

Our results apply to a narrower set of true models $g_j(x, \beta)$ than the earlier consistency theorems. In particular, we require that $g_j(x, \beta)$ be smooth, so we allow our Example 1, the logit case, but not the other examples, where $g_j(x, \beta)$ involves an indicator function. Nevertheless, the random coefficients logit is a leading model used in empirical work and the results exploiting the smoothness of $g_j(x, \beta)$ are illustrative of issues, such as the curse of dimensionality, that apply also to nonsmooth choices for $g_j(x, \beta)$.

Our results relate to the literature on sieve estimations (e.g., Newey (1997) and Chen (2007)). Of course, the major difference is our choice of discrete basis functions, motivated by our estimator's computational advantages and our desire to easily constrain our estimate to be a valid CDF. We cannot directly apply previous results because of our different sieve space. Our proof technique to derive our error bounds uses results on quadrature, as we will explain.

We restrict the true distribution F_0 to lie in a class of distributions that have smooth densities on a compact space of heterogeneous parameters. We define the parameter space for the choice probabilities $P(x, F)$ as the collection of those generated by such smooth densities:

$$\mathcal{H} = \left\{ P(x, F) \mid \max_{0 \leq s \leq \overline{s}} \sup_{\beta \in \mathcal{B}} |D^s f(\beta)| \leq \bar{C}, \mathcal{B} = \prod_{k=1}^{K} \left[\underline{\beta}_k, \overline{\beta}_k \right], f = dF \in \mathcal{C}^{\overline{s}}[\mathcal{B}] \right\}, \tag{13}$$

where $D^s = \frac{\partial^s}{\partial \beta_1^{\alpha_1} \ldots \partial \beta_K^{\alpha_K}}$, $s = \alpha_1 + \cdots + \alpha_K$ with $D^0 f = f$, giving the collection of all derivatives of order s. Also, $\mathcal{C}^{\overline{s}}[\mathcal{B}]$ is a space of $\overline{s}$-times continuously differentiable density functions defined on $\mathcal{B}$. Therefore we assume any element of the class of density functions that generates $\mathcal{H}$ is defined on a Cartesian product $\mathcal{B}$, is uniformly bounded by $\bar{C} < \infty$, is $\overline{s}$-times continuously differentiable, and has

all own and partial derivatives uniformly bounded by $\bar{C}$. The definition of $\mathcal{H}$ depends on $\bar{C}$ and $\bar{s}$; the degree of smoothness $\bar{s}$ will show up in our convergence rate results.

Let the space of approximating functions be

$$\mathcal{H}_{R(N)} = \left\{ P(x,F) \mid F \in \mathcal{F}_{R(N)} \right\}. \tag{14}$$

We require that the grid points accumulate in $\mathcal{F}_{R(N)}$ such that $\mathcal{H}_R \subseteq \mathcal{H}_{R+1} \subseteq \ldots$. Our approach uses results from quadrature to pick approximation choices θ^r such that we can approximate the choice probability $P(x, F_0)$ arbitrarily well using approximating functions in $\mathcal{H}_{R(N)}$ as $R(N) \to \infty$.

In this section and in the corresponding proofs, we let $\|v\|_E = \sqrt{v'v}$, $\|h\|_{L_{2,N}}^2 = \frac{1}{N}\sum_{i=1}^N ||h(x_i)||_E^2$, $\|h\|_{L_2}^2 = \int ||h||_E^2 d\varpi$ (the norm in L_2), and $\|h\|_\infty^2 = \sup_{x\in\mathcal{X}} ||h(x)||_E^2$ for any function $h : \mathcal{X} \to \mathbb{R}$, where ϖ denotes a probability measure on $\mathcal{X}$. We introduce the linear probability model error $e_{i,j}$, as in $y_{i,j} = P_j(x_i, F) + e_{i,j}$ and $E[e_{i,j} \mid X_1, \ldots, X_N] = 0$. In addition to restricting the class of true distributions, we make the following additional assumptions.

Assumption 4. *(i)* $\left(e_i = (e_{i,1}, \ldots, e_{i,J})'\right)_{i=1}^N$ *are independently distributed; (ii)* $E[e_i \mid X_1, \ldots, X_N] = 0$*; (iii)* $(X_i)_{i=1}^N$ *are i.i.d. with a density function bounded above;* (iv) $g(x,\beta)$ is $\bar{s}$*-times continuously differentiable w.r.t.* β *and its (all own and partial) derivatives are uniformly bounded*; (v) *the sieve space defined in (*14*) satisfies* $\mathcal{H}_R \subseteq \mathcal{H}_{R+1} \subset \ldots$.

Assumptions 4 (i)-(iii) are about the structure of the data and in particular they allow for heteroskedasticity for the linear probability error, which is necessary for linear probability models. As previewed earlier, Assumption 4 (iv) assumes that the true model $g(x,\beta)$ is differentiable. We use Assumption 4 (iv) to approximate $P(x,F)$ using a sieve method for F combined with a quadrature method for the choice of weights θ^r. Assumption 4 (iv) is satisfied by Example 1. Assumption 4 (v) was mentioned previously.

Any asymptotic error bound consists of two terms: the order of bias and the variance. While obtaining the variance term is rather standard in the sieve estimation literature, deriving the bias term depends on the specific choice of basis function (e.g., power series or splines in the previous literature). Because our choice of basis functions is new in the sieve literature, we first state the order of bias, meaning the approximation error rate of our sieve approximation to arbitrary conditional choice probabilities $P(x,F)$ in $\mathcal{H}$. Keep in mind this bias result is primarily about the flexibility of a class of approximations and has less to do with using a finite sample of data.

Lemma 1. *Suppose* $P(x,F) \in \mathcal{H}$ *and suppose Assumptions 4 (iv) and (v) hold. Then there exist* $F^* \in \mathcal{F}_{R(N)}$ *such that for all* $x \in \mathcal{X}$, $\|P(x,F^*) - P(x,F)\|_E^2 = O\left(R^{-2\bar{s}/K}\right)$. *Further suppose Assumptions 4 (i)-(iii) hold. Then,* $\|P(x_i,F^*) - P(x_i,F)\|_{L_{2,N}}^2 = O_{\mathbb{P}}\left(R^{-2\bar{s}/K}\right)$.

To prove this result, we combine a quadrature approximation with a power series approximation to approximate $P(x,F) = \int g(x,\beta) f(\beta) d\beta$. We first approximate $g(x,\beta) f(\beta)$ using a tensor product power series in β. Then we approximate the integral of the tensor products approximation with respect to β using quadrature. The complete proof is in Appendix C.3.

Lemma 1 is the key ingredient that allows us to use machinery from the sieve literature for our estimator. Let C be a (generic) positive constant. Define $\Psi_R \equiv \left(E\left[\sum_j g_j(X_i,\beta^r) g_j\left(X_i,\beta^{r'}\right)\right]\right)_{1\le r,r'\le R}$

(a $R \times R$ matrix) and its smallest eigenvalue as $\xi_{\min}(R)$. Then we obtain the following estimation error bounds.

Theorem 5. *Suppose Assumptions 1.1, 1.3, and 1.4 and Condition 1 hold. Suppose Assumption 4 holds. Suppose further that* $\frac{R(N)^2 \log R(N)}{N} \to 0$ *and* $\xi_{\min}(R) > 0$ *for all finite* R*. Then if* $P(x, F_0) \in \mathcal{H}$ *and* C *is a particular constant,*

$$\left\| P\left(x_i, \hat{F}_N\right) - P(x_i, F_0) \right\|_{L_{2,N}}^2 \quad \leq \quad C \cdot \max\left\{ \frac{R(N) \log R(N)}{\xi_{\min}(R(N)) \cdot N}, R(N)^{-2\overline{s}/K} \right\} \ \textit{w.p.a.1.}$$

Theorem 5 establishes a bound on the distance between the true conditional choice probability and the approximated conditional choice probability. It shows how the finite-sample estimator is able to fit data generated by a nonparametric choice of a heterogeneity distribution. Roughly speaking, in the estimation error bound the first term in the max operator corresponds to an asymptotic variance and the second term corresponds to an asymptotic bias (Chen 2007). Fixing N, a larger $R(N)$ reduces the bias but increases the variance and an optimal choice of $R(N)$ is obtained by balancing the bias and variance. Obtaining the variance term is standard in the sieve estimation literature and obtaining the bias term requires approximation results that depend on the type of sieves (in the previous literature, e.g., power series or splines). In the proof, we obtain this bias term using Lemma 1.

There are several lessons from this error bound. The approximation error rate in the bias term shows that we have faster convergence with a smoother density function $\overline{s}$ and slower convergence with a higher dimensional β, K. These results are intuitive. In many other settings, smoother distributions are easier to approximate and higher dimensional distributions are harder to approximate. Practically, one might want to use caution when estimating an unrestricted joint distribution of β when K is large.

Note that the condition $\xi_{\min}(R) > 0$ for all finite R in Theorem 5 means the regressors $g_j(X_i, \beta^r)$ are not linearly dependent across different β^r in least squares estimation. Theorem 5 allows the case $\lim_{R\to\infty} \xi_{\min}(R) = 0$ but this term drops out in the convergence rate if $\lim_{R\to\infty} \xi_{\min}(R) > 0$. The case $\lim_{R\to\infty} \xi_{\min}(R) = 0$ means that Ψ_R, which is the probability limit (as $N \to \infty$ with fixed R) of the sample matrix $\left(\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^J g_j(x_i, \beta^r) g_j(x_i, \beta^{r'}) \right)_{1 \leq r, r' \leq R}$, tends to be singular as R grows to infinity. In this case, as often proposed in the literature (e.g. Hastie, Tibshirani, and Friedman 2009), we can potentially reduce the variances of resulting estimators using a subset selection method or a shrinkage method, such as LASSO and ridge/Tikhonov regression. On the other hand, these approaches can cause additional bias in finite samples.

The previous theorem was for choice probabilities, but we are also interested in the distribution function itself. Using the result in Theorem 5, we can approximate the distribution of heterogeneous parameters with the following convergence rate.

Theorem 6. *Suppose the assumptions and conditions in Theorem 5 hold. Then for a particular constant* C*, w.p.a.1*

$$\left| \widehat{F}_N(\beta) - F_0(\beta) \right| \leq C \sqrt{\frac{R(N)}{\xi_{\min}(R(N))}} \max\left\{ \sqrt{\frac{R(N) \log R(N)}{\xi_{\min}(R(N)) \cdot N}}, R(N)^{-\overline{s}/K} \right\} \ \textit{a.e. } \beta \in \mathcal{B}.$$

Not surprisingly, the bias term in this bound suggests that the estimator of the distribution function

suffers from a curse of dimensionality in the number of heterogeneous parameters K and has a faster rate of convergence if the true generating process is smoother, as indexed by $\bar{s}$.

7.2 Continuous Distribution Function with Compact Support

Here we consider the asymptotics for the smooth density estimator proposed in Section 5.2. In this approach, we approximate the vector of choice probabilities $P(x_i) = (P_1(x_i), \ldots, P_J(x_i))$ using smooth basis functions as

$$P(x_i) \approx \sum_{r=1}^{R} \theta^r \int g(x_i, \beta) d\Phi(\beta \mid \lambda^r),$$

where $\Phi(\beta \mid \lambda^r)$ denotes the rth basis distribution function. We focus on simulation as the numerical integration technique. Therefore,

$$P(x_i) \approx \sum_{r=1}^{R} \theta^r \int g(x_i, \beta) d\Phi(\beta \mid \lambda^r) \approx \sum_{r=1}^{R} \theta^r \left(\frac{1}{S} \sum_{s=1}^{S} g(x_i, \beta^{r,s}) \right), \tag{15}$$

where the $(\beta^{r,s})_{s=1}^{S}$ are drawn from $\Phi(\beta \mid \lambda^r)$. We then can rewrite (15) as

$$P(x_i) \approx \sum_{r=1}^{R} \theta^r \left(\frac{1}{S} \sum_{s=1}^{S} g(x_i, \beta^{r,s}) \right) = \sum_{r=1}^{R} \sum_{s=1}^{S} \frac{\theta^r}{S} g(x_i, \beta^{r,s}) = \sum_{\tilde{r}=1}^{R \times S} \tilde{\theta}^{\tilde{r}} g(x_i, \beta^{\tilde{r}})$$

for some $\beta^{\tilde{r}}$ and $\tilde{\theta}^{\tilde{r}}$. Therefore, we can interpret (15) as the discrete approximation of $P(x_i)$ with $R \cdot S$ grid points and $R \cdot (S-1)$ restrictions such that the first group of S coefficients on the first group of S regressors are identical to $\frac{\theta^1}{S}$, the second group of S coefficients on the second group of S regressors are identical to $\frac{\theta^2}{S}$, and so on.

Therefore, the smooth mixture model can be nested in the discrete approximation case if the support of β is bounded. To see this note that we have $\hat{F}_{N,S}(\beta) \equiv \sum_{r=1}^{R(N)} \hat{\theta}_S^r \Phi(\beta \mid \lambda^r)$ and the simulation-approximated distribution estimator by $\tilde{F}_{N,S}(\beta) \equiv \sum_{r=1}^{R(N)} \hat{\theta}_S^r \frac{1}{S} \sum_{s=1}^{S} 1[\beta^{r,s} \leq \beta]$ where $\widehat{\theta}_S$ solves (12). Then we can obtain

$$\begin{aligned} \left| \hat{F}_{N,S}(\beta) - F_0(\beta) \right| &\leq \left| \hat{F}_{N,S}(\beta) - \tilde{F}_{N,S}(\beta) \right| + \left| \tilde{F}_{N,S}(\beta) - F_0(\beta) \right| \\ &\leq \sum_{r=1}^{R(N)} \hat{\theta}_S^r \left| \frac{1}{S} \sum_{s=1}^{S} 1[\beta^{r,s} \leq \beta] - \Phi(\beta \mid \lambda^r) \right| + \left| \tilde{F}_{N,S}(\beta) - F_0(\beta) \right| \xrightarrow{p} 0 \end{aligned} \tag{16}$$

as $N \to \infty$ and $S \to \infty$. The first term in (16) goes to zero because for any given r the empirical distribution obtained from the simulation draws converges to the CDF of the draws (by the Glivenko–Cantelli theorem) and the second term in (16) goes to zero because $\tilde{F}_{N,S}(\beta)$ can be seen exactly as our discrete approximation estimator. We can also bound the convergence rate of the first term in (16) by $R(N)/\sqrt{S}$. Note that the term $\left| \frac{1}{\sqrt{S}} \sum_{s=1}^{S} \{1[\beta^{r,s} \leq \beta] - \Phi(\beta \mid \lambda^r)\} \right| = O_{\mathbb{P}^*}(1)$ for each given λ^r by a central limit theorem because $\beta^{r,s}$ are iid draws from $\Phi(\beta|\lambda^r)$ and $E^*[1[\beta^{r,s} \leq \beta]] = \Phi(\beta|\lambda^r)$ for each r, where $E^*[\cdot]$ denotes expectation and $\mathbb{P}^*$ denotes probability measure with respect to the simulation

draw. It follows that $\sum_{r=1}^{R(N)} \hat{\theta}_S^r \frac{1}{\sqrt{S}} \left| \frac{1}{\sqrt{S}} \sum_{s=1}^{S} \{1[\beta^{r,s} \leq \beta] - \Phi(\beta \mid \lambda^r)\} \right| \leq C \cdot R(N)/\sqrt{S}$ w.p.a.1, which is the bound for the first term in (16). As discussed above, Theorem 6 gives the convergence rate of the second term in (16), again because $\tilde{F}_{N,S}(\beta)$ can be seen exactly as our discrete approximation estimator.

Of course, the approach of nesting the mixture of continuous densities into the discrete approximation does not handle the full set of models allowed by the consistency theorem, Theorem 3. In particular, mixtures of normals are not allowed. Future work could explore more general results for the rate of convergence of the estimator using a mixture of continuous densities.

8 Conclusion

Previous papers have introduced fixed grid estimators for distributions of heterogeneity in structural models. We explore the asymptotic properties of the nonparametric distribution estimators. We show consistency in the function space of all distributions under the weak topology by viewing our estimators as sieve estimators and verifying high-level conditions in Chen and Pouzo (2012). We also show consistency for i) a model with homogenous parameters and ii) an estimator based on a mixture over smooth basis distribution functions on possibly unbounded supports. We verify some of the conditions for consistency for three example discrete choice models, each of which is widely used in empirical work. We also derive the convergence rates of the least squares nonparametric estimator for a class of differentiable models, for which we can derive the approximation error rates of our nonparametric sieve space.

A 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 $d_{\text{LP}}(F_n, F_0) \nrightarrow 0$.[6] We now argue that our problem is well-posed.

Note that $\mathcal{F}$ is compact in the weak topology (Assumption 1.1). 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

$$\inf_{F \in \mathcal{F}_{R(N)}: d_{\text{LP}}(F,F_0) \geq \epsilon} (Q(F) - Q(F_0)) \geq \inf_{F \in \mathcal{F}: d_{\text{LP}}(F,F_0) \geq \epsilon} (Q(F) - Q(F_0)) > 0, \tag{17}$$

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.

B Proofs of Consistency

B.1 Proofs of Theorems 1 and 2

We provide proof of the consistency theorem for models with homogenous parameters because the essentially same proof can be applied to the models without homogeneous parameters.

We verify the conditions of CP's Lemma A.2 (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 A.2, which does not incorporate a penalty function. We let $\alpha = (\gamma, F) \in \mathcal{A} \equiv \Gamma \times \mathcal{F}$, $\mathcal{A}_{R(N)} \equiv \Gamma \times \mathcal{F}_{R(N)}$, and with possible abuse of notation $d_{\text{LP}}(\alpha_1, \alpha_2) \equiv \|\gamma_1 - \gamma_2\|_E + d_{\text{LP}}(F_1, F_2)$ for $\alpha_1, \alpha_2 \in \mathcal{A}$. Define $\hat{Q}_N(\alpha) = \frac{1}{NJ} \sum_{i=1}^{N} \left\| y_i - \int g(x_i, \beta, \gamma)\, dF \right\|_E^2$ and $Q(\alpha) \equiv E\left[\left\| y - \int g(x, \beta, \gamma)\, dF \right\|_E^2 / J \right]$.

Lemma 2. ***Lemma A.2 of CP:*** *Let $\hat{\alpha}_N = (\hat{\gamma}, \hat{F}_N)$ be such that $\hat{Q}_N(\hat{\alpha}_N) \leq \inf_{\alpha \in \mathcal{A}_{R(N)}} \hat{Q}_N(\alpha) + O_p(\nu_N)$ with $\nu_N \to 0$. Suppose the following conditions (B.2.1)–(B.2.4) hold:*

- (B.2.1) (i) $Q(\alpha_0) < \infty$; (ii) there is a positive function $\delta(N, R(N), \varepsilon)$ such that for each $N \geq 1$, $R \geq 1$, and $\varepsilon > 0$, $\inf_{\alpha \in \mathcal{A}_{R(N)}: d_{LP}(\alpha, \alpha_0) \geq \varepsilon} Q(\alpha) - Q(\alpha_0) \geq \delta(N, R(N), \varepsilon)$ and $\liminf_{N \to \infty} \delta(N, R(N), \varepsilon) \geq 0$ for all $\varepsilon > 0$.
- (B.2.2) (i) $(\mathcal{A}, d_{\text{LP}}(\cdot))$ is a metric space; (ii) $\mathcal{A}_R \subseteq \mathcal{A}_{R+1} \subseteq \mathcal{A}$ for all $R \geq 1$, and there exists a sequence $\Pi_N \alpha_0 \in \alpha_{R(N)}$ such that $d_{\text{LP}}(\Pi_N \alpha_0, \alpha_0) \to 0$.

[6] 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.

- (B.2.3) (i) $\hat{Q}_N(\alpha)$ is a measurable function of the data $((y_i, x_i))_{i=1}^N$ for all $\alpha \in \mathcal{A}_{R(N)}$; (ii) $\hat{\alpha}_N$ is well-defined and measurable with respect to the Borel σ-field generated by the topology induced by the metric $d_{\text{LP}}(\cdot, \cdot)$.
- (B.2.4) (i) Let $\hat{c}^Q(R(N)) = \sup_{\alpha \in \mathcal{A}_{R(N)}} \left| \hat{Q}_N(\alpha) - Q(\alpha) \right| \xrightarrow{p} 0$;
  (ii) $\max \left\{ \hat{c}^Q(R(N)), \nu_N, |Q(\Pi_N \alpha_0) - Q(\alpha_0)| \right\} / \delta(N, R(N), \varepsilon) \xrightarrow{p} 0$ for all $\varepsilon > 0$.

Then $d_{\text{LP}}(\hat{\alpha}_N, \alpha_0) \xrightarrow{p} 0$.

Using Lemma 2 we now provide our consistency proof for the least squares estimator. Because our estimator is an extremum estimator, we can take ν_N to be arbitrarily small. We start with the condition (B.2.1). The condition $Q(\alpha_0) < \infty$ holds because $Q(\alpha) \leq 1$ for all $\alpha \in \mathcal{A}$, because we have a linear probability model. Next we will verify the condition

$$\inf_{a \in \mathcal{A}_{R(N)} : d_{\text{LP}}(\alpha, \alpha_0) \geq \varepsilon} Q(\alpha) - Q(\alpha_0) \quad \geq \quad \delta(N, R(N), \varepsilon) > 0 \tag{18}$$

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 $m(x, \alpha) = P(x, \alpha_0) - P(x, \alpha)$, where $P(x, \alpha) = \int g_j(x, \beta, \gamma) \, dF(\beta)$. Note that we have

$$Q(\alpha) \quad = \quad E\left[||y - P(x, \alpha_0) + m(x, \alpha)||_E^2 / J \right] = E\left[||y - P(x, \alpha_0)||_E^2 / J \right] + E\left[||m(x, \alpha)||_E^2 / J \right] \tag{19}$$

because $E[(y - P(x, \alpha_0))' m(x, \alpha)] = 0$ by the law of iterated expectation and $E[y - P(x, \alpha_0) \mid x] = 0$. Therefore, for each $\alpha \in \mathcal{A}$, we have

$$Q(\alpha) - Q(\alpha_0) \quad = \quad E\left[||m(x, \alpha)||_E^2 / J \right] - E\left[||m(x, \alpha_0)||_E^2 / J \right] = E\left[||m(x, \alpha)||_E^2 / J \right] \tag{20}$$

because $m(x, \alpha_0) = 0$. The condition (18) now holds due to our assumption of identification, as the following argument shows.

Consider $E[||m(x, \alpha)||_E^2]$, with $m(x, \alpha)$ defined above, as a map from $\mathcal{A}$ to $\mathbb{R}^+ \cup \{0\}$. For any $\alpha \neq \alpha_0$, $E[||m(x, \alpha)||_E^2]$ takes on positive values for each $\alpha \in \mathcal{A}$, because the model is identified on a set $\tilde{\mathcal{X}}$ with positive probability. Then note that $E[||m(x, \alpha)||_E^2]$ is continuous in α and also note that $\mathcal{A}_{R(N)}$ is compact. Therefore $E[||m(x, \alpha)||_E^2]$ attains some strictly positive minimum on $\{\alpha \in \mathcal{A}_{R(N)} : d_{\text{LP}}(\alpha, \alpha_0) \geq \varepsilon\}$. Then we can take $\delta(N, R(N), \varepsilon) = \inf_{\alpha \in \mathcal{A}_{R(N)} : d_{\text{LP}}(\alpha, \alpha_0) \geq \varepsilon} E\left[||m(x, \alpha)||_E^2 / J \right] > 0$ for all $R(N) \geq 1$ with $\varepsilon > 0$.

We have shown that $\delta(N, R(N), \varepsilon) > 0$ for all N and hence $\liminf_{N \to \infty} \delta(N, R(N), \varepsilon) \geq 0$. This is enough for (B.2.1). However, under our assumptions indeed $\liminf_{N \to \infty} \delta(N, R(N), \varepsilon) > 0$ because

$$\begin{aligned} \delta(N, R(N), \varepsilon) \quad &= \quad \inf_{\alpha \in \mathcal{A}_{R(N)} : d_{\text{LP}}(\alpha, \alpha_0) \geq \varepsilon} (Q(\alpha) - Q(\alpha_0)) \geq \inf_{\alpha \in \mathcal{A} : d_{\text{LP}}(\alpha, \alpha_0) \geq \varepsilon} (Q(\alpha) - Q(\alpha_0)) \\ &= \quad \inf_{\alpha \in \mathcal{A} : d_{\text{LP}}(\alpha, \alpha_0) \geq \varepsilon} E[||m(x, \alpha)||_E^2 / J] > 0, \end{aligned}$$

where the first inequality holds because $\mathcal{A}_{R(N)} \subseteq \mathcal{A}$ by construction and the second, strict inequality holds because the model is identified (Assumption 1.4). Here, $\liminf_{N \to \infty} \delta(N, R(N), \varepsilon) > 0$ holds because the last term in the right-hand side of the above inequality does not depend on N and the

term is strictly positive.[7] This result makes it convenient to verify (B.2.4)(ii) below.

Next we consider (B.2.2). First note that $(\mathcal{A}, d_{\text{LP}})$ is a metric space and we have $\mathcal{A}_R \subseteq \mathcal{A}_{R+1} \subseteq \mathcal{A}$ for all $R \geq 1$ by construction of our sieve space. Then we claim that there exists a sequence of functions $\Pi_N \alpha_0 \in \mathcal{A}_{R(N)}$ such that $d_{\text{LP}}(\Pi_N \alpha_0, \alpha_0) \to 0$ as $N \to \infty$. First, $\mathcal{B}_{R(N)}$ becomes dense in $\mathcal{B}$ by assumption. Second, $\mathcal{A}_{R(N)}$ becomes dense in $\mathcal{A}$ because the set of distributions $\mathcal{F}_{R(N)}$ 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 we consider (B.2.3). As Theorem 1 is a special case of Theorem 2, we focus explicitly on Theorem 2. Consider first Assumption 2.5.a. Remark A.1.(1) (a) of CP says that (B.2.3) holds if each sieve space $\mathcal{A}_R$ is compact and the finite-sample objective function is lower semicontinuous in (γ, F). 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 $\mathcal{B}$ and hence $\mathcal{A}_R$ is also a compact subset of $\mathcal{A}$.[8] Second we show that for any data $((y_i, x_i))_{i=1}^N$, $\hat{Q}_N(\alpha)$ is continuous on $\mathcal{A}_R$ for each $R \geq 1$. Since $\mathcal{A}_R$ is compact, this continuity means our estimator is well defined as the minimum in (6). Because checking the continuity in γ is trivial for given models (e.g. logit model), we focus on the continuity on $\mathcal{F}_R$. For any $F_1, F_2 \in \mathcal{F}_{R(N)}$, applying the triangle inequality, we obtain

$$\begin{aligned}
\left|\hat{Q}_N(\gamma, F_1) - \hat{Q}_N(\gamma, F_2)\right| &\leq 2\sum_{i=1}^N \sum_{j=1}^J y_{i,j} \left|\int g_j(x_i, \beta, \gamma)(dF_1 - dF_2)\right| / NJ \\
&+ \sum_{i=1}^N \sum_{j=1}^J \left\{\int g_j(x_i, \beta, \gamma)(dF_1 + dF_2)\right\} \left|\int g_j(x_i, \beta, \gamma)(dF_1 - dF_2)\right| / NJ \\
&\leq 4\sum_{i=1}^N \sum_{j=1}^J \left|\int g_j(x_i, \beta, \gamma)(dF_1 - dF_2)\right| / NJ,
\end{aligned} \tag{21}$$

where the second inequality holds because $y_{i,j}$, $g_j(x_i, \beta, \gamma)$, and $\int g_j(x_i, \beta, \gamma) dF(\beta)$ are uniformly bounded by 1 for all j and x_i. Then because $g_j(x_i, \beta, \gamma)$ is uniformly bounded by 1 and F_1 and F_2 are discrete distributions with the finite support $\mathcal{B}_R$, in this case the weak convergence implies that almost surely $\hat{Q}_N(\gamma, F)$ 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 $|\hat{Q}_N(\gamma, F_1) - \hat{Q}_N(\gamma, F_2)| \to 0$ almost surely.[9] Therefore (B.2.3) holds by Remark A.1.(1) (a) of CP.

[7] 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.

[8] 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, $(\mathcal{F}_R, d_{\text{LP}})$ 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 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. □

[9] 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.

For the case of Assumption 2.5.b, we directly assume (B.2.3) for the reasons stated in Remark 8.

Next there are two conditions to verify in (B.2.4). We first focus on the uniform convergence of the sample criterion function in (B.2.4). Here we need to verify the uniform convergence over $\mathcal{A}_{R(N)}$ such that

$$\sup_{(\gamma,F)\in\Gamma\times\mathcal{F}_{R(N)}} \left|\hat{Q}_N(\gamma,F) - Q(\gamma,F)\right| \xrightarrow{p} 0. \tag{22}$$

For this purpose, it is convenient to view $\hat{Q}_N(\gamma,F)$ and $Q(\gamma,F)$ as functions of γ and $\theta \in \Delta_{R(N)}$ and write them as $\hat{Q}_N(\gamma,\theta)$ and $Q(\gamma,\theta)$, respectively. Then define, for any R, the class of measurable functions

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

and note that $\hat{Q}_N(\gamma,\theta) = N^{-1}\sum_{i=1}^N l(y_i,x_i,\theta,\gamma)$. Then again by Pollard (1984, Theorem II.24), the uniform convergence (22) holds if and only if the entropy satisfies $\log \mathbf{N}(\varepsilon,\tilde{\mathcal{G}}_R,||\cdot||_{L_1,N}) = o_p(N)$ for all $\varepsilon > 0$. Note that the entropy measure of $\tilde{\mathcal{G}}_R$ is bounded by the sum of two entropies, one associated with $\mathcal{F}_{R(N)}$ and the other one associated with Γ. Below we show that the former is $o_p(N)$. We also note that the latter satisfies the entropy condition (and so is $o_p(N)$) under Assumption 2.5a-2.5b by Theorem 2.7.11 of van der Vaart and Wellner (1996) for the Lipschitz case and because the class of indicator functions belongs to the Vapnik-Červonenkis class and has a uniformly bounded entropy (Theorem 2.6.7 of van der Vaart and Wellner 1996).

Now we verify the entropy condition associated with $\mathcal{F}_{R(N)}$. Let $\Delta_{R(N)}$ be the $R(N)$ unit simplex. 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. For ease of notation, below we suppress γ (the result holds for any given $\gamma \in \Gamma$) and define for any R, the class of measurable functions

$$\mathcal{G}_R = \left\{ l(y,x,\theta) = ||y - \sum_r \theta^r g(x,\beta^r)||_E^2 / J : \theta \in \Delta_R \right\}. \tag{23}$$

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 entropy condition to satisfy becomes $\log \mathbf{N}(\varepsilon,\mathcal{G}_R,||\cdot||_{L_1,N})/N \xrightarrow{p} 0$ 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

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

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 entropy condition associated with $\mathcal{F}_{R(N)}$ 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 (B.2.4), 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 (B.2.4) (i)) and we can take ν_N to be small enough. We also have $|Q(\Pi_N\alpha_0) - Q(\alpha_0)| \to 0$, which is trivially satisfied by the continuity of $Q(\alpha)$ at α_0 and $d_{\mathrm{LP}}\left(\Pi_N\alpha_0, \alpha_0\right) \to 0$. Therefore because $\liminf_{N\to\infty} \delta\left(N, R(N), \varepsilon\right) > 0$, the condition (B.2.4) (ii) is satisfied.

We have verified all the conditions in Lemma 2 (Lemma A.2 of CP) and this completes the consistency proof.

B.2 Proof of Corollary 1

Similarly to the baseline least squares estimator, using Lemma 2 we show the consistency of the ML estimator. For ease of notation, we consider the model with heterogeneous parameters only. The following proof is parallel to the least squares case.

We start with (B.2.1). First, the condition $Q(F_0) < \infty$ holds under the assumption that $P_j(x, F_0)$ is bounded away from zero for all $j \leq J$. Similarly to the least squares case, next we show that for $\varepsilon > 0$

$$\liminf_{F\in\mathcal{F}_{R(N)}: d_{\mathrm{LP}}(F,F_0)\geq\varepsilon} Q(F) - Q(F_0) \quad > 0. \tag{24}$$

Note that for any $F \neq F_0$, we have

$$\begin{aligned} Q(F) - Q(F_0) &= -\left(E\left[\sum_{j=1}^J y_j \log P_j(x,F)\right] - E\left[\sum_{j=1}^J y_j \log P_j(x,F_0)\right]\right) \\ &= -E\left[\log\left(\prod_{j=1}^J P_j(x,F)/P_j(x,F_0)^{y_j}\right)\right] \\ &> -\log\left(E\left[\prod_{j=1}^J \left(P_j(x,F)/P_j(x,F_0)\right)^{y_j}\right]\right) = -\log\left(E\left[\sum_{j=1}^J P_j(x,F)\right]\right) = 0, \end{aligned} \tag{25}$$

where the inequality holds by Jensen's inequality. Here the strict inequality holds because $P\left(x, F_0\right) \neq P\left(x, F\right)$ with positive probability for any $F \neq F_0$ (Assumption 1.4). The third equality holds by the law of iterated expectation and $\Pr[y_j = 1|x] = P_j(x, F_0)$. The last result holds because $\sum_{j=1}^J P_j(x, F) = 1$ by construction. Therefore, (24) is satisfied because (25) holds for any $F \in \mathcal{F}$ such that $d_{\mathrm{LP}}(F, F_0) \geq \varepsilon$, by essentially the same argument for the least squares case.

Next, (B.2.2) holds by the same argument for the proof of the least squares estimator.

Next, we show (B.2.3) holds. We use Remark A.1.(i)(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 $\mathcal{B}$. Second we need to show for any data

$((y_i, x_i))_{i=1}^N$, $\hat{Q}_N(F)$ is continuous on $\mathcal{F}_R$ for each $R \geq 1$. Using the inequality that for $0 < c \leq a, b$, $|\log a - \log b| \leq |a - b|/c$, we obtain for some constant C

$$\begin{aligned}
\left|\hat{Q}_N(F_1) - \hat{Q}_N(F_2)\right| &\leq \sum_{i=1}^N \sum_{j=1}^J y_{i,j} \left|\log P_j(x_i, F_1) - \log P_j(x_i, F_2)\right| / N \\
&\leq C \sum_{i=1}^N \sum_{j=1}^J y_{i,j} \left|P_j(x_i, F_1) - P_j(x_i, F_2)\right| / N \\
&= C \sum_{i=1}^N \sum_{j=1}^J y_{i,j} \left|\int g_j(x_i, \beta)(dF_1 - dF_2)\right| / N
\end{aligned}$$

and therefore (B.2.3) holds by the essentially same argument to the proof of the least squares case (see the discussion below (21)).

Next there are two conditions to verify in (B.2.4). We first focus on the uniform convergence of the sample criterion function. Here we need to verify the uniform convergence over $\mathcal{F}_{R(N)}$ such that

$$\sup_{F \in \mathcal{F}_{R(N)}} \left|\hat{Q}_N(F) - Q(F)\right| \xrightarrow{p} 0. \tag{26}$$

For this purpose, it is convenient to view $\hat{Q}_N(F)$ and $Q(F)$ as functions of $\theta \in \Delta_{R(N)}$ and write them as $\hat{Q}_N(\theta)$ and $Q(\theta)$, respectively. Then define, for any R, the class of measurable functions

$$\mathcal{G}_R^{ML} = \left\{ l(y, x, \theta) = \sum_{j=1}^J y_j \log\left(\sum_r \theta^r g_j(x, \beta^r)\right) : \theta \in \Delta_R \right\}. \tag{27}$$

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) $E\left[\sup_{\theta \in \Delta_{R(N)}} |l(y, x, \theta)|\right] < \infty$. Then by Pollard (1984, Theorem II.24) (also see Chen (2007, Section 3.1, page 5592) for related discussion), the entropy condition becomes $\log \mathbf{N}(\varepsilon, \mathcal{G}_R^{ML}, ||\cdot||_{L_1,N})/N \xrightarrow{p} 0$ for all $\varepsilon > 0$. Using the inequality that for $0 < c \leq a, b$, $|\log a - \log b| \leq |a - b|/c$, we obtain for some constant C

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

with some function $E[M(\cdot)^2] < \infty$. Therefore, the first condition in (B.2.4) holds by the essentially same argument to the proof of the least squares case. Finally, the second condition in (B.2.4) also holds by the essentially same argument to the proof of the least squares case. We have verified all the conditions in Lemma 2 (Lemma A.2 of CP) for the ML estimator.

B.3 Proof of Theorem 3

We can prove Theorem 3 by verifying conditions (B.2.1)–(B.2.4) in Lemma 2, just as in the proof of Theorem 1. Observe that (B.2.1)–(B.2.2) are clearly satisfied because the arguments that (B.2.1)–(B.2.2) are satisfied within the proof of Theorem 1 follow almost exactly if we replace $g(x, \beta^r)$, $\mathcal{F}_R$, $\mathcal{F}$, and $Q(F)$ with $\tilde{g}(x, \lambda^r)$, $\mathcal{P}_{\lambda,R}$, $\mathcal{P}_\lambda$, and $Q(P_\lambda)$. The verification of (B.2.3) also follows the arguments in the proof of Theorem 1, instead using $\hat{Q}_N(P_\lambda)$ and Λ_R. Condition (B.2.4), regarding the uniform convergence of $\hat{Q}_N(P_\lambda)$ to $Q(P_\lambda)$, is satisfied by invoking the same arguments as in the proof Theorem 1, using $\hat{Q}_N(P_\lambda)$ and $Q(P_\lambda)$. Other arguments are also essentially identical to Theorem 1. Therefore $d_{\text{LP}}\left(\hat{P}_{\lambda,N}, P_{\lambda,0}\right) \xrightarrow{p} 0$, which also implies $d_{\text{LP}}\left(\hat{F}_N, F_0\right) \xrightarrow{p} 0$ because we assume $F_0 \in \mathcal{F}^M$ and because $F = \int \Phi(\cdot|\lambda) dP_\lambda(d\lambda)$ is continuous on $\mathcal{P}_\lambda$ in the Lévy-Prokhorov metric. We omit a complete proof of Theorem 3 to eliminate redundancy.

B.4 Proof of Theorem 4

Define $\tilde{g}^S(x, \lambda^r) = \frac{1}{S}\sum_{s=1}^S g(x, \beta^{r,s})$ where $\beta^{r,s}$ are drawn from $\Phi(\beta|\lambda^r)$ and define $\hat{Q}_{N,S}(P_\lambda) = N^{-1}\sum_{i=1}^N \left\|y_i - \sum_r \theta^r \tilde{g}^S(x_i, \lambda^r)\right\|_E^2 / J$. Then define the estimator of $P_{\lambda,0}$ as

$$\hat{P}_{\lambda,N,S} = \text{argmin}_{P_\lambda \in \mathcal{P}_{\lambda,R(N)}} \hat{Q}_{N,S}(P_\lambda) + C \cdot \nu_N \tag{28}$$

with some tolerance of minimization, $C \cdot \nu_N$, that tends to zero (if this is necessary).

We prove the consistency by verifying conditions in Lemma 2 for the simulated distribution function estimator. Note that (B.2.1)–(B.2.2) are clearly satisfied because the proofs of (B.2.1)-(B.2.2) in Theorem 1 hold by replacing $g(x, \beta^r)$, $\mathcal{F}_R$, $\mathcal{F}$, and $Q(F)$ with $\tilde{g}^S(x, \lambda^r)$, $\mathcal{P}_{\lambda,R}$, $\mathcal{P}_\lambda$, and $Q(P_\lambda)$, respectively, when necessary. We focus on (B.2.3) and (B.2.4).

To show (B.2.3) holds, we use Remark A.1.(i)(a) of CP. First note that $\mathcal{P}_{\lambda,R}$ is a compact subset of $\mathcal{P}_\lambda$ for each R because Λ_R is a compact subset of Λ. Second we need to show that for any data $((y_i, x_i))_{i=1}^N$, $\hat{Q}_{N,S}(P_\lambda)$ is continuous on $P_{\lambda,R}$ for each $R \geq 1$. Since $P_{\lambda,R}$ is compact, this continuity means our estimator is well defined as the minimum in (28). Note $\int \tilde{g}^S(x, \lambda)\, dP_{\lambda,l} = \sum_{r=1}^R \theta_l^r \tilde{g}^S(x, \lambda^r)$ for $P_{\lambda,l} \in \mathcal{P}_{\lambda,R}, l = 1, 2$. Then, for any $P_{\lambda,1}, P_{\lambda,2} \in \mathcal{P}_{\lambda,R(N)}$, applying the triangle inequality, we obtain

$$\begin{aligned}
\left|\hat{Q}_{N,S}(P_{\lambda,1}) - \hat{Q}_{N,S}(P_{\lambda,2})\right| &\leq 2\sum_{i=1}^N \sum_{j=1}^J y_{i,j} \left|\int \tilde{g}_j^S(x_i, \lambda)(dP_{\lambda,1} - dP_{\lambda,2})\right| / NJ \\
&+ \sum_{i=1}^N \sum_{j=1}^J \left\{\int \tilde{g}_j^S(x_i, \lambda)(dP_{\lambda,1} + dP_{\lambda,2})\right\} \left|\int \tilde{g}_j^S(x_i, \lambda)(dP_{\lambda,1} - dP_{\lambda,2})\right| / NJ \\
&\leq 4\sum_{i=1}^N \sum_{j=1}^J \left|\int \tilde{g}_j^S(x_i, \lambda)(dP_{\lambda,1} - dP_{\lambda,2})\right| / NJ,
\end{aligned}$$

where the second inequality holds because $y_{i,j}$, $\tilde{g}_j^S(x_i, \lambda)$, and $\int \tilde{g}_j^S(x_i, \lambda) dP_\lambda$ are uniformly bounded by 1 for all j and x_i. Then because $\tilde{g}_j^S(x_i, \lambda)$ is uniformly bounded by 1 and $P_{\lambda,1}$ and $P_{\lambda,2}$ are discrete distributions with the finite support Λ_R, weak convergence implies that almost surely $\hat{Q}_{N,S}(P_\lambda)$ is continuous on $\mathcal{P}_{\lambda,R}$, i.e. for any $P_{\lambda,1}, P_{\lambda,2} \in \mathcal{P}_{\lambda,R}$ such that $d_{LP}(P_{\lambda,1}, P_{\lambda,2}) \to 0$, it follows that $|\hat{Q}_{N,S}(P_{\lambda,1}) - \hat{Q}_{N,S}(P_{\lambda,2})| \to 0$ almost surely. Therefore, by Remark A.1.(i) (a) of CP, (B.2.3) holds with simulated basis functions with any S.

Next we verify (B.2.4). (B.2.4) (ii) is clearly satisfied as in Theorem 1. We focus on (B.2.4) (i): the uniform convergence of $\hat{Q}_{N,S}(P_\lambda)$ to $Q(P_\lambda)$ for $P_\lambda \in \mathcal{P}_{\lambda,R(N)}$, $\sup_{P_\lambda \in \mathcal{P}_{\lambda,R(N)}} |\hat{Q}_{N,S}(P_\lambda) - Q(P_\lambda)| \xrightarrow{p} 0$. As in Theorem 1 it is convenient to view $\hat{Q}_{N,S}(P_\lambda)$ and $Q(P_\lambda)$ for $P_\lambda \in \mathcal{P}_{\lambda,R(N)}$ as functions of $\theta \in \Delta_{R(N)}$ and then write them as $\hat{Q}_{N,S}(\theta)$ and $Q(\theta)$, respectively. Then the uniform convergence condition to verify becomes

$$\sup_{\theta \in \Delta_{R(N)}} \left| \hat{Q}_{N,S}(\theta) - Q(\theta) \right| \xrightarrow{p} 0. \tag{29}$$

Define, for any S, $Q_S(\theta) \equiv E\left[\left\| y_i - \sum_r \theta^r \tilde{g}^S(x_i, \lambda^r) \right\|_E^2 / J \right]$, where the expectation is taken with respect to (y_i, x_i) while fixing the simulation draws in $\tilde{g}^S(x_i, \lambda^r)$. We first show that

$$\sup_{\theta \in \Delta_{R(N)}} \left| \hat{Q}_{N,S}(\theta) - Q_S(\theta) \right| \xrightarrow{p} 0.$$

Later we show $\sup_{\theta \in \Delta_{R(N)}} |Q_S(\theta) - Q(\theta)| \xrightarrow{p} 0$; therefore invoking the triangle inequality we obtain (29).

For any R define the class of measurable functions

$$\mathcal{G}_{R,S} = \left\{ l(y, x, \theta) = \left\| y - \sum_r \theta^r \tilde{g}^S(x, \lambda^r) \right\|_E^2 / J : \theta \in \Delta_R \right\}.$$

Note (i) $\hat{Q}_{N,S}(\theta) = N^{-1} \sum_{i=1}^N l(y_i, x_i, \theta)$, (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 (29) holds if and only if $\log \mathbf{N}(\varepsilon, \mathcal{G}_{R,S}, ||\cdot||_{L_1,N})/N \xrightarrow{p} 0$ 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 $\{(y_i, x_i)\}_{i=1}^N$. Then following the same steps in the proof of Theorem 1, we conclude as long as $R(N) \log R(N)/N \to 0$, the uniform convergence of $\hat{Q}_{N,S}(\theta)$ to $Q_S(\theta)$ holds for any given S.

Next we show $\sup_{\theta \in \Delta_{R(N)}} |Q_S(\theta) - Q(\theta)| \xrightarrow{p} 0$ in $\mathbb{P}^*$ where $\mathbb{P}^*$ denotes probability measure w.r.t.

the simulation draws. Consider that

$$
\begin{aligned}
& |Q_S(\theta) - Q(\theta)| \\
= \;& E\left[\left\|y_i - \sum_r \theta^r \tilde{g}^S(x_i,\lambda^r)\right\|_E^2 - \left\|y_i - \sum_r \theta^r \tilde{g}(x_i,\lambda^r)\right\|_E^2\right] \\
\leq \;& J^{-1}\sum_{j=1}^J E\left[\left|\left(y_{i,j} - \sum_r \theta^r \tilde{g}_j^S(x_i,\lambda^r)\right)^2 - \left(y_{i,j} - \sum_r \theta^r \tilde{g}_j(x_i,\lambda^r)\right)^2\right|\right] \\
\leq \;& J^{-1}\sum_{j=1}^J E\left[\left|\left(2y_{i,j} - \sum_r \theta^r \tilde{g}_j^S(x_i,\lambda^r) - \sum_r \theta^r \tilde{g}_j(x_i,\lambda^r)\right)\left(\sum_r \theta^r \tilde{g}_j^S(x_i,\lambda^r) - \sum_r \theta^r \tilde{g}_j(x_i,\lambda^r)\right)\right|\right] \\
\leq \;& 2J^{-1}\sum_{j=1}^J E\left[\left|\sum_r \theta^r \tilde{g}_j^S(x_i,\lambda^r) - \sum_r \theta^r \tilde{g}_j(x_i,\lambda^r)\right|\right] \\
\leq \;& 2J^{-1}\sum_{j=1}^J \sum_r \theta^r E\left[\left|\tilde{g}_j^S(x_i,\lambda^r) - \tilde{g}_j(x_i,\lambda^r)\right|\right] \\
\leq \;& 2J^{-1}\sum_{j=1}^J \sum_r \theta^r \max_r E\left[\left|\frac{1}{S}\sum_{s=1}^S \left\{g_j(x_i,\beta^{r,s}) - \int g_j(x_i,\beta)d\Phi(\beta|\lambda^r)\right\}\right|\right] = o_{\mathbb{P}^*}(1),
\end{aligned}
$$

where the third inequality holds because $0 < \sum_r \theta^r \tilde{g}_j^S(x_i,\lambda^r) \leq 1$ and $0 < \sum_r \theta^r \tilde{g}_j(x_i,\lambda^r) \leq 1$ and the last result holds because $\sum_{r=1}^R \theta^r = 1$ and $\beta^{r,s}$ are iid draws from $\Phi(\beta|\lambda^r)$, so $E^*[g_j(x,\beta^{r,s})] = \int g_j(x,\beta)d\Phi(\beta|\lambda^r) < \infty$ for each r and we applied a LLN to the term inside the $|\cdot|$ bracket for each r, which holds for any $x \in \mathcal{X}$. Note that the convergence of the partial sum inside the $|\cdot|$ bracket is independent of r because we use S simulation draws independently drawn for each λ^r.

Here $E^*[\cdot]$ denotes expectation w.r.t. the simulation draw. Note that the above $o_{\mathbb{P}^*}(1)$ result does not depend on θ and therefore we also have $\sup_{\theta\in\Delta_{R(N)}} |Q_S(\theta) - Q(\theta)| \xrightarrow{p} 0$. Then invoking the triangle inequality we obtain the uniform convergence in (29).

We have verified all the conditions in Lemma 2 (Lemma A.2 of CP) and therefore showed that $\hat{P}_{\lambda,N,S}$ is a consistent estimator for $P_{\lambda,0}$. This in turn implies the consistency of $\hat{F}_{N,S} = \int \Phi(\cdot|\lambda)\hat{P}_{\lambda,N,S}(d\lambda)$ to $F_0 = \int \Phi(\cdot|\lambda)P_{\lambda,0}(d\lambda)$ because $\Phi(\beta|\lambda)$ is continuous in λ for all β and thus $F = \int \Phi(\cdot|\lambda)dP_\lambda(d\lambda)$ is also continuous on $\mathcal{P}_\lambda$ in the Lévy-Prokhorov metric.

C Proofs for Asymptotic Bounds Results

We let C, C_1, C_2, ... denote generic positive constants. We use $\text{diag}\,(A)$ to denote a diagonal matrix composed of diagonal elements of a matrix A. We often use the following inequality (denoted by RCS for the Cauchy-Schwarz inequality for R terms in a sum): $\sum_{r=1}^R W_r \leq \sqrt{R}\sqrt{\sum_{r=1}^R W_r^2}$ for a sequence W_r's. We note $\|P\,(x,F_0)\|_\infty \leq \zeta_0$ and $\|g\,(x,\beta^r)\|_\infty \leq \zeta_0$ for some constant $\zeta_0 > 0$ uniformly over $r \leq R(N)$ because probabilities are bounded by one. We also note for some $c_0 > 0$, $\|g(x,\beta^r)\|_{L_2} \geq c_0$ uniformly over $r \leq R(N)$ because $\mathcal{X}$, the support of X, has positive measure and the support of β is bounded. We first present preliminary lemmas that are useful to prove Theorem 5 and Theorem 6.

We define $\Psi_{N,R} = \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} g_j(x_i,\beta^r)g_j(x_i,\beta^{r'})\right)_{1\le r,r'\le R}$ (a $R\times R$ matrix).

Lemma 3. *Suppose $\mathcal{X}$ has positive measure and Assumption 4(i)-(iii) hold. Then*

$$\min\left\{\Pr\left\{||g(\cdot,\beta^r)||^2_{L_{2,N}} \le 2||g(\cdot,\beta^r)||^2_{L_2},\ \forall r\right\}, \Pr\left\{||g(\cdot,\beta^r)||_{L_2} \le 2||g(\cdot,\beta^r)||_{L_{2,N}},\ \forall r\right\}\right\} \ge 1 - R\exp\left(-C_1 N c_0^4/\zeta_0^4\right).$$

Proof. The claim follows from the union bound applied to the r-specific events and Hoeffding's inequality. □

Lemma 3 implies that $\mathrm{diag}(\Psi_{N,R}) \le 2\mathrm{diag}(\Psi_R)$ holds with probability approaching one (w.p.a.1) because $\|g(\cdot,\beta^r)\|^2_{L_{2,N}}$'s are diagonal elements of $\Psi_{N,R}$ and $\|g(\cdot,\beta^r)\|^2_{L_2}$'s are diagonal elements of Ψ_R. The purpose of the following Lemma 4 is to show that $\Psi_{N,R} \ge \Psi_R/2$ holds w.p.a.1.

Lemma 4. *Let $\mathcal{G} = span\{g(\cdot,\beta^1),\ldots,g(\cdot,\beta^R)\}$ be the linear space spanned by functions $g(\cdot,\beta^1),\ldots,g(\cdot,\beta^R)$. Suppose Assumptions 4 (i)-(iii) hold. Then* $\Pr\left\{\sup_{\mu\in\mathcal{G}\setminus\{0\}}(\|\mu\|^2_{L_2}/\|\mu\|^2_{L_{2,N}}) > 2\right\} \le R^2\exp\left(-C_2 N/(\zeta_0^4 R^2)\right)$.

Proof. Let $\phi_1,\ldots,\phi_M$ be an orthonormal basis of $\mathcal{G}$ in $L_2(\varpi)$ with $M\le R$. Also let $\overline{\rho}(D)$ denote the following quantity for a symmetric matrix D: $\overline{\rho}(D) = \sup\sum_l |a_l|\sum_{l'}|a_{l'}|\left|D_{l,l'}\right|$, where the sup is taken over sequences $\{a_l\}_{l=1}^M$ with $\sum_{l=1}^M a_l^2 = 1$. Then, following Lemma 5.2 in Baraud (2002), we have

$$\Pr\left\{\sup_{\mu\in\mathcal{G}\setminus\{0\}}\frac{\|\mu\|^2_{L_2}}{\|\mu\|^2_{L_{2,N}}} > c\right\} \le M^2\exp\left(-N\frac{(\varpi_0 - c^{-1})^2}{4\varpi_1\max\left\{\overline{\rho}^2(A),\overline{\rho}(B)\right\}}\right) \tag{30}$$

where $A_{l,l'} = \sqrt{E[||\phi_l||^2_E||\phi_{l'}||^2_E]}$ and $B_{l,l'} = \left\|\|\phi_l\|_E\|\phi_{l'}\|_E\right\|_\infty$ for $l,l' = 1,\ldots,M$ and ϖ_0 and ϖ_1 denote the lower bound and upper bound of the density of X, respectively. We find $\left|A_{l,l'}\right| \le \zeta_0^2$ and $\left|B_{l,l'}\right| \le \zeta_0^2$. It follows that

$$\overline{\rho}(A) \le \zeta_0^2\sup\sum_l|a_l|\sum_{l'}|a_{l'}| = \zeta_0^2\sup\left(\sum_l|a_l|\right)^2 \le \zeta_0^2\sup M\sum_l|a_l|^2 = \zeta_0^2 M \le \zeta_0^2 R$$

where $(\sum_l|a_l|)^2 \le M\sum_l|a_l|^2$ holds by the Cauchy-Schwarz inequality. Similarly we have $\overline{\rho}(B)\le\zeta_0^2 R$. The conclusion follows from (30) and $M\le R$. □

Now define an event $\mathbf{E}_0 = \left\{\Psi_{N,R} - \left(\xi_{\min}(R)/4\zeta_0^2\right)\mathrm{diag}(\Psi_{N,R}) \ge 0\right\}$.

Lemma 5. *Suppose $\mathcal{X}$ has positive measure and Assumption 4(i)-(iii) hold. Then,*
$\Pr\{\mathbf{E}_0\} \ge 1 - R\exp\left(-C_1 N c_0^4/\zeta_0^4\right) - R^2\exp\left(-C_2 N/\zeta_0^4 R^2\right)$

Proof. First, note that $\Psi_R - (\xi_{\min}(R)/\zeta_0^2)\mathrm{diag}(\Psi_R) \ge 0$. Now let $\mathcal{G}$ be the linear space spanned by $g(\cdot,\beta^1),\ldots,g(\cdot,\beta^R)$. Now note that, under the event $A = \left\{||g(\cdot,\beta^r)||^2_{L_{2,N}} \le 2||g(\cdot,\beta^r)||^2_{L_2}\ \forall r = 1,\ldots,R\right\}$, we have $\mathrm{diag}(\Psi_{N,R}) \le 2\mathrm{diag}(\Psi_R)$. Also, under the event $B = \left\{\sup_{\mu\in\mathcal{G}\setminus\{0\}}\left(\|\mu\|^2_{L_2}/\|\mu\|^2_{L_{2,N}}\right) \le 2\right\}$, we have $\Psi_{N,R} \ge \Psi_R/2$. Therefore, under the intersection of the two events, A and B, we have

$$\Psi_{N,R} - \xi_{\min}(R)\mathrm{diag}(\Psi_{N,R})/(4\zeta_0^2) \ge \Psi_R/2 - 2\xi_{\min}(R)\mathrm{diag}(\Psi_R)/(4\zeta_0^2) \ge 0,$$

or event $\mathbf{E}_0$. Then by Lemma 3 and Lemma 4, we find $1 - \Pr\{\mathbf{E}_0\} \leq R \cdot \exp\left(-C_1 N c_0^4/\zeta_0^4\right) + R^2 \cdot \exp\left(-C_2 N/\zeta_0^4 R^2\right)$ because $\Pr\{\mathbf{E}_0^c\} \leq \Pr\{A^c \cup B^c\} \leq \Pr\{A^c\} + \Pr\{B^c\}$. □

Lemma 6. *Suppose $\mathcal{X}$ has positive measure and Assumption 4(i)-(iii) hold. Then, for given positive sequence η_N*

$$\begin{aligned} & \Pr\left\{\left|\sum\nolimits_i e_i' g(X_i, \beta^r)/N\right| \leq \eta_N ||g(\cdot, \beta^r)||_{L_{2,N}} \text{ for all } r = 1, \ldots, R\right\} \\ \geq \; & 1 - R \cdot \exp\left(-CN\eta_N^2 c_0^2/\zeta_0^2\right) - R \cdot \exp\left(-C_1 N c_0^4/\zeta_0^4\right). \end{aligned}$$

Proof. Hoeffding (1963)'s inequality implies that

$$\begin{aligned} & E_X\left[\Pr\left\{\left|\sum\nolimits_i e_i' g(X_i, \beta^r)/N\right| \geq \eta_N ||g(\cdot, \beta^r)||_{L_{2,N}}, \forall r \leq R \,|X_1, \ldots, X_N\right\}\right] \\ \leq \; & E_X\left[\sum\nolimits_r \exp\left(-2N\eta_N^2 ||g(\cdot, \beta^r)||_{L_{2,N}}^2/4J\zeta_0^2\right)\right] \end{aligned} \tag{31}$$

because $E\left[e_i' g\left(X_i, \beta^r\right) \mid X_1, \ldots, X_N\right] = 0$, choice probabilities lie in $[0,1]$, and $-\sqrt{J}\zeta_0 \leq e_i' g(X_i, \beta^r) \leq \sqrt{J}\zeta_0$ (by the Cauchy-Schwarz inequality) uniformly.

Now note under the event $\left\{||g(\cdot, \beta^r)||_{L_2} \leq 2||g(\cdot, \beta^r)||_{L_{2,N}},\ \forall r = 1, \ldots, R\right\}$,

$$\begin{aligned} \sum\nolimits_{r=1}^R \exp\left(-N\eta_N^2 \left\|g\left(\cdot, \beta^r\right)\right\|_{L_{2,N}}^2/2J\zeta_0^2\right) &\leq \sum\nolimits_{r=1}^R \exp\left(-N\eta_N^2 \left\|g\left(\cdot, \beta^r\right)\right\|_{L_2}^2/8J\zeta_0^2\right) \\ &\leq \sum\nolimits_{r=1}^R \exp\left(-N\eta_N^2 c_0^2/8J\zeta_0^2\right) = R\exp\left(-CN\eta_N^2 c_0^2/\zeta_0^2\right). \end{aligned} \tag{32}$$

From (31)-(32) and Lemma 3, the claim follows. □

The following Lemma 7 decomposes the error bound into a bias term and a variance term.

Lemma 7. *Suppose $\mathcal{X}$ has positive measure and Assumption 4(i)-(iii) hold. Then, for any $N \geq 1$, $R \geq 2$, and $a > 1$, we have for all $F \in \mathcal{F}_{R(N)}$,*

$$\left\|P(x_i, \hat{F}_N) - P\left(x_i, F_0\right)\right\|_{L_{2,N}}^2 \leq \frac{a+1}{a-1}\left\|P(x_i, F) - P\left(x_i, F_0\right)\right\|_{L_{2,N}}^2 + C\frac{\eta_N^2 R\zeta_0^2}{\xi_{\min}(R)}\frac{a^2}{a-1},$$

where the inequality holds with probability greater than $1 - p_{N,R}$,
$p_{N,R} \equiv R\exp\left(-CN\eta_N^2 c_0^2/\zeta_0^2\right) + 2R\exp\left(-C_1 N c_0^4/\zeta_0^4\right) + R^2\exp\left(-C_2 N/\zeta_0^2 R^2\right)$.

Proof. Because $P(x_i, \hat{F}_N)$ (i.e., $\hat{F}_N$) is the solution of the minimization problem in (6), we have

$$||P(x_i, \hat{F}_N) - y_i||_{L_{2,N}}^2 \leq ||P(x_i, F) - y_i||_{L_{2,N}}^2 \tag{33}$$

for any $F \in \mathcal{F}_{R(N)}$. Now note that

$$\begin{aligned} \left\|P(x_i, \hat{F}_N) - y_i\right\|_{L_{2,N}}^2 &= \left\|P(x_i, \hat{F}_N) - P(x_i, F_0) + P(x_i, F_0) - y_i\right\|_{L_{2,N}}^2 \\ &= \left\|P(x_i, \hat{F}_N) - P(x_i, F_0)\right\|_{L_{2,N}}^2 - 2\sum\nolimits_i e_i'\left(P(x_i, \hat{F}_N) - P(x_i, F_0)\right)/N + ||e_i||_{L_{2,N}}^2, \end{aligned} \tag{34}$$

where we use the definition $e_i = y_i - P(x_i, F_0)$. Similarly we obtain

$$\left\|P(x_i,\hat{F}_N) - y_i\right\|_{L_{2,N}}^2 = \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2 - 2\sum\nolimits_i e_i' \left(P(x_i,F) - P(x_i,F_0)\right)/N + ||e_i||_{L_{2,N}}^2. \tag{35}$$

Subtracting (35) from (34) and by (33), we obtain

$$\left\|P(x_i,\hat{F}_N) - P(x_i,F_0)\right\|_{L_{2,N}}^2 \leq \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2 + 2\sum_i e_i'(P(x_i,\hat{F}_N) - P(x_i,F))/N.$$

Let $V_{N,r} = \frac{1}{N}\sum_{i=1}^N e_i' g(x_i,\beta^r)$. Then, $\frac{1}{N}\sum_i e_i'(P(x_i,\hat{F}_N) - P(x_i,F)) = \sum_{r=1}^R V_{N,r}(\widehat{\theta}^r - \theta^r)$ by construction of $P(x_i,\hat{F}_N)$ and $P(x_i,F)$ and we obtain by the triangle inequality,

$$\left\|P(x_i,\hat{F}_N) - P(x_i,F_0)\right\|_{L_{2,N}}^2 \leq \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2 + 2\sum_r |V_{N,r}| \cdot |\widehat{\theta}^r - \theta^r|. \tag{36}$$

Define the event $\mathbf{E}_1 = \bigcap_{r=1}^R \left\{|V_{N,r}| \leq \eta_N ||g(\cdot,\beta^r)||_{L_{2,N}}\right\}$. Then under $\{\mathbf{E}_0 \cap \mathbf{E}_1\}$, we have

$$\begin{aligned}
&\sum\nolimits_{r=1}^R V_{N,r}^2(\widehat{\theta}^r - \theta^r)^2 \leq \eta_N^2 \sum\nolimits_{r=1}^R ||g(\cdot,\beta^r)||_{L_{2,N}}^2 (\widehat{\theta}^r - \theta^r)^2 \qquad (37)\\
= \;& \eta_N^2 \frac{1}{N}\sum\nolimits_{i=1}^N \sum\nolimits_{r=1}^R (\widehat{\theta}^r - \theta^r)^2 ||g(x_i,\beta^r)||_E^2 = \eta_N^2 (\widehat{\theta} - \theta)' \mathrm{diag}(\Psi_{N,R})(\widehat{\theta} - \theta)\\
\leq \;& \eta_N^2 \left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1} (\widehat{\theta} - \theta)' \Psi_{N,R} (\widehat{\theta} - \theta) = \eta_N^2 \left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1} ||P(x_i,\hat{F}_N) - P(x_i,F)||_{L_{2,N}}^2.
\end{aligned}$$

From (36) and (37), it follows that under the event $\{\mathbf{E}_0 \cap \mathbf{E}_1\}$,

$$\begin{aligned}
&\left\|P(x_i,\hat{F}_N) - P(x_i,F_0)\right\|_{L_{2,N}}^2 \leq \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2 + 2\sum\nolimits_{r=1}^R |V_{N,r}| \cdot |\widehat{\theta}^r - \theta^r|\\
\leq \;& \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2 + 2\sqrt{R}\left(\sum\nolimits_{r=1}^R V_{N,r}^2 \left(\widehat{\theta}^r - \theta^r\right)^2\right)^{1/2}\\
\leq \;& \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2 + 2\eta_N\sqrt{R}\left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1/2} ||P(x_i,\hat{F}_N) - P(x_i,F)||_{L_{2,N}}\\
\leq \;& \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2\\
&+2\eta_N\sqrt{R}\left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1/2}\left(\left\|P(x_i,\hat{F}_N) - P(x_i,F_0)\right\|_{L_{2,N}} + \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}\right)
\end{aligned}$$

by the Cauchy-Schwarz inequality, RCS, and the triangle inequality. Applying the inequality $2xy \leq x^2 a + y^2/a$ (any x,y, $a > 0$) to $x = \eta_N\sqrt{R}(\xi_{\min}(R)/4\zeta_0^2)^{-1/2}$ and $y = ||P(x_i,\hat{F}_N) - P(x_i,F_0)||_{L_{2,N}}$ and to $x = \eta_N\sqrt{R}(\xi_{\min}(R)/4\zeta_0^2)^{-1/2}$ and $y = ||P(x_i,F) - P(x_i,F_0)||_{L_{2,N}}$, respectively, we obtain under the event $\{\mathbf{E}_0 \cap \mathbf{E}_1\}$,

$$\begin{aligned}
\left\|P(x_i,\hat{F}_N) - P(x_i,F_0)\right\|_{L_{2,N}}^2 \leq \;& \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2 + \left\|P(x_i,\hat{F}_N) - P(x_i,F_0)\right\|_{L_{2,N}}^2/a\\
&+ \|P(x_i,F) - P(x_i,F_0)\|_{L_{2,N}}^2/a + 2a\eta_N^2 R\left(\xi_{\min}(R)/4\zeta_0^2\right)^{-1}
\end{aligned}$$

It follows that under the event $\{\mathbf{E_0} \cap \mathbf{E}_1\}$, for all $a > 1$,

$$\left\| P(x_i, \hat{F}_N) - P(x_i, F_0) \right\|_{L_{2,N}}^2 \leq \frac{a+1}{a-1} \left\| P(x_i, F) - P(x_i, F_0) \right\|_{L_{2,N}}^2 + C \frac{\eta_N^2 R \zeta_0^2}{\xi_{\min}(R)} \frac{a^2}{a-1}.$$

The conclusion follows from Lemma 6. □

C.1 Proof of Theorem 5

We first obtain the convergence rates of the choice probability as

Lemma 8. *Suppose the assumptions and the conditions in Theorem 5 hold. Then, we have*

$$\left\| P(x_i, \hat{F}_N) - P(x_i, F_0) \right\|_{L_{2,N}}^2 \leq O_{\mathbb{P}} \left(\frac{R(N) \log R(N)}{\xi_{\min}(R(N)) \cdot N} \right) + C \cdot \inf_{F \in \mathcal{F}_{R(N)}} \left\| P(x_i, F) - P\left(x_i, F_0\right) \right\|_{L_{2,N}}^2 .$$

Then the result of Theorem 5 follows from Lemma 8 combined with Lemma 1. Lemma 8 derives the variance term of the asymptotic bounds and Lemma 1 derives the bias term. We prove Lemma 8 below.

C.2 Proof of Lemma 8

From Lemma 7, the fastest convergence rate will be obtained when the order of η_N is as small as possible while keeping $p_{N,R} \to 0$. By inspecting $p_{N,R}$, we note that the optimal rate is obtained when we choose $\eta_N = C\sqrt{\log R(N)/N}$ since the first term in $p_{N,R}$ dominates the second term in $p_{N,R}$ when η_N is small enough and $p_{N,R} \to 0$ with this choice of η_N. The inspection of the third term in $p_{N,R}$ reveals that we also require $R(N)$ should satisfy $R(N)^2 \log R(N)/N \to 0$ so that $p_{N,R} \to 0$. The result of Lemma 8 follows from these requirements, Lemma 7 and

$$\left\| P(x_i, \hat{F}_N) - P(x_i, F_0) \right\|_{L_{2,N}}^2 \leq O_{\mathbb{P}} \left(\frac{R(N) \log R(N)}{\xi_{\min}(R(N)) \cdot N} \right) + C \cdot \left\| P(x_i, F) - P\left(x_i, F_0\right) \right\|_{L_{2,N}}^2$$

because $C \frac{\eta_N^2 R(N) \zeta_0^2}{\xi_{\min}(R(N))} \frac{a^2}{a-1} = O\left(\frac{R(N) \log R(N)}{\xi_{\min}(R(N)) \cdot N} \right)$ under our choice of η_N and because Lemma 7 holds for any $F \in \mathcal{F}_{R(N)}$.

C.3 Proof of Lemma 1

First we construct approximating power series with the length of L tensor products of higher order polynomials of β_k's in β as $(\varphi_1(\beta), \ldots, \varphi_l(\beta), \ldots, \varphi_L(\beta))$, where $\varphi_l(\beta)$ is the l^{th} element in the L number of tensor product polynomials. The tensor products are defined by the functions $\varphi_l(\beta) = \beta_1^{l_1} \beta_2^{l_2} \cdots \beta_K^{l_K}$ with $\beta = (\beta_1, \beta_2, \ldots, \beta_K) \in \mathcal{B} = \prod_{k=1}^K [\underline{\beta}_k, \overline{\beta}_k]$ and l_k's are exponents of each β_k. For example, we can let $\varphi_1(\beta) = 1$, $\varphi_2(\beta) = \beta_1$, ... , and $\varphi_l(\beta) = \beta_1^{l_1} \beta_2^{l_2} \cdots \beta_K^{l_K}$.

Now note that each $g_j(x, \beta) f(\beta)$ is a member of the Hölder class ($\overline{s}$-smooth) of functions since it is uniformly bounded and all of its own and partial derivatives up to the order of $\overline{s}$ are also uniformly bounded by our restriction on f in $\mathcal{H}$ and Assumption 4 (iv). Therefore, we can approximate

$g_j(x,\beta)f(\beta)$ well using power series (see Chen, 2007) and obtain the approximation error rate due to Timan (1963) as (we suppress dependence of a_l and φ_l on j for notational simplicity)

$$\sup_{\beta\in\mathcal{B}}\left|g_j(x,\beta)f(\beta)-\sum\nolimits_{l=1}^{L}a_l(x)\varphi_l(\beta)\right|=O\left(L^{-\overline{s}/K}\right) \tag{38}$$

for all $x\in\mathcal{X}$. Let $\left(\underline{\beta}_k=b_{k,1}<b_{k,2}<\cdots<b_{k,r_k+1}=\overline{\beta}_k,\quad k=1,\ldots,K\right)$ be partitions of the intervals $[\underline{\beta}_k,\overline{\beta}_k]$, $k=1,\ldots,K$, into $r_1,\ldots,r_K$ subintervals, respectively. Then we can define the $R=r_1r_2\cdots r_K$ number of subcubes as $\{C_{\iota_1,\ldots,\iota_K}=\prod_{k=1}^{K}[b_{k,\iota_k},b_{k,\iota_k+1}],\ \iota_k=1,2,\ldots,r_k\}$, which become a partition $P(\mathcal{B})$ of $\mathcal{B}$. For any choice of R points

$$(b_{\iota_1,\ldots,\iota_K}\in C_{\iota_1,\ldots,\iota_K}\mid \iota_k=1,2,\ldots,r_k,k=1,\ldots,K)$$

(one $b_{\iota_1,\ldots,\iota_K}$ for each of R subcubes), now we can approximate a Riemann integral of $\sum_{l=1}^{L}a_l(x)\varphi_l(\beta)$ using a quadrature method with R distinct weights

$$(c(\iota_1,\ldots,\iota_K)\equiv c(b_{\iota_1,\ldots,\iota_K})\mid \iota_k=1,\ldots,r_k,k=1,\ldots,K)\ \text{ such that}$$

$$\begin{aligned}\int\sum\nolimits_{l=1}^{L}a_l(x)\varphi_l(\beta)d\beta &= \sum\nolimits_{l=1}^{L}a_l(x)\int\varphi_l(\beta)d\beta\\ &= \sum\nolimits_{l=1}^{L}a_l(x)\sum\nolimits_{C_{\iota_1,\ldots,\iota_K}\in P(\mathcal{B})}c(\iota_1,\ldots,\iota_K)\varphi_k(b_{\iota_1,\ldots,\iota_K})+\mathcal{R}(\delta_R)\end{aligned}$$

where $\mathcal{R}(\delta_R)$ denotes a remainder term with $\delta_R=\max\{\text{diam}(C_{\iota_1,\ldots,\iota_K}):C_{\iota_1,\ldots,\iota_K}\in P(\mathcal{B})\}$. Without loss of generality, we will pick $\delta_R=C\cdot R^{-1/K}$. Noting that $\varphi_l(\beta)$ is a product of polynomials in β_k's by construction, we can apply Theorem 6.1.2 (Generalized Cartesian Product Rules) of Krommer and Ueberhuber (1998) and so we can approximate multivariate integrals with products of univariate integrals. Note that $\int\varphi_l(\beta)d\beta=\prod_{k=1}^{K}\int\varphi_{l,k}(\beta_k)d\beta_k$ with $\varphi_l(\beta)=\prod_{k=1}^{K}\varphi_{l,k}(\beta_k)$. Therefore if we approximate $\int\varphi_{l,k}(\beta_k)d\beta_k$ using a univariate quadrature with weights $\{c_k(1),\ldots,c_k(r_k)\}$, generally we obtain $\int\varphi_{l,k}(\beta_k)d\beta_k=\sum_{\iota_k=1}^{r_k}c_k(\iota_k)\varphi_{l,k}(b_{k,\iota_k})+\mathcal{R}_{l,k}(\delta_R)$ where $\mathcal{R}_{l,k}(\delta_R)$ denotes a possible remainder term. Now we can make the univariate quadrature even become accurate (or exact) at least up to the order of r_k (Theorem 5.2.1 in Krommer and Ueberhuber (1998)) i.e.,$\int\beta_k^p d\beta_k=\sum_{\iota_k=1}^{r_k}c_k(\iota_k)b_{k,\iota_k}^p$with suitable choice of $c_k(\iota_k)\equiv c_k(b_{k,\iota_k})$ for all $p\le r_k$ such that $\mathcal{R}_{l,k}(\delta_R)=0$.

For notational simplicity, we take $r_k=r_1$ for all k. Then with the $L=(r_1+1)^K=(R^{1/K}+1)^K$ number of power series, we can include powers and cross products of β_k's at least up to the order of r_1. With the choice of $L=(r_1+1)^K$ and $c_k(\iota_k)$'s that make the univariate quadrature exact at least up to the order of r_1, we can let

$$\int\varphi_l(\beta)d\beta=\prod\nolimits_{k=1}^{K}\int\varphi_{l,k}(\beta_k)d\beta_k=\prod\nolimits_{k=1}^{K}\sum\nolimits_{\iota_k=1}^{r_k}c_k(\iota_k)\varphi_{l,k}(b_{k,\iota_k}) \tag{39}$$

for $l = 1, \ldots, L$. By adding and subtracting terms, it follows that for some F^* defined later

$$P_j(x, F) - P_j(x, F^*) = \int f(\beta) g_j(x, \beta)\, d\beta - \int \sum_{l=1}^{L} a_l(x) \varphi_l(\beta) d\beta \tag{40}$$

$$+ \sum_{l=1}^{L} a_l(x) \int \varphi_l(\beta) d\beta - \sum_{l=1}^{L} a_l(x) \prod_{k=1}^{K} \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{k,\iota_k}) \tag{41}$$

$$+ \sum_{l=1}^{L} a_l(x) \prod_{k=1}^{K} \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{k,\iota_k}) - P_j(x, F^*). \tag{42}$$

We first bound (40) by the triangle inequality and (38), (40) $\leq \int \left| f(\beta) g_j(x, \beta) - \sum_{l=1}^{L} a_l(x) \varphi_l(\beta) \right| d\beta = O(L^{-\overline{s}/K} \cdot \mathrm{vol}(\mathcal{B}))$. Second note that (41) becomes zero due to (39).

Next construct $\widetilde{\varphi}_l(b^r)$, $r = 1, \ldots, R$ such that $\sum_{r=1}^{R} \widetilde{\varphi}_l(b^r) = \prod_{k=1}^{K} \left\{ \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{k,\iota_k}) \right\}$ with $R = r_1 \cdots r_K$ and $b^r = (b_1^r, b_{2,}^r, \ldots, b_K^r)$ in $\mathbf{b} \equiv \{b : b = (b_{1,\iota_1}, \ldots, b_{K,\iota_K}), \iota_1 = 1, \ldots, r_1, \ldots, \iota_K = 1, \ldots, r_K\}$. Then, for any choice of $b^r \in \mathbf{b}$ we can always write $\widetilde{\varphi}_l(b^r) = c_1(b_{1,\iota_1}) \cdots c_K(b_{K,\iota_K}) \varphi_l(b)$ for some $b = (b_{1,\iota_1}, \ldots, b_{K,\iota_K})$ in $\mathbf{b}$. Therefore without loss of generality write

$$\sum_{r=1}^{R} \widetilde{\varphi}_l(b^r) = \prod_{k=1}^{K} \left\{ \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{k,\iota_k}) \right\} = \sum_{r=1}^{R} c_1(b^r) \cdots c_K(b_K^r) \varphi_l(b^r). \tag{43}$$

Define $\theta^{*r} = c_1(b_1^r) \cdots c_K(b_K^r) f(b^r) / \left\{ \sum_{r=1}^{R} c_1(b_1^r) \cdots c_K(b_K^r) f(b^r) \right\}$ and let F^* be constructed using the weights θ^*. It follows that

$$\begin{aligned}
& \left| \sum_{l=1}^{L} a_l(x) \prod_{k=1}^{K} \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{k,\iota_k}) - P_j(x, F^*) \right| \\
\leq & \sum_{r=1}^{R} \left| \sum_{l=1}^{L} a_l(x) \widetilde{\varphi}_l(b^r) - \theta^{*r} g_j(x, b^r) \right| \\
= & \sum_{r=1}^{R} \left| \sum_{l=1}^{L} a_l(x) c_1(b_1^r) \cdots c_K(b_K^r) \varphi_l(b^r) - \frac{c_1(b_1^r) \cdots c_K(b_K^r) f(b^r)}{\sum_{r=1}^{R} c_1(b_1^r) \cdots c_K(b_K^r) f(b^r)} g_j(x, b^r) \right| \\
\leq & \sum_{r=1}^{R} |c_1(b_1^r) \cdots c_K(b_K^r)| \left| \sum_{l=1}^{L} a_l(x) \varphi_l(b^r) - \frac{f(b^r)}{\sum_{r=1}^{R} c_1(b_1^r) \cdots c_K(b_K^r) f(b^r)} g_j(x, b^r) \right| \\
\leq & \sum_{r=1}^{R} |c_1(b_1^r) \cdots c_K(b_K^r)| \, | \sum_{l=1}^{L} a_l(x) \varphi_l(b^r) - f(b^r) g_j(x, b^r) | \\
& + \sum_{r=1}^{R} |c_1(b_1^r) \cdots c_K(b_K^r)| \left| \frac{\sum_{r=1}^{R} c_1(b_1^r) \cdots c_K(b_K^r) f(b^r) - 1}{\sum_{r=1}^{R} c_1(b_1^r) \cdots c_K(b_K^r) f(b^r)} \right| |f(b^r) g_j(x, b^r)| \\
& = O\left(R \cdot R^{-1} L^{-\overline{s}/K} \right)
\end{aligned}$$

where the first inequality holds by the triangle inequality and (43) and the first equality holds by (43) and by the definition of θ^{*r}. In the above note that $|c_1(b_1^r) \cdots c_K(b_K^r)|$ is at most $O(R^{-1})$, which is obtained if one uses the uniform weights, i.e., $c_k(b_k^1) = \ldots = c_k(b_k^R)$ for $k = 1, \ldots, K$. Second note that $\left| \sum_{l=1}^{L} a_l(x) \varphi_l(b^r) - f(b^r) g_j(x, b^r) \right| = O(L^{-\overline{s}/K})$ due to the bound in (38). Third note that

$\sum_{r=1}^{R} c_1(b_1^r)\cdots c_K(b_K^r)f(b^r)$ is another quadrature approximation of the integral $\int_{\mathcal{B}} f(\beta)d\beta = 1$. Because $f(\beta)$ itself belongs to a Hölder class, the approximation error rate of this integral becomes $\left|\sum_{r=1}^{R} c_1(b_1^r)\cdots c_K(b_K^r)f(b^r) - 1\right| = O\left(L^{-\overline{s}/K}\right)$. To see this, note that

$$
\begin{aligned}
& \left|\int_{\mathcal{B}} f(\beta)d\beta - \sum_{r=1}^{R} c_1(b_1^r)\cdots c_K(b_K^r)f(b^r)\right| \\
\leq\ & \left|\int_{\mathcal{B}} f(\beta)d\beta - \int_{\mathcal{B}} \sum_{l=1}^{L} \tilde{a}_l\varphi_l(\beta)d\beta\right| && (44)\\
& + \left|\sum_{l=1}^{L} \tilde{a}_l \int_{\mathcal{B}} \varphi_l(\beta)d\beta - \sum_{l=1}^{L} \tilde{a}_l \prod_{k=1}^{K} \sum_{\iota_k=1}^{r_k} c_k(\iota_k)\varphi_{l,k}(b_{k,\iota_k})\right| && (45)\\
& + \left|\sum_{l=1}^{L} \tilde{a}_l \prod_{k=1}^{K} \sum_{\iota_k=1}^{r_k} c_k(\iota_k)\varphi_{l,k}(b_{k,\iota_k}) - \sum_{l=1}^{L} \tilde{a}_l \sum_{r=1}^{R} c_1(b^r)\cdots c_K(b_K^r)\varphi_l(b^r)\right| && (46)\\
& + \left|\sum_{r=1}^{R} c_1(b^r)\cdots c_K(b_K^r) \sum_{l=1}^{L} \tilde{a}_l\varphi_l(b^r) - \sum_{r=1}^{R} c_1(b_1^r)\cdots c_K(b_K^r)f(b^r)\right|, && (47)
\end{aligned}
$$

where $\tilde{a}_l$ satisfies $\sup_{\beta\in\mathcal{B}}\left|f(\beta) - \sum_{l=1}^{L} \tilde{a}_l\varphi_l(\beta)\right| = O\left(L^{-\overline{s}/K}\right)$ (such $\tilde{a}_l$'s exist since $f(\beta)$ belongs to a Hölder class). Then we find (44)$= O\left(L^{-\overline{s}/K}\cdot \mathrm{vol}(\mathcal{B})\right)$ by the triangle inequality, (45)$= 0$ by (39), and (46)$= 0$ by 43. Finally note (47)$\leq \sum_{r=1}^{R} c_1(b^r)\cdots c_K(b_K^r)\left|\sum_{l=1}^{L} \tilde{a}_l\varphi_l(b^r) - f(b^r)\right| = O\left(R\cdot R^{-1}\cdot L^{-\overline{s}/K}\right)$ by the triangle equality, $|c_1(b^r)\cdots c_K(b_K^r)|$ is at most $O(R^{-1})$, and because $\tilde{a}_l$ satisfies $\sup_{\beta\in\mathcal{B}}\left|f(\beta) - \sum_{l=1}^{L} \tilde{a}_l\varphi_l(\beta)\right| = O\left(L^{-\overline{s}/K}\right)$.

Combining these results, we bound (42) as $O(L^{-\overline{s}/K})$. Combining these bounds, we then conclude

$$
\begin{aligned}
\left|P_j\left(x,F\right) - P_j\left(x,F^*\right)\right| &= O\left(L^{-\overline{s}/K}\cdot vol(\mathcal{B})\right) + 0 + O\left(L^{-\overline{s}/K}\right) = O\left(L^{-\overline{s}/K}\right) \\
&= O\left(\left(\left(R^{1/K}+1\right)^{K}\right)^{-\overline{s}/K}\right) = O\left(\left(R^{1/K}+1\right)^{-\overline{s}}\right) \leq O\left(R^{-\overline{s}/K}\right).
\end{aligned}
$$

The second conclusion in the lemma is trivial since the above holds for all $x\in\mathcal{X}$ and $j = 1,\ldots,J$.

C.4 Proof of Theorem 6

First we derive the distance between $\hat{\theta}$ and θ_0^* where θ_0^* (i.e. F_0^*) satisfies $\left\|P\left(x_i,F_0^*\right) - P\left(x_i,F_0\right)\right\|_{L_{2,N}}^2 = O_{\mathbb{P}}\left(R^{-2\overline{s}/K}\right)$. Note that such F_0^* exists by Lemma 1. Note that with probability approaching to one, we have $2\left\|g\left(\cdot,\beta^r\right)\right\|_{L_{2,N}} \geq \left\|g\left(\cdot,\beta^r\right)\right\|_{L_2} \geq c_0 > 0$ for all $r = 1,\ldots,R$ by Lemma 3. It follows that

$$
\begin{aligned}
& \frac{c_0}{2}\sum_{r=1}^{R}|\widehat{\theta}^r-\theta_0^{*r}| \qquad (48)\\
\leq\ & \sum_{r=1}^{R}||g(\cdot,\beta^r)||_{L_{2,N}}|\widehat{\theta}^r-\theta_0^{*r}| \leq \sqrt{R}\left(\sum_{r=1}^{R}||g(\cdot,\beta^r)||_{L_{2,N}}^2(\widehat{\theta}^r-\theta_0^{*r})^2\right)^{1/2}\\
\leq\ & \left(4\zeta_0^2/\xi_{\min}(R)\right)^{1/2}\sqrt{R}||P(x_i,\hat{F}_N)-P(x_i,F_0^*)||_{L_{2,N}}\\
\leq\ & \left(4\zeta_0^2/\xi_{\min}(R)\right)^{1/2}\sqrt{R}\left(||P(x_i,\hat{F}_N)-P(x_i,F_0)||_{L_{2,N}}+||P(x_i,F_0^*)-P(x_i,F_0)||_{L_{2,N}}\right)\\
=\ & \sqrt{R(N)/\xi_{\min}(R(N))}\max\left\{O_{\mathbb{P}}\left(\sqrt{\varrho_{R,N}}\right),O_{\mathbb{P}}(R(N)^{-\overline{s}/K})\right\},
\end{aligned}
$$

where the second inequality holds by RCS, the third inequality holds similarly with (37), and the fourth inequality holds by the triangle inequality. Therefore, by Lemma 1, the result of Theorem 5, and the Cauchy-Schwarz inequality, the conclusion follows.

Next we have $\left|\widehat{F}_N(\beta)-F_0(\beta)\right| \leq \left|\widehat{F}_N(\beta)-F_0^*(\beta)\right|+|F_0^*(\beta)-F_0(\beta)|$ by the triangle inequality. It is not difficult to see that

$$
\begin{aligned}
\sup_{\beta\in\mathcal{B}}\left|\widehat{F}_N(\beta)-F_0^*(\beta)\right| &= \sup_{\beta\in\mathcal{B}}\left|\sum_{r=1}^{R}\widehat{\theta}^r 1[\beta^r\leq\beta]-\sum_{r=1}^{R}\theta_0^{*r}1[\beta^r\leq\beta]\right|\\
&= \sup_{\beta\in\mathcal{B}}\left|\sum_{r=1}^{R}(\widehat{\theta}^r-\theta_0^{*r})1[\beta^r\leq\beta]\right| \leq \sum_{r=1}^{R}|\widehat{\theta}^r-\theta_0^{*r}|,
\end{aligned}
$$

where the last inequality holds by the triangle inequality. Note that $F_0(\beta)=\int f_0(b)1[b\leq\beta]\,db$ and $F_0^*(\beta)$ becomes a quadrature approximation. We use a similar strategy with the proof of Lemma 1 where we approximate $f_0(b)$ with $\sum_{l=1}^{L}a_{f_0,l}\varphi_l(b)$ such that $\sup_{\beta\in\mathcal{B}}\left|f_0(b)-\sum_{l=1}^{L}a_{f_0,l}\varphi_l(b)\right|=O\left(L^{-\overline{s}/K}\right)$ due to Timan (1963). We find $F_0(\beta)-F_0^*(\beta)=O\left(R^{-\overline{s}/K}\right)$ a.e. $\beta\in\mathcal{B}$. Therefore, the conclusion follows from this and Theorem 5.

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] Bajari, P., J.T. Fox and S. Ryan (2007), "Linear Regression Estimation of Discrete Choice Models with Nonparametric Distributions of Random Coefficients", *American Economic Review*, 72, 2, 459–463.

[4] Baraud, Y. (2002), "Model Selection for Regression on a Random Design*", ESAIM Probability & Statistics 7*, 127-146.

[5] 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.

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

[7] 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.

[8] Chen, X. and D. Pouzo (2012), "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals", *Econometrica*, 80, 1, 277–321.

[9] Fox, J.T. and A. Gandhi (2015), "Nonparametric Identification and Estimation of Random Coefficients in Multinomial Choice Models", University of Michigan working paper.

[10] Fox J.T., K. Kim, S. Ryan, and P. Bajari (2011), "A Simple Estimator for the Distribution of Random Coefficients", *Quantitative Economics*, 2, 381-418.

[11] Fox, J.T., K. Kim, S. Ryan, and P. Bajari (2012), "The Random Coefficients Logit Model Is Identified", *Journal of Econometrics*, 166(2), 204-212.

[12] Geweke, J. and M. Keane (2007), "Smoothly mixing regressions", *Journal of Econometrics*, 138, 2007.

[13] Hastie, T, R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics.

[14] Heiss, F. and Winschel, V. (2008), "Likelihood approximation by numerical integration on sparse grids", *Journal of Econometrics*, 144(1), 62–80.

[15] Hoeffding, W. (1963), "Probability Inequalities for Sums of Independent Random Variables*", Journal of the American Statistical Association*, 58, 13-30.

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

[17] 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.

[18] Jacobs, R. , Jordan, M. I., Nowlan, S., and Hinton, G (1991), "Adaptive mixtures of local experts," *Neural Comput.*, 3(1), 79–87.

[19] Koenker, R. and I. Mizera (2014), "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules", Journal of the American Statistical Association, 109, 506, 674–685.

[20] Krommer, A.R. and C.W. Ueberhuber (1998), *Computation Integration*, SIAM.

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

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

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

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

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

[26] Petersen, B.E. (1983), *Introduction to the Fourier Transform and Pseudo-Differential Operators*, Pitman Publishing, Boston.

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

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

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

[30] Timan, A.F. (1963), *Theory of Approximation of Functions of a Real Variable*, MacMilan, New York.

[31] Train, K. (2008), "EM Algorithms for Nonparametric Estimation of Mixing Distributions", *Journal of Choice Modeling*, 1, 1, 40–69.

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

[33] Zeevi, A. J. and R. Meir (1997), "Density Estimation Through Convex Combinations of Densities: Approximation and Estimation Bounds," *Neural Network*s, 10(1), 99–109.