

# A Simple Nonparametric Estimator for the Distribution of Random Coefficients

Patrick Bajari

University of Minnesota & NBER

Jeremy T. Fox

University of Chicago & NBER

Kyoo il Kim

University of Minnesota

Stephen P. Ryan

MIT & NBER\*

May 2009

## Abstract

We propose a simple nonparametric mixtures estimator for recovering the joint distribution of parameter heterogeneity in economic models, such as the random coefficients logit. The estimator is based on linear regression subject to linear inequality constraints, and is robust, easy to program and computationally attractive compared to alternative estimators for mixture models. We prove consistency and provide the rate of convergence under deterministic and stochastic choices for the sieve approximating space. We present a Monte Carlo study and an empirical application to dynamic programming discrete choice with a serially-correlated unobserved state variable.

---

\*Fox thanks the National Science Foundation, the Olin Foundation, and the Stigler Center for generous funding. Thanks to helpful comments from seminar participants at the AEA meetings, Chicago, Chicago GSB, UC Davis, European Commission antitrust, Far East Econometric Society meetings, LSE, Mannheim, MIT, Northwestern, Paris I, Quantitative Marketing and Economics, Queens, Rochester, Rutgers, Stanford, Stony Brook, Toronto, UCL, USC, Virginia, and Yale. Thanks to comments from Xiaohong Chen, Andrew Chesher, Philippe Fevrier, Amit Gandhi, David Margolis, Andrés Musalem, Peter Reiss, Jean-Marc Robin, Andrés Santos, Azeem Shaikh and Harald Uhlig. Thanks to research assistance from Chenchuan Li.

# 1 Introduction

In economics, it is common to observe that otherwise identical agents behave differently when faced with identical choice environments, due to such factors as heterogeneity in preferences. A growing econometric literature has addressed this problem by providing estimators that allow the coefficients of the economic model to vary across agents (for recent work in demand estimation, see Berry, Levinsohn, and Pakes (1995), Nevo (2001), Petrin (2002), and Rossi, Allenby, and McCulloch (2005)). In this paper, we describe a general method for estimating such *random coefficient* models that is both nonparametric with respect to the underlying distribution of heterogeneity and easy to compute. Our estimator exploits a reparameterization of the underlying model so that the parameters enter linearly. This is in contrast to previous approaches in the literature, such as the EM algorithm, Markov Chain Monte Carlo (MCMC), simulated maximum likelihood and simulated method of moments, which are highly nonlinear. Linearity simplifies the computation of the parameters.

To motivate our approach, consider the following simple example. Suppose that the econometrician is interested in estimating a binary logit model with a scalar random coefficient. Let  $y = 1$  if the first option is chosen and let  $y = 0$  otherwise. Suppose that this model has a single independent variable,  $x$ . Furthermore, suppose that the random coefficient is known to have support on the  $[0,1]$  interval. Fix a large but finite grid of  $R$  equally spaced points. Suppose that the grid points take on the values  $\frac{1}{R}, \frac{2}{R}, \dots, \frac{R-1}{R}, 1$ . The parameters of our model are  $\theta^r$ , the frequencies of each random coefficient  $\frac{r}{R}$ . It then follows that the empirical probability that the dependent variable  $y$  is 1 conditional on  $x$  can be approximated by the linear combination

$$\Pr(y = 1 | x) \approx \sum_{r=1}^R \theta^r \frac{\exp\left(\frac{r}{R} \cdot x\right)}{1 + \exp\left(\frac{r}{R} \cdot x\right)}.$$

The key insight of our approach is that the dependent variable in our model,  $\Pr(y = 1 | x)$ , is linearly related to the model parameters  $\theta^r$ , irrespective of the nonlinear model used to compute the probability under a given type  $r$ . Instead of optimizing over that nonlinear model, we compute the probability under each type as if it were the true parameter, and then find the proper mixture of those models that best approximates the actual data. In the paper, we demonstrate that the  $\theta^r$  parameters can be consistently estimated using inequality constrained least squares. This estimator has a single, global optimum and widely-available specialized minimization approaches are guaranteed to converge to that point. This contrasts with alternative approaches

to estimating random coefficient models, where the objective functions can have multiple local optima and the econometrician is not guaranteed to find the global solution.

We also note that our approach does not require a parametric specification for the distribution of the random coefficients. We can closely approximate any well behaved distribution on  $[0,1]$  by making  $R$  large and by choosing  $\theta^r$  appropriately. Intuitively, our framework allows us to manipulate the cell size in a histogram. Many alternative estimators require computationally expensive nonlinear optimization, and as a result researchers frequently use tightly specified parametric models in applied work because of computational constraints. For example, applied researchers frequently assume that the random coefficients are mutually independent and normal. Our approach allows us to estimate the joint distribution of random coefficients without having to impose restrictions on the family of distributions.

We show how to extend this simple intuition to a more general framework. First, modeling the random coefficients using equally spaced grid points does not lead to a smooth estimated density. We suggest alternative methods for discretizing the model that give smooth densities, while still maintaining the linear relationship between the dependent variable and model parameters. Second, we discuss how our approach can be extended to more complex economic choice models. As an example, we show how to include a nonparametric distribution of random coefficients in a dynamic programming discrete choice model such as Rust (1987). In our application, we demonstrate that our approach can dramatically reduce the computational burden of these models. Third, we extend our approach to accommodate the case where the support of the basis functions is not known, and the econometrician must search over the parameter space.

Our approach can include random coefficients on multiple independent variables. Because of the curse of dimensionality in estimating infinite dimensional objects, the rate of convergence of the estimator of the distribution (with a deterministic choice of grid points) is inversely related to the number of random coefficients. With a stochastic choice of grid points, the rate of convergence is invariant to the number of random coefficients.

We do not claim that our estimator dominates all existing methods in all potential demand-estimation applications. In the paper, we discuss some strengths and weaknesses of our approach. First, if the researcher has aggregate data on market shares with price endogeneity, the approach of Berry, Levinsohn and Pakes (1995) is likely to be preferable. Our approach can accommodate aggregate data, even if market shares are measured with error. However, we need to specify a reduced-form pricing function as in Petrin and Train (2009) in order to account for price endogeneity. While this approach has worked well in some examples, it is not possible to prove the existence of this pricing function in the general case. Second, in small samples Bayesian

methods that use prior information will likely have better finite-sample performance (Rossi, Allenby and McCulloch 2005). Our methods are intended for applications where the researcher has access to large sample sizes.

Our approach is a general, nonparametric mixtures estimator. The most common frequentist, nonparametric estimator is nonparametric maximum likelihood or NPMLE (Laird 1978, Böhning 1982, Lindsay 1983, Heckman and Singer 1984). Often the EM algorithm is used for computation (Dempster, Laird and Rubin 1977), but this approach is not guaranteed to find the global maximum. The literature worries about the strong dependence of the output of the EM algorithm on initial starting values and well as the difficulty in diagnosing convergence (Verbeek, Vlassis and Kröse 2002). Li and Barron (2000) introduce another alternative, but again our approach is computationally simpler.<sup>1</sup> The discrete-grid idea (called the “histogram” approach) is found outside of economics in Kamakura (1991), who uses a discrete grid to estimate an ideal-point model. He does not discuss identification, the nonparametric statistical properties of his approach, or any of our extensions. Of course, mixtures themselves have a long history in economics (Quandt and Ramsey 1978).

We prove the consistency of our estimator for the distribution of random parameters, in function space and under a potential ill-posed inverse problem. We note that many if not most of the alternative estimators discussed above do not have general consistency theorems for the estimator for the distribution of random parameters.

The outline of our paper is as follows. In section 2, we start by precisely describing our model and estimator. We focus on the motivating example of discrete choice with individual data. However, we provide other examples such as discrete choice with aggregate data, dynamic discrete choice and mixed continuous and discrete choice (selection). In section 3, we review some results from the literature on nonparametric identification. In section 4, we apply results from Andrews (1999, 2002) to derive the asymptotic distribution of our estimator and discuss practical inference. In section 5, we discuss the case where the true distribution of types is unrestricted. Following Chen and Pouzo (2008), we view our estimator as a sieve estimation problem. We prove consistency in a function space if the true distribution is any arbitrary

---

<sup>1</sup>There is a literature on customized estimators for discrete-choice problems, as opposed to general mixtures estimators. Ichimura and Thompson (1998) and Gautier and Kitamura (2008) provide estimators for models with two choices and where covariates enter linearly. Lewbel (2000) considers the identification and estimation of mean preferences using a special regressor and a mean independence assumption. Manski’s (1975) maximum score estimator is consistent for mean preferences in the presence of random coefficients for the case of two choices only. Briesch, Chintagunta and Matzkin (2007) allow utility to be a nonparametric function of  $x$  and a scalar unobservable. Hoderlein, Klemelä and Mammen (2008) estimate the distribution of heterogeneity in a linear regression model.

distribution function. In section 6, under additional assumptions we derive a pointwise rate of convergence for our distribution estimator. In section 7, we conduct a Monte Carlo experiment to investigate the finite sample performance of our estimator. In section 8, we apply our estimator to a dynamic programming, discrete choice empirical problem studied in Dufló, Hanna and Ryan (2008). The dynamic programming problem has a serially correlated, unobserved state variable. We allow for a nonparametric distribution of random parameters.

## 2 The Estimator

### 2.1 General Notation

We first introduce general notation. The econometrician observes covariates  $x$  and the probability of some binary outcome,  $P(x)$ . For a model with a more complex outcome (including a continuous outcome  $y$ ), we can always consider whether some event ( $y \in A$  for some set  $A$ , say) happened or did not happen.  $P_A(x)$  is the probability of the event  $A$  happening.  $x$  is independent of  $\beta$ . Let  $g_A(x, \beta)$  be the probability of an agent with characteristics  $\beta$  taking an action in  $A$ . Our goal is to estimate the distribution function  $F(\beta)$  (and sometimes the corresponding density  $f(\beta)$ ) in the equation

$$P_A(x) = \int g_A(x, \beta) dF(\beta). \quad (1)$$

Identification means that a unique  $F(\beta)$  solves this equation for all  $x$  and all  $A$ .<sup>2</sup> In operator notation, let (1) be written as  $\mathcal{P} = L(F)$ , where  $\mathcal{P} = \{P_A(\cdot) \forall A\}$ . Identification means  $F = L^{-1}(\mathcal{P})$ . This paper focuses on estimating  $F$  using a finite sample of data. Because  $L^{-1}(\mathcal{P})$  may not be continuous in  $\mathcal{P}$  (the estimation problem may be ill-posed), we will adopt a sieve approach to estimating  $F$ .

### 2.2 The Logit Model

As we discussed in the introduction, the motivating example for our paper is the logit with random coefficients. In the logit, agents  $i = 1, \dots, N$  can choose between  $j = 1, \dots, J$  mutually exclusive alternatives. The exogenous variables for choice  $j$  are in the  $K \times 1$  vector  $x_{i,j}$ . In the example of demand estimation,  $x_{i,j}$  might include the product characteristics, the price of good

---

<sup>2</sup>This is the definition used in the statistics literature, see Teicher (1963).

$j$  and the demographics of agent  $i$ . We shall let  $x_i = (x'_{i,1}, \dots, x'_{i,J})$  denote the stacked vector of the all the  $x_{i,j}$ .

In the model, there are  $r = 1, \dots, R$  types of consumers. The preference parameters of type  $r$  are equal to  $\beta^r$ . We shall discuss how to choose the types  $\beta^r$  below. For the moment, assume that the  $\beta^r$  are fixed and exogenously specified. As in our example in the introduction, it might be helpful to think of the  $\beta^r$  being defined using a fixed grid on a compact set. Later we will view the  $R$  types of consumers as giving a nonparametric, sieve approximation to an arbitrary distribution for  $\beta$ ,  $F(\beta)$ . The random variable  $\beta$  is distributed independently of  $x$ . The probability of type  $r$  in the population is  $\theta^r$ . Let  $\theta = (\theta^1, \dots, \theta^R)$  denote the corresponding vector. This will be a parameter in our model. The  $\theta$  must lie on the simplex, or

$$\sum_{r=1}^R \theta^r = 1 \quad (2)$$

$$\theta^r \geq 0. \quad (3)$$

If agent  $i$  is of type  $r$ , her utility for choosing good  $j$  is equal to

$$u_{i,j} = x'_{i,j} \beta^r + \epsilon_{i,j}.$$

There is an outside good with utility  $\epsilon_{i,0}$ . Assume that  $\epsilon_{i,j}$  is distributed as Type I extreme value. Agents in the model are assumed to be utility maximizers. The dependent variable  $y_{i,j}$  is defined as

$$y_{i,j} = \begin{cases} 1 & \text{if } u_{i,j} > u_{i,j'} \text{ for all } j' \neq j \\ 0 & \text{otherwise.} \end{cases}$$

The probability of type  $r$  picking choice  $j$  at  $x_i$  is

$$g_j(x_i, \beta^r) = \frac{\exp(x_{i,j} \beta^r)}{1 + \sum_{j'=1}^J \exp(x_{i,j'} \beta^r)}.$$

Because the error term is extreme value, it follows that

$$\Pr(y_{i,j} = 1 \mid x_i) = \sum_{r=1}^R \theta^r g_j(x_i, \beta^r) = \sum_{r=1}^R \theta^r \frac{\exp(x_{i,j} \beta^r)}{1 + \sum_{j'=1}^J \exp(x_{i,j'} \beta^r)}. \quad (4)$$

## 2.3 Linear Regression

A first method for estimating the parameters  $\theta$  is by ordinary least squares. To construct the estimator, begin by adding  $y_{i,j}$  to both sides of (4) and moving  $\Pr(y_{i,j} = 1 \mid x_i)$  to the right side of this equation, which gives

$$y_{i,j} = \left( \sum_{r=1}^R \theta^r g_j(x_i, \beta^r) \right) + (y_{i,j} - \Pr(y_{i,j} = 1 \mid x_i)).$$

Define the  $R \times 1$  vector  $z_{i,j} = (z_{i,j,1}, \dots, z_{i,j,R})'$  with individual elements  $z_{i,j,r} = g_j(x_i, \beta^r)$  and let  $z_i = (z_{i,1}, \dots, z_{i,J})$ . Recall that  $\beta^r$  is fixed and is not a parameter to be estimated. As a result, given  $x_i$ , the term  $z_{i,j,r}$  is a known constant. Next, by the definition of a choice probability,

$$E[(y_{i,j} - \Pr(y_{i,j} = 1 \mid x_i)) \mid z_i] = 0. \quad (5)$$

This implies that the following ordinary least squares problem is a consistent estimator of the  $\theta^r$ ,

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J (y_{i,j} - z'_{i,j} \theta)^2.$$

Let  $Y$  denote the  $NJ \times 1$  vector formed by stacking the  $y_{i,j}$  and let  $Z$  be the  $NJ \times R$  matrix formed by stacking the  $z_{i,j}$ . Then our estimator is  $\hat{\theta} = (Z'Z)^{-1} Z'Y$ .<sup>3</sup>

(5) implies that the standard exogeneity condition is satisfied for least squares. The “error term” in our least squares regression is  $y_{i,j} - \Pr(y_{i,j} = 1 \mid x_i)$ . This is the difference between the  $y_{i,j}$  and the probability that  $y_{i,j} = 1$ . This prediction error only depends on the realization of the  $\varepsilon_{i,j}$ 's, which are iid and as a consequence independent of  $z_i$ . The least squares estimator will have a unique solution so long as  $Z$  has rank  $R$ .

---

<sup>3</sup>We focus on the least squares criterion, rather than a likelihood or pseudo-likelihood, for computational simplicity. However, it is not necessarily the case that the least squares criterion will result in a less efficient estimator. The least squares objective function enforces all of the parametric assumptions inherent in the logit model, which arises in the definition of  $z_{i,j,r}$ . There is no sense in which the likelihood alternative uses more structure. Second, in a nonparametric world, the true distribution lies in an infinite-dimensional space, and the limiting distribution of any estimator for the infinite-dimensional object is not known. Therefore, we avoid discussing efficiency, which relies on the estimator having an asymptotically normal distribution.

## 2.4 Inequality Constrained Least Squares

A limitation of the ordinary least squares estimator is that  $\hat{\theta}$  need not satisfy (2) and (3). In practice, one might wish to constrain  $\hat{\theta}$  to be a well-defined probability measure. This would be useful in making sure that our model predicts probabilities that always lie between zero and one. Also, this may be important if the economist wishes to interpret the distribution of  $\beta$  as a structural parameter. We suggest estimating  $\theta$  using inequality constrained least squares,

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J (y_{i,j} - z'_{i,j} \theta)^2 \quad (6)$$

subject to (2) and (3).

This minimization problem is a quadratic programming problem subject to linear inequality constraints.<sup>4</sup> The minimization problem is convex and routines like MATLAB's `lsqin` guarantee finding a global optimum. One can construct the estimated cumulative distribution function for the random coefficients as

$$\hat{F}(\beta) = \sum_{r=1}^R \hat{\theta}^r 1[\beta^r \leq \beta],$$

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

## 2.5 Smooth Basis Densities

A limitation of the method above is that the CDF of the random parameters will be a step function. For human visualization, it is attractive to have a smooth distribution of random parameters. In this subsection, we describe one approach to estimating a density instead of a CDF. Our approach is easily extended to this case: instead of modeling the distribution of the random parameters as a mixture of point masses, we instead model the density as a mixture of normal densities.

Let a basis  $r$  be a normal distribution with mean the  $K \times 1$  vector  $\mu^r$  and standard deviation

---

<sup>4</sup>Lindsay (1983) discusses how if there are at most  $M$  distinct values of the  $N$  observations  $(x_i, y_i)$ , then the nonparametric maximum likelihood estimator can maximize the likelihood using at most  $M$  points of support for the beta space,  $\beta$ . Adding more points of support would not increase the statistical fit given the finite sample. Here, we fix the points of support and estimate only the weights, so we do not appeal to Lindsay's results. In our simulations and empirical work, we often find that only a few of the many points of support have nonzero weights in the final estimates.

the  $K \times 1$  vector  $\sigma^r$ . Let  $N(\beta_k | \mu_k^r, \sigma_k^r)$  denote the normal density of the  $k$ th random parameter. Under normal basis functions, the joint density for a given  $r$  is just the product of the marginals, or

$$N(\beta | \mu^r, \sigma^r) = \prod_k N(\beta_k | \mu_k^r, \sigma_k^r).$$

Let  $\theta^r$  denote the probability mass that  $\beta$  is drawn from  $N(\beta | \mu^r, \sigma^r)$ . As in the previous subsection, it is desirable to constrain  $\theta$  to lie in the simplex, that is, (2) and (3).

For a given  $r$ , make  $S$  simulation draws from  $N(\beta | \mu^r, \sigma^r)$ . Let a particular draw  $s$  be denoted as  $\beta^{r,s}$ . We can then simulate  $\Pr(y_{i,j} = 1 | x_i)$  as:

$$\Pr(y_{i,j} = 1 | x_i) \approx \sum_{r=1}^R \theta^r \left( \frac{1}{S} \sum_{s=1}^S g_j(x_i, \beta^{r,s}) \right) = \sum_{r=1}^R \theta^r \left( \frac{1}{S} \sum_{s=1}^S \frac{\exp(x_{i,j} \beta^{r,s})}{1 + \sum_{j'=1}^J \exp(x_{i,j'} \beta^{r,s})} \right).$$

Other numerical-integration approaches such as quadrature can be used in place of simulation (Judd 1998).

We can then estimate  $\theta$  using the inequality constrained least squares problem

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J \left( y_{i,j} - \sum_{r=1}^R \theta^r \left( \frac{1}{S} \sum_{s=1}^S g_j(x_i, \beta^{r,s}) \right) \right)^2$$

subject to (2) and (3).

This is once again inequality constrained least squares, a globally convex optimization problem.<sup>5</sup> The resulting density estimate is

$$\hat{f}(\beta) = \sum_{r=1}^R \hat{\theta}^r N(\beta | \mu^r, \sigma^r).$$

## 2.6 Location Scale Model

In many applications, the econometrician may not have good prior knowledge about the support region where most of the random coefficients  $\beta^r$  lie. Return to the discrete grid estimator. Let the unscaled basis vectors  $\{\beta^r\}_{r=1}^R$  lie in the set  $[0, 1]^K$ , that is, the  $K$ -fold Cartesian product of

---

<sup>5</sup>If the true distribution has a continuous density function, using an estimator that imposes that the density estimate is smooth may provide better statistical performance than an estimator that imposes no restrictions on the true distribution function. We do not formally derive the rate of convergence of the estimator for densities or compare the rate to the rate for the earlier estimator for CDFs, which is derived in Section 6.

the unit interval. We include a set of location and scale parameters  $\mu_k$  and  $\sigma_k$ ,  $k = 1, \dots, K$  and define the  $r$ th random coefficient for the  $k$ th characteristic as  $\mu_k + \sigma_k \beta_k^r$ .

In numerical optimization, we now search over  $R + 2K$  parameters corresponding to  $\theta$  and  $\mu = (\mu_1, \dots, \mu_K)'$  and  $\sigma = (\sigma_1, \dots, \sigma_K)'$ . Market shares predictions for type  $r$  are

$$g_j(x_i, \mu + \sigma \beta^r) = \frac{\exp\left(\sum_{k=1}^K x_{k,i,j} (\mu_k + \sigma_k \beta_k^r)\right)}{1 + \sum_{j'=1}^J \exp\left(\sum_{k=1}^K x_{k,i,j'} (\mu_k + \sigma_k \beta_k^r)\right)},$$

where  $\sigma \beta^r$  is assumed to represent element-by-element multiplication. Our estimator for the weights solves the nonlinear least squares problem

$$\begin{aligned} \min_{\mu, \sigma, \theta} \frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J \left( y_{i,j} - \sum_{r=1}^R \theta^r g_j(x_i, \mu + \sigma \beta^r) \right)^2 \\ \text{subject to (2) and (3)}. \end{aligned} \tag{7}$$

The appropriate MATLAB routine is `lsqnonlin`, and there is no theorem that the routine will converge to a global minimum. However, it is worth noting that the derivatives to our objective function can be expressed in closed form up to any order, at least for the logit example. Note that the model has  $2K$  nonlinear parameters and the remaining  $R$  parameters still enter linearly. Also,  $\mu$  and  $\sigma$  do not enter the constraints, which are still linear. Using exact, rather than numerical derivatives will improve the speed and convergence of the estimator.<sup>6</sup> In numerical experimentation with the logit example, we have found the convergence of the estimator to be robust.

In many applied problems, the number of characteristics  $K$  may be large and it may not be practical to estimate a  $K$ -dimensional nonparametric distribution of random coefficients, and therefore it may be desirable to only let a subset of the most important characteristics have random coefficients. This is possible by a trivial modification of (7).

## 2.7 Choosing $R$ and the $\beta^r$ 's

We have discussed the location and scale model for when the support of  $\beta^r$  is not known and the  $\beta^r$  are picked on a grid. The  $\beta^r$  in this case can be chosen by using a random number generator

---

<sup>6</sup>Automatic differentiation is an option to compute derivatives.

to make draws from  $[0, 1]^K$ . We have also experimented with Halton and Weyl sequences; they tend to improve Monte Carlo performance over evenly-spaced grids. The rate of convergence derived in Section 6 uses results from the theory of quadrature. Another rate of convergence, derived in Section 6.1, relies on choosing a grid using random draws from a chosen distribution.

It is also natural to ask how to choose  $R$ , the number of basis components. The nonparametric rates of convergence in section 6 are not useful for choosing  $R$  in a finite sample. However, the rates do suggest that  $R$  should not increase too quickly with  $N$ , the number of observations. In Monte Carlo experiments in section 7, we set  $R = \frac{N}{40}$ . Our Monte Carlos show a small  $R$  (say a few hundred basis vectors) can give a good approximation to  $F(\beta)$  in a finite sample.

A researcher may prefer an iterative procedure where he or she chooses  $R$  grid points, discards those with zero points, and chooses new grid points near the apparent areas of the support of  $\beta$  with more mass. As statistical sampling error enters the first set of estimates, this iterative procedure introduces pre-testing and will alter the asymptotic distribution.

## 2.8 Aggregate Data

Our estimator can still be used if the researcher has access to data on market shares  $s_{j,t}$ , rather than individual level choice data  $y_{i,j}$ . Index markets by  $t = 1, \dots, T$ . In this framework, we assume that the utility of person  $i$  when the type of  $i$  is  $r$  is

$$u_{i,j} = x'_{j,t} \beta^r + \epsilon_{i,j}.$$

In this model, the utility of individuals is a function of product- and market-varying attributes  $x_{j,t}$ . In applied work, even if product characteristics are fixed, prices will vary across markets in the sample. If a market has a continuum of consumers, our modeling framework implies that the market share of product  $j$  should satisfy

$$s_{j,t} = \sum_{r=1}^R \theta^r g_j(x_t, \beta^r) = \sum_{r=1}^R \theta^r \frac{\exp(x_{j,t} \beta^r)}{1 + \sum_{j'=1}^J \exp(x_{j',t} \beta^r)}$$

where we let  $x_t = (x'_{1,t}, \dots, x'_{J,t})$  denote the stacked vector of the all the  $x_{j,t}$ . Suppose that the economist only observes a noisy measure  $\widehat{s}_{j,t}$  of the true share. This is common in applied work. For example, there may be a finite number of consumers in a market. Let the actual share be

denoted as  $\widehat{s}_{j,t}$ . Simple algebra implies that

$$\widehat{s}_{j,t} = \left( \sum_{r=1}^R \theta^r g_j(x_t, \beta^r) \right) + (\widehat{s}_{j,t} - s_{j,t}).$$

Under standard assumptions,  $E[\widehat{s}_{j,t} - s_{j,t} | x_t] = 0$ . That is, the difference between the measured shares and the true shares is independent of the product characteristics in our model  $x_t$ . This would be the case if the difference between  $\widehat{s}_{j,t}$  and  $s_{j,t}$  is accounted for by random sampling. Then we can estimate  $\theta$  using the regression

$$\widehat{\theta} = \arg \min_{\theta} \frac{1}{JT} \sum_{t=1}^T \sum_{j=1}^J \left( \widehat{s}_{j,t} - \left( \sum_{r=1}^R \theta^r g_j(x_t, \beta^r) \right) \right)^2$$

subject to (2) and (3).

One advantage of our approach is that it can accommodate measurement error in the market shares. We note that the method of Berry, Levinsohn and Pakes (1995) assumes that  $s_{j,t}$  is observed without error by the economist. If the number of persons in a survey or in a market is small, this assumption may be a poor approximation.

A drawback of our approach is that it may require less attractive assumptions than Berry, Levinsohn and Pakes (1995) to account for price endogeneity and aggregate demand shocks. One could use the control function approach of Petrin and Train (2009) to account for this problem. Petrin and Train argue that the control function works well in the applied examples that they consider. However, it has two potential drawbacks. The first is that this approach assumes that there is a reduced-form pricing equation that permits the econometrician to infer the omitted product attributes from the observed prices. In general, it is unknown whether this reduced form exists. The second limitation is that the control function approach requires inserting a generated regressor that depends on a nonparametric function into our estimator. This complication is beyond the scope of our asymptotic theory.

## 2.9 Dynamic Programming Models

In this section, we demonstrate that our approach can be applied to dynamic discrete choice models as in Rust (1987, 1994). We generalize the framework he considers by allowing for a nonparametric distribution of random coefficients. Suppose that the flow utility of agent  $i$  in a

period  $t$  from choosing action  $j$  is

$$u_{i,j,t} = x'_{i,j,t}\beta_i + \epsilon_{i,j,t}.$$

The error term  $\epsilon_{i,j,t}$  is a preference shock for agent  $i$ 's utility to choice  $j$  at time period  $t$ . The error term is iid extreme value across agents, choices and time periods. For simplicity, make  $\epsilon_{i,j,t}$  logit. Agent  $i$ 's decision problem is dynamic because there is a link between the current and future values of  $x_{i,t} = (x_{i,1,t}, \dots, x_{i,J,t})$  through current decisions. Let  $\pi(x_{i,t+1} | x_{i,t}, j)$  denote the transition probability for the state variable  $x_{i,t}$ . This does not involve random coefficients and we assume that it can be estimated in a first stage.

The goal is to estimate  $F(\beta)$ , the distribution of the random coefficients. Again we draw  $R$  basis vectors  $\beta^r$ . For each of the  $R$  basis vectors, we can solve the corresponding single-agent dynamic programming problem for the state  $x_{i,t}$  value functions,  $V^r(x_{i,t})$ . Once all value functions  $V^r(x_{i,t})$  are known, the choice probabilities  $g_j(x_i, \beta^r)$  for all combinations of choices  $j$  and states  $x_{i,t}$  can be calculated as

$$g_j(x_{i,t}, \beta^r) = \frac{\exp(x'_{i,j,t}\beta^r + \delta E[V^r(x_{i,t+1}) | x_{i,t}, j])}{\sum_{j'=1}^J \exp(x'_{i,j',t}\beta^r + \delta E[V^r(x_{i,t+1}) | x_{i,t}, j'])} = \frac{\exp(v^r(j, x_{i,t}))}{\sum_{j'=1}^J \exp(v^r(j', x_{i,t}))},$$

where  $v^r(j, x_{i,t})$  is Rust's choice-specific continuation value, here implicitly defined in the above equation.

We use panel data on  $N$  panels of length  $T$  each. Suppose that we observe  $y_{i,j,t}$ , an indicator variable equal to one if  $j$  is chosen at time  $t$  by agent  $i$ . Let  $x_{i,t}$  be the state of agent  $i$  at time  $t$ . We can then estimate

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{NJT} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^J \left( y_{i,j,t} - \sum_{r=1}^R \theta^r g_j(x_{i,t}, \beta^r) \right)^2$$

subject to (2) and (3).

As in the earlier examples, we could use normal distributions as basis functions in order to smooth the estimates of the random coefficients. However, it is not desirable to use the location scale model since modifying these parameters would require us to re-solve the model. The idea of presolving a complex economic model for only  $R$  types before a nonlinear (in our case quadratic) search commences is also found in Akerberg (2009), although Akerberg's approach is parametric

instead of nonparametric on the distribution of random coefficients.

It is often the case that the information on a panel can provide more information on heterogeneity than  $T$  repeated cross sections. We can explicitly incorporate this use of panel data into our estimator. Let  $w$  index a sequence of choices for each time period  $t = 1, \dots, T$  called  $w_1, \dots, w_T$ . For example, a choice sequence  $w$  could be  $w_1 = 5, w_2 = 2, w_3 = 3, \dots$ . If there are  $J$  choices per period, there are  $W = J^T$  sequences that could occur. Let  $y_{i,w}$  be equal to 1 if agent  $i$  takes action sequence  $w$  over the  $T$  periods. This minimization problem when panel data is used for extra information on heterogeneity is

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{NW} \sum_{i=1}^N \sum_{w=1}^W \left( y_{i,w} - \prod_{t=1}^T \left( \sum_{r=1}^R \theta^r g_{w_t}(x_{i,t}, \beta^r) \right) \right)^2$$

subject to (2) and (3). (8)

In this case, the estimator matches sequences of choices.

## 2.10 Joint Discrete and Continuous Demand

Other economic models can be estimated using our methods. Say a consumer purchases a particular type of air conditioner according to the logit model. Conditional on purchase, we observe the electricity consumption of the air conditioner, a measure of usage. The electricity usage equation of consumer  $i$  of type  $r$  for air conditioner  $j$  is

$$a_{i,j} = w'_{i,j} \gamma_j^r + \eta_j^r.$$

Let  $w_i = (w_{i,1}, \dots, w_{i,J})$ ,  $\gamma = (\gamma_1, \dots, \gamma_J)$  and  $\eta = (\eta_1, \dots, \eta_J)$ . Let  $y_{i,j}^l$  be 1 when consumer  $i$  purchases air conditioner  $j$  and consumes electricity between the lower and upper bounds  $a_j^l$  and  $a_j^{l+1}$ . Then

$$\Pr \left( y_{i,j}^l = 1 \mid x_i, w_i; \beta^r, \gamma^r, \eta^r \right) = g_{j,l}(x_i, w_i, \beta^r, \gamma^r, \eta^r) = \frac{\exp(x_{i,j} \beta^r)}{1 + \sum_{j'=1}^J \exp(x_{i,j'} \beta^r)} \mathbf{1} \left[ a_j^l \leq w'_{i,j} \gamma_j^r + \eta_{i,j}^r < a_j^{l+1} \right].$$

The unknown object of interest is the joint distribution  $F(\beta, \gamma, \eta)$  of the multinomial choice parameters  $\beta$ , the electricity usage parameters  $\gamma$ , and the additive errors in utility usage,  $\eta$ . In this case one can choose a grid of taste parameters  $\{\beta^r, \gamma^r, \eta^r\}_{r=1}^R$ . The researcher must also discretize

the continuous outcome variable  $a_j$  by choosing  $L$  bins  $[a_j^0, a_j^1]$ ,  $[a_j^1, a_j^2]$ , through  $[a_j^{L-1}, a_j^L]$ . A higher  $L$  increases the computational burden and the closeness of the approximation to the continuous outcome model.

This joint continuous and discrete demand model is an example of a selection model. A statistical observation is  $(j_i, a_{i,j}, x_i, w_i)$ , which can be transformed into  $\left(\left\{y_{i,j}^l\right\}_{1 \leq j \leq J, 0 \leq l < L}, x_i, w_i\right)$  by a change of variables. The outcome  $a_{i,j}$  is selected because the researcher only observes the electricity usage for air conditioner  $j$  for those individuals who purchases air conditioner  $j$ , or  $j_i = j$ . Regressing  $a_{i,j}$  on  $w_{i,j}$  for those individuals who choose  $j$  will not consistently estimate  $E[\gamma_j]$  if  $\gamma_j$  is not statistically independent of  $\beta$ . Those agents who choose air conditioner  $j$  will have certain  $\beta$  preferences, and the correlation of  $\beta$  with  $\gamma$  and the correlation of  $x$  with  $w$  will induce correlation between  $w$  and  $\gamma$ . However, jointly modeling both the choice of air conditioner and electricity usage removes this selection problem. Note that we allow random coefficients in both the selection (multinomial choice) and outcome (usage) parts of the model.

The estimator is

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{NJJL} \sum_{i=1}^N \sum_{j=1}^J \sum_{l=1}^L \left( y_{i,j}^l - \sum_{r=1}^R \theta^r g_{j,l}(x_i, w_i, \beta^r, \gamma^r, \eta^r) \right)^2$$

subject to (2) and (3).

The estimator of the joint distribution of the random parameters is

$$\hat{F}(\beta, \gamma, \eta) = \sum_{r=1}^R \hat{\theta}^r \mathbf{1}[\beta^r \leq \beta, \gamma^r \leq \gamma, \eta^r \leq \eta].$$

Fox and Gandhi (2009b) discuss the identification of selection models, such as a discrete-continuous model of demand.

## 2.11 Endogenous Regressors

In some cases, the regressors may be correlated with omitted factors. Consider the multinomial choice model. Fox and Gandhi (2009a) address the correlation of  $x_{i,j}$  with  $\beta$  using instrumental variables and an auxiliary equation, in which the values of the endogenous regressors are given as a function of exogenous regressors and instruments. There can be random coefficients in the auxiliary equation. Berry and Haile (2008) investigate a multinomial choice model where the utility to each choice  $j$  includes an additive  $\xi_{j,t}$  term, perhaps reflecting an unobserved product

characteristic of product  $j$  in market  $t$ . By exploiting individual-level characteristic variation within a market and instruments that vary across markets, they are able to identify the values of  $\xi_{j,t}$ , allowing them to treat  $\xi_{j,t}$  as an observable.

The estimator for the distribution of heterogeneity in this paper can be an ingredient into the identification strategies of Fox and Ghandi (2009a) and Berry and Haile (2008). See those papers for more details.

### 3 Identification

Before one estimates a particular model's distribution of random coefficients nonparametrically, one must theoretically verify that the distribution of random coefficients is identified. Identification will be required for a complete proof of the consistency of any nonparametric estimator, including ours. In this section, we review results on the problem of nonparametrically identifying the density of random coefficients  $f(\beta)$  or its distribution,  $F(\beta)$ . We use the general notation of Section 2.1.

#### 3.1 Identification using Linear Independence / Bounded Completeness

A common identification strategy is to show that the space of allowable models,  $\{g_A(x, \beta) \mid \beta \in \mathcal{B}\}$ , is linearly independent.

**Assumption 3.1** *The family of measurable functions  $\{g_A(x, \beta) \mid \beta \in \mathcal{B}\}$  in both  $x \in \mathcal{X}$  and  $\beta \in \mathcal{B}$  is linearly independent (or bounded complete) with respect to continuous functions if when  $\int g_A(x, \beta) h(\beta) d\beta = 0$  for all  $x \in \mathcal{X}$  and  $h(\cdot)$  is bounded measurable, then  $h(\beta) = 0$ .*

**Theorem 3.1** *Suppose Assumption 3.1 holds. Then the true  $f^0$  in the class of all continuous mixtures that admit a density and are bounded measurable,*

$$\left\{ \int g_A(x, \beta) f(\beta) d\beta \mid C \geq f(\beta) \geq 0, \int f(\beta) d\beta = 1 \right\},$$

*is identified.*

**Proof.** Identification is equivalent to the nonexistence of any function  $\delta(\beta) = \tilde{f}(\beta) - f(\beta) \neq 0$  such that  $\int g_A(x, \beta) \delta(\beta) d\beta = 0$  for all  $x \in \mathcal{X}$ , where  $\mathcal{X}$  is the support of  $x$ . Therefore, identification immediately follows by Assumption 3.1 and the fact that  $f(\beta)$  belongs to a class of uniformly bounded density functions. ■

That linear independence implies identification is nearly tautological, as the two definitions are similar.<sup>7</sup> However, the condition may be useful to the extent linear independence or bounded completeness has been shown for various classes of functions in the literature. When  $g_A(x, \beta)$  belongs to the exponential family of distributions, completeness is proved in Lehmann and Romano (2005, p. 117). Completeness implies bounded completeness. There are a few other classes of functions that are known to be bounded complete. Any location family of absolutely continuous distributions with compact supports is bounded complete. It is possible to use results on distribution functions for the identification of economic choice models by considering the transformed choice model

$$\tilde{g}_A(x, \beta) = \frac{g_A(x, \beta)}{\int g_A(x, \beta) dx},$$

which is a density that integrates to 1 over the space of  $x$ .

Teicher (1963) and Yakowitz and Spragins (1968) were the first to explore a related linear independence argument for the class of finite mixtures where the mass points with positive support are not known before identification. Verifying these conditions for particular economic models is difficult. Fox and Gandhi (2009a) introduce an intermediate condition for identification that is expressed in terms of the properties of an economic choice model. They show the identification of models where, for example,  $g_A(x, \beta)$  is an indicator for making multinomial choice  $j \in A = \{1, \dots, J\}$  or having a continuous outcome  $y \in A \subseteq \mathbb{R}$  less than some threshold. In other words, they do not consider the logit case with parametric distributions for  $\epsilon_{i,j}$ ; all heterogeneity is embedded in the random coefficients or, even, random nonparametric functions. Fox and Gandhi (2009b) prove identification of selection and mixed continuous-discrete models without relying on identification at infinity. Our mixtures estimator could be applied to these selection and mixed continuous-discrete models as well.

### 3.2 Other Approaches to Identification

Our leading example is the random coefficients logit. In Bajari, Fox, Kim and Ryan (2009), we develop a calculus-based, constructive identification theorem for the density of random coefficients in the logit model. Thus, the random coefficients logit model is identified. Our calculus-based proof identifies all the moments of the random coefficients. This technique can be applied to other differentiable choice models as well. Combined with the coming consistency theorem in this

---

<sup>7</sup>A distinct definition of bounded completeness is often referred to in the nonparametric instrumental variables literature. For an IV setting, d'Haultfoeuille (2009) derives bounded completeness using primitive economic models.

paper, our identification theorem for the logit model provides a complete proof of consistency of our estimator than does not relying on assuming that the logit model is identified.

Ichimura and Thompson (1997) study binary choice and use the Cramér-Wold (1936) theorem for identification. Gautier and Kitamura (2008) present alternative arguments for the same set of results. Hoderlein, Klemelä, and Mammen (2008) study the identification and estimation of the density of random coefficients in the liner regression model.

For a multinomial choice model, Berry and Haile (2008) use a regressor that enters additively separably with a known sign to trace out the CDF of the total utility values conditional on other covariates. Identifying the total utility values conditional on  $x$ , rather than the distribution of random utility functions of  $x$  or the distribution of random coefficients, prevents welfare analysis. For welfare analysis, one needs to know, say, the willingness to pay for each “type” of consumer and the two budget sets. Briesch, Chintagunta and Matzkin (2008) study identification in a model of multinomial choice where the subutility function for each choice has a nonparametric function of observables and a scalar unobservable.

## 4 Estimating Confidence Regions with Finite Types

The literature on sieve estimation (Chen and Pouzo 2008) does not have asymptotic distribution theory for the infinite-dimensional unknown parameter,  $F(\beta)$ . Given the state of the literature, we focus on computing standard errors using a parametric method, which requires the assumption that the set of  $R$  types used in estimation is the true set of  $R$  types that generates the data. Under this assumption, one can use the common OLS heteroskedasticity-consistent standard errors for the unknown weights,  $\theta^1, \dots, \theta^R$ .

In empirical work, we often find that many of the included  $R$  types have estimated weights of  $\hat{\theta}^r = 0$ . Thus, one can construct smaller confidence intervals by recognizing that the parameters on the boundary of the parameter space cannot have an asymptotically normal distribution.<sup>8</sup> Let  $\hat{\theta}$  be the vector of estimated weights and  $\theta_0$  the vector of true weights in the data generating process. Following Andrews (1999, 2002), the asymptotic distribution of  $\hat{\theta}^r - \theta_0^r$  will take on only nonnegative values when  $\theta_0^r = 0$  for some  $r \in R$ .

We recommend estimation using inequality-constrained least squares and then computing standard errors using the usual formulas, with  $\hat{\theta}$  used to construct the estimated residuals. Andrews and Guggenberger (2009b) study the case of a regression with one inequality constraint

---

<sup>8</sup>Statistical inference for linear regression subject to a set of inequality constraints has been studied by Judge and Takayama (1966), Liew (1976), Geweke (1986), and Wolak (1987).

and i.i.d. observations.<sup>9</sup> They show the traditional OLS confidence intervals, using a fixed critical value such as 1.96 for a two-tailed, 95% confidence interval, have what they define to be correct asymptotic coverage under some nonstandard asymptotics designed to better approximate the problem of a parameter on the boundary. Our estimator is more complex. One needs to use heteroskedasticity-consistent standard errors because the errors in a linear probability model such as (6) are heteroskedastic. One also should cluster the standard errors at the level of the “regression observations”  $j = 1, \dots, J$  for each statistical observation  $i = 1, \dots, N$ . Recall that for each  $i$ , there are  $J$  “regression observations” in (6) because each inside good  $j$  has a separate term in the least-squares objective function. After constructing the OLS confidence intervals for  $\hat{\theta}^r$ , one should remove infeasible values by reporting, for a 95%, two-sided confidence interval,

$$[0, 1] \cap \left[ \hat{\theta}^r - 1.96 \cdot \text{SE} \left( \hat{\theta}^r \right), \hat{\theta}^r + 1.96 \cdot \text{SE} \left( \hat{\theta}^r \right) \right],$$

where  $\text{SE} \left( \hat{\theta}^r \right)$  is the standard error adjusted for heteroskedasticity and clustered across the  $J$  “regression observations” for each statistical observation  $i$ .

Researchers are often not directly interested in confidence intervals for the weights  $\theta$  but rather for functions  $m(\theta; X)$ , where  $X$  is some arbitrary data. For example, researchers may wish to construct confidence intervals for the distribution  $\hat{F}(\beta) = m(\hat{\theta}; X) = \sum_{r=1}^R \hat{\theta}^r 1[\beta^r \leq \beta]$  evaluated at a particular value of  $\beta$  ( $X$  does not enter  $m$  here). In our empirical example below, we construct confidence intervals for out-of-sample predictions. To construct standard errors for  $m(\hat{\theta}; X)$ , one first constructs the distribution of  $\hat{\theta}^r - \theta_0^r$  as above and then uses the delta method. A 95% confidence interval is then

$$\left[ \min_{\theta} m(\theta; X), \max_{\theta} m(\theta; X) \right] \cap \left[ m(\hat{\theta}; X) - 1.96 \cdot \text{SE} \left( m(\hat{\theta}; X) \right), m(\hat{\theta}; X) + 1.96 \cdot \text{SE} \left( m(\hat{\theta}; X) \right) \right].$$

Here the minimum and maximum are taking over the values of  $\theta$  that satisfy (2) and (3). This is a compact set, so the minimum and maximum are obtained. In many examples, it will be possible to deduce the feasible upper and lower bounds for  $m(\theta; X)$  without resorting to computation.

---

<sup>9</sup>A parameter on the boundary is challenging for inference because the limit distribution of the estimator is discontinuous at the parameter value on the boundary. Andrews and Guggenberger (2009b, footnote 7) argue that the traditional OLS confidence intervals with the unconstrained (not imposing (2) and (3)) point estimates have correct coverage, even if they are conservative. The limit distribution of the unconstrained estimator is not discontinuous in the value of the true parameter.

Resampling procedures are a possibility, but one that Andrews and Guggenberger do not currently recommend when a true parameter may lie on a boundary. Andrews (2000) shows that the standard bootstrap is inconsistent but that subsampling and the  $m$ -out-of- $n$  bootstrap are pointwise consistent. Andrews and Guggenberger (2010) show that the latter two resampling procedures are not uniformly consistent and may have poor finite-sample coverage. Andrews and Guggenberger (2009a, 2009b) discuss a hybrid method where the maximum of a traditional critical value for a  $t$ -statistic and a subsampled critical value is used to construct confidence intervals for  $\theta$ . The hybrid subsampling method has correct asymptotic coverage under the definition of Andrews and Guggenberger, but, as Andrews and Guggenberger (2009b) show, so does one of its ingredients, the traditional fixed critical value method that Andrews and Guggenberger recommend.

In an online appendix, we show consistency and derive the sampling distribution of the inequality-constrained nonlinear least squares estimator of the location and scale model. The nonlinear distribution also applies if some parameters in the model are homogeneous. As the sampling distribution is derived, we have verified the only regularity condition needed for the pointwise consistency of subsampling (Politis, Romano and Wolf 1999).

## 5 Consistency for the Distribution in Function Space

In this section, we assume the true distribution function  $F$  lies in the space  $\mathcal{F}$  of all distribution functions on a parameter space  $\mathcal{B}$ . Let  $R(N)$  be the number of grid points chosen with  $N$  observations. 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  $F_0 \in \mathcal{F}$ . To do this, we use the recent results for sieve estimators developed by Chen and Pouzo (2008), hereafter referred to as CP. We view each set of  $R$  points as a sieve space

$$\mathcal{F}_R = \left\{ F \mid \exists \theta^1, \dots, \theta^R, \theta^r \geq 0, \sum_{r=1}^R \theta^r = 1 \text{ s.t. } F(\beta) = \sum_{r=1}^R \theta^r 1[\beta^r \leq \beta] \right\},$$

for some fixed choice of grid  $\mathcal{B}^R = \{\beta^1, \dots, \beta^R\}$ . Note that this sieve space automatically imposes that the estimated function is a distribution function. CP suggest other bases for unknown functions; we prefer our choice of sieve spaces in the interests of computational simplicity and the ease of constraining the estimated function to be a distribution function.

CP develop conditions under which one can prove that the estimator  $\hat{F}_N$  converges to the true  $F_0$  in function space, i.e. in the space of distributions. CP study estimators defined by a two-

stage procedure. Let  $y$  be the dependent variable, which represents some discrete outcome. We suppress the notation  $A$  for the outcome. In our version of CP's first stage, one uses observations on  $(y_i, x_i)_{i=1}^N$  and a nonparametric least-squares procedure to estimate  $\hat{P}(x)$ . In the demand context, the researcher estimates choice probabilities by smoothing across agents with similar  $x$ 's. In a second stage, the researcher minimizes the objective function

$$\frac{1}{N} \sum_{i=1}^N \left( \hat{P}(x_i) - \sum_{r=1}^R \theta^r g(x_i, \beta^r) \right)^2$$

over the unknown weights  $\theta^r$  and subject to the constraints  $\theta^r \geq 0 \forall r$  and  $\sum_{r=1}^{R(N)} \theta^r = 1$ . One can add to the least squares objective function a function that penalizes values of  $F$  that are less smooth, but we prove consistency using only the properties of the sieve space, not the penalization function. See CP for more on optional penalization functions.

We must choose a metric to show the consistency of  $\hat{F}_N$  for  $F_0$ . We choose the Lévy-Prokhorov metric, a metrization of the weak topology for the space of multivariate distributions. The space  $\mathcal{F}$  of distributions on  $\mathcal{B}$  is compact in the weak topology if  $\mathcal{B}$  itself is compact (Parthasarathy 1967, Theorem 6.4). We do not formally use compactness of the true function space  $\mathcal{F}$  in our proof of consistency. It is possible that compactness rules out the ill-posed inverse problem that CP investigate and may be present in the mixtures problem under other metrics. We assume identification is shown using the tools in Section 3 to focus on consistency here. We require that the grid of points be chosen so that the grid  $\mathcal{B}^{R(N)}$  becomes dense in  $\mathcal{B}$ . The theorem is one of consistency and so does not offer any additional insight on how choices of grids relate to statistical performance.

### Theorem 5.1

- Let  $\mathcal{F}$  be the space of all distribution functions on a finite-dimensional real space  $\mathcal{B}$ , where  $\mathcal{B}$  is compact.
- Let  $\mathcal{B}^{R(N)}$  become dense in  $\mathcal{B}$  as  $N \rightarrow \infty$ . Let  $\{y_i, x_i\}_{i=1}^N$  be iid.
- Let the space of  $X$ ,  $\mathcal{X}$ , be compact and let any open ball in  $\mathcal{X}$  have positive measure.
- Let each  $g(x, \beta)$  be continuous and bounded in both  $x$  and  $\beta$ .
- Assume the model  $\{g(x, \beta) \mid \beta \in \mathcal{B}\}$  is identified with positive probability, meaning  $F_0$  and any  $F_1 \neq F_0$ ,  $F_1 \in \mathcal{F}$  give  $P_0(x) \neq P_1(x)$  on a set  $\tilde{\mathcal{X}}$  with positive probability.

- Let the first-stage estimate of  $\hat{P}(x)$  satisfy the requirements in CP's Assumptions 3.3(i) and either 3.8 or both of 3.9 and 3.10.

Then  $\left\| \hat{F}_N - F_0 \right\|_L \rightarrow o_P(1)$ , where  $\|\cdot\|_L$  represents the Lévy-Prokhorov metric.

## 6 Pointwise Rate of Convergence

In this section, we derive convergence rates, both for the approximation of market shares (or conditional choice probabilities) and for the approximation of the underlying distribution of types  $F(\beta)$ . Here we explicitly account for a number of  $J$  product choices or choice probabilities. To simplify our notation, let  $X_i = (X'_{1,i}, \dots, X'_{J,i})'$  and denote by  $\mathcal{X} \equiv \mathcal{X}_1 \times \dots \times \mathcal{X}_J$  the support of the distribution of  $X_i$ . Here we focus on the logit but the asymptotic theory we develop in this section can be applied to other cases as long as equivalents of Assumption 6.1 and 6.2 below hold. Our results in this section relate to those in Newey (1997) and Chen (2006), although the Riemann and quadrature arguments are more novel.

We maintain the assumption of the nonparametric random utility model with logit errors. The market share equation or the conditional choice probability for alternative  $j$  is

$$P(j | x_i, f) = \int \left( \frac{\exp(x'_{j,i}\beta)}{1 + \sum_{j'=1}^J \exp(x'_{j',i}\beta)} \right) f(\beta) d\beta \text{ for any } x_i = (x'_{1,i}, \dots, x'_{J,i})' \in \mathcal{X}, \quad (9)$$

where  $f$  is a density and the true market share or the true conditional choice probability is denoted by  $P(j | x_i, f_0)$ . We further let the true choice probabilities in the data be  $P(j | x_i) \equiv P(j | x_i, f_0)$ .

We assume that a consistent estimate of the market share or a consistent estimate of the conditional choice probability  $\hat{P}(j | x_i)$  is available, so that

$$\begin{aligned} \hat{P}(j | x_i) &= P(j | x_i) + e_{j,i} \text{ and} \\ E[e_{j,i} | X_1, \dots, X_N] &= 0. \end{aligned}$$

We introduce additional simplifying notation. Define a simplex

$$\Delta_{R(N)} = \left\{ \theta = (\theta^1, \dots, \theta^{R(N)}) \mid \theta^r \geq 0, \sum_{r=1}^{R(N)} \theta^r = 1 \right\}.$$

Recall that we require  $\theta^r \geq 0$  and  $\sum_{r=1}^{R(N)} \theta^r = 1$  as the  $\theta^r$ 's are type frequencies in our approximation. We approximate  $P(j | x_i)$  using  $\sum_{r=1}^R \theta^r \gamma(r, j, x_i)$  and  $\gamma(r, j, x_i) \equiv \frac{\exp(x'_{j,i} \beta^r)}{1 + \sum_{j'=1}^J \exp(x'_{j',i} \beta^r)}$ . Let the approximation to choice probabilities be

$$\mu_\theta(j, x_i) = \sum_{r=1}^{R(N)} \theta^r \gamma(r, j, x_i) \text{ for } \theta \in \Delta_{R(N)}.$$

The estimated weights are

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Delta_{R(N)}} \frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J \left( \hat{P}(j | x_i) - \mu_\theta(j, x_i) \right)^2. \quad (10)$$

Let  $K$  be the number of random coefficients. Then we define the parameter space for choice  $j$  as the collection of market shares or conditional choice probabilities,

$$\mathcal{H}^{(j)} = \left\{ P(j | x_i, f) \mid M \geq \max_{s=0, \dots, \bar{s}} \sup_{\beta \in \mathcal{B}} |D^s f(\beta)|, \int_{\mathcal{B}} f(\beta) d\beta = 1, \mathcal{B} = \prod_{k=1}^K [\underline{\beta}_{(k)}, \bar{\beta}_{(k)}], f \in C^{\bar{s}}[\mathcal{B}] \right\},$$

where  $D^s = \frac{\partial^s}{\partial \beta_{(1)}^{\alpha_1} \dots \partial \beta_{(K)}^{\alpha_K}}$ ,  $s = \alpha_1 + \dots + \alpha_K$  with  $D^0 f = f$ , giving the collection of all derivatives of order  $s$ . Also,  $C^{\bar{s}}[\mathcal{B}]$  is a space of  $\bar{s}$ -times continuously differentiable functions defined on  $\mathcal{B}$ . We assume  $P(j | x_i, f_0) \in \mathcal{H}^{(j)}$  for  $j = 1, \dots, J$ . Therefore we assume any element of the class of density functions that generates  $\mathcal{H}^{(j)}$  is defined on a Cartesian product  $\mathcal{B}$ , is uniformly bounded by  $M$ , is  $\bar{s}$ -times continuously differentiable, and its (all own and partial) derivatives are uniformly bounded by  $M$ .

We define the space of approximating functions to choice probabilities,

$$\mathcal{H}_{R(N)}^{(j)} = \{ \mu_\theta(j, x_i) \mid \theta \in \Delta_{R(N)} \}, \quad (11)$$

with  $R(N)$  tending to infinity as  $N \rightarrow \infty$ . In order to show that any  $P(j | x_i, f) \in \mathcal{H}^{(j)}$  can be approximated by an element of  $\mathcal{H}_{R(N)}^{(j)}$ , we can put an additional structure on  $\theta$  such that

$$\theta^r = \frac{c(\beta^r) f(\beta^r)}{\sum_{r=1}^R c(\beta^r) f(\beta^r)}, r = 1, \dots, R,$$

for some chosen weights  $c(\beta^r)$ 's and where we pick the  $R$  grid points  $\{\beta^1, \dots, \beta^R\}$  to become dense

in  $\mathcal{B}$  as  $R(N) \rightarrow \infty$ . Then we have  $\mathcal{H}^{(j)} = \mathcal{H}_\infty^{(j)}$  since any element in  $\mathcal{H}_{R(N)}^{(j)}$  is a corresponding Riemann sum (where we let  $c(\beta^r) = 1/R$ ) or more generally a quadrature approximation for an integral like (9) such that

$$\mu_\theta(j, x_i) = \sum_{r=1}^R \theta^r \gamma(r, j, x_i) = \frac{1}{\sum_{r=1}^R c(\beta^r) f(\beta^r)} \sum_{r=1}^R c(\beta^r) f(\beta^r) \gamma(r, j, x_i),$$

where we expect  $\sum_{r=1}^R c(\beta^r) f(\beta^r) \xrightarrow{R \rightarrow \infty} \int f(\beta) d\beta = 1$ ,  $\sum_{r=1}^R c(\beta^r) f(\beta^r) \gamma(r, j, x_i) \xrightarrow{R \rightarrow \infty} P(j | x_i, f)$ , and so  $\mu_\theta(j, x_i) \xrightarrow{R \rightarrow \infty} P(j | x_i, f)$ . So we can approximate the observed choice probabilities  $P(j | x_i)$  arbitrarily well using the logit functions  $\gamma(r, j, x_i)$ .

In what follows, even though our notation is specific to the logit example, we develop the asymptotic theory for general basis functions that satisfy a set of conditions. We let  $\|g\|_{L_2, N} = \sqrt{\frac{1}{N} \sum_{i=1}^N g^2(x_i)}$ ,  $\|g\|_{L_2} = \sqrt{\int_{\mathcal{X}} g^2(x) d\varpi(x)}$  (the norm in  $L_2$ ), and  $\|g\|_\infty = \sup_{x \in \mathcal{X}} |g(x)|$  for any function  $g : \mathcal{X} \rightarrow \mathbb{R}$ , where  $\varpi(\cdot)$  denotes the distribution of  $X$ .

**Assumption 6.1** (i)  $\{e_{j,i}, \dots, e_{J,i}\}$  are independent across  $i = 1, \dots, N$ ; (ii)  $E[e_{j,i} | X_1, \dots, X_N] = 0$  for all  $j = 1, \dots, J$ ; (iii)  $\{X_{1,i}, \dots, X_{J,i}\}$  are iid across  $i = 1, \dots, N$ ; (iv)  $\mathcal{X}$  is compact; (v) The density of  $X$  is bounded above and is bounded away from zero on  $\mathcal{X}$ .

**Assumption 6.2** (i)  $\|P(j | x_i, f_0)\|_\infty \leq \zeta_0$  and  $\|\gamma(r, j, x)\|_\infty \leq \zeta_0$  uniformly over  $r \leq R(N)$  and for all  $j$ ; (ii)  $\|\gamma(r, j, X)\|_{L_2} \geq c_0 > 0$  uniformly over  $r \leq R(N)$  and for all  $j$ ; (iii) the  $R(N) \times R(N)$  matrix  $(E[\gamma(r, j, \cdot) \gamma(r', j, \cdot)])_{1 \leq r, r' \leq R(N)}$  is positive definite and the smallest eigenvalue of  $(E[\gamma(r, j, \cdot) \gamma(r', j, \cdot)])_{1 \leq r, r' \leq R(N)}$  is bounded away from zero uniformly in  $R(N)$  and for all  $j$ .

Assumption 6.1 is about the structure of the data. In Assumption 6.1 (i), we allow for heteroskedasticity of the measurement errors in choice probabilities,  $\{e_{j,i}, \dots, e_{J,i}\}$ , across different  $i$ 's. For the logit case, we satisfy Assumption 6.2 (i) trivially since

$$\gamma(r, j, x) = \frac{\exp(x'_j \beta^r)}{1 + \sum_{j'=1}^J \exp(x'_{j'} \beta^r)} \leq 1 \text{ uniformly over } \beta \text{ and } x$$

and  $P(j | x_i, f_0) \leq 1$  uniformly by construction. We also satisfy Assumption 6.2 (ii) trivially unless  $\mathcal{X}$  has no positive mass. In practice, Assumption 6.2 (iii) requires that the  $(R(N) \times R(N))$  matrix  $\left(\frac{1}{N} \sum_{i=1}^N \gamma(r, j, x_i) \gamma(r', j, x_i)\right)_{1 \leq r, r' \leq R(N)}$  is nonsingular for sufficiently large  $N$ .

Under these regularity conditions, the rate of convergence of the approximation to choice probabilities is obtained by the following theorem.

**Theorem 6.1** *Suppose Assumptions 6.1 and 6.2 hold. Further suppose  $\frac{R(N)^2 \log R(N)}{N} \rightarrow 0$ . Then, we have*

$$\begin{aligned} & \frac{1}{J} \sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - P(j \mid x_i, f_0) \right\|_{L_{2,N}}^2 \\ = & O_{\mathbb{P}} \left( \frac{R(N) \log(R(N)N)}{N} \right) + O_{\mathbb{P}} \left( \inf_{\mu_{\theta} \in \mathcal{H}_{R(N)}^{(j)}} \frac{1}{J} \sum_{j=1}^J \left\| \mu_{\theta}(j, x_i) - P(j \mid x_i, f_0) \right\|_{L_{2,N}}^2 \right). \end{aligned}$$

The condition  $\frac{R(N)^2 \log R(N)}{N} \rightarrow 0$  is required so that the  $(R(N) \times R(N))$  matrix

$$\left( \frac{1}{N} \sum_{i=1}^N \gamma(r, j, x_i) \gamma(r', j, x_i) \right)_{1 \leq r, r' \leq R(N)}$$

be nonsingular with sufficiently large  $N$  (see Lemma B.3 in Appendix). Roughly speaking, in view of asymptotic theory for sieve estimators, Theorem 6.1 tells that the first term corresponds to an asymptotic variance ( $\approx R(N)/N$ ) and the second term corresponds to an asymptotic bias term (Chen 2006). To derive convergence rates, we need to obtain the order of bias, i.e., the approximation error rate of our sieve approximation to arbitrary conditional choice probabilities.

**Lemma 6.1** *For all  $x \in \mathcal{X}$  and for any  $P(j \mid x_i, f) \in \mathcal{H}^{(j)}$ , there exist  $\theta^* \in \Delta_R$  such that*

$$|\mu_{\theta^*}(j, x) - P(j \mid x, f)| = O\left(R^{-\bar{s}/K}\right)$$

and that  $\frac{1}{J} \sum_{j=1}^J \left\| \mu_{\theta^*}(j, x_i) - P(j \mid x_i, f) \right\|_{L_{2,N}}^2 = O_{\mathbb{P}}\left(R^{-2\bar{s}/K}\right)$ .

The approximation error rate in Lemma 6.1 shows that we have faster convergence with a smoother true density function and slower convergence with a higher dimensional  $\beta$ .

Now we consider approximating the true conditional choice probability function. Then, from Theorem 6.1 and Lemma 6.1, we can find  $\theta_0^* \in \Delta_{R(N)}$  (i.e.  $\mu_{\theta_0^*}(j, \cdot) \in \mathcal{H}_{R(N)}^{(j)}$ ) and  $R(N)$  such that

$$\frac{1}{J} \sum_{j=1}^J \left\| \mu_{\theta_0^*}(j, x_i) - P(j \mid x_i, f_0) \right\|_{L_N^2}^2 = O_{\mathbb{P}}\left(R^{-2\bar{s}/K}\right) \asymp O_{\mathbb{P}}\left(\frac{R(N) \log(R(N)N)}{N}\right). \quad (12)$$

With this choice of  $R(N)$ , one will obtain the optimal convergence rate that balances the bias and variance of the sieve estimator. We conclude:

**Theorem 6.2** *Suppose Assumptions 6.1 and 6.2 hold. Further suppose  $\frac{R(N)^2 \log R(N)}{N} \rightarrow 0$  and  $\bar{s} > K/2$ . Then, we have*

$$\begin{aligned} \frac{1}{J} \sum_{j=1}^J \|\mu_{\hat{\theta}}(j, x_i) - P(j | x_i, f_0)\|_{L_N^2}^2 &= O_{\mathbb{P}} \left( \frac{R(N) \log(R(N)N)}{N} \right) \text{ and} \\ \sum_{r=1}^{R(N)} |\hat{\theta}^r - \theta_0^{*r}| &= O_{\mathbb{P}} \left( R(N) \sqrt{\frac{\log(R(N)N)}{N}} \right). \end{aligned}$$

One can let  $R(N) = C \cdot N^\rho$ . Combining the rate conditions on  $R(N)$ , we note that  $\rho$  should be slightly larger than  $\frac{K}{2\bar{s}+K}$  (from (12)) and be smaller than  $1/2$  (from  $\frac{R(N)^2 \log R(N)}{N} \rightarrow 0$ ). Also note that  $\bar{s} > K/2$  is required to have such a  $\rho$  exist.

Theorem 6.2 establishes the convergence rates of the distance between the true conditional choice probability and the approximated conditional choice probability as well as the  $L_1$  distance between  $\hat{\theta}$  and  $\theta_0^*$ . The  $L_1$  metric for the discrepancy between  $\hat{\theta}$  and  $\theta_0^*$  is reasonable because  $\hat{\theta}$  and  $\theta_0^*$  are frequency parameters (Devroye and Györfi 1985, Chapter I).

We can also estimate the distribution function using our estimates  $\hat{\theta} = (\hat{\theta}^1, \dots, \hat{\theta}^R)$ . Denote by  $F_0(\beta)$  the true distribution of  $\beta$ . Also let  $\mathcal{B}_0$  be the support (or compact subset of the support) of  $F_0(\beta)$ . One can estimate the distribution of the random coefficients using the following empirical distribution based on our estimates

$$\hat{F}_N(\beta) = \sum_{r=1}^{R(N)} \hat{\theta}^r \mathbf{1}[\beta^r \leq \beta].$$

We show that  $\hat{F}_N(\beta)$  can approximate the true distribution function  $F_0(\beta)$ . The approximation rate holds pointwise in  $\beta$  rather than in a function space.

**Theorem 6.3** *Suppose Assumptions in Theorem 6.2 hold. Then,*

$$\left| \hat{F}_N(\beta) - F_0(\beta) \right| = O_{\mathbb{P}} \left( R(N) \sqrt{\frac{\log(R(N)N)}{N}} \right) + O \left( R(N)^{-\bar{s}/K} \right) \text{ a.e. } \beta \in \mathcal{B}_0.$$

The rate of convergence is divided into a rate involving the statistical sampling error and one involving the approximation of a true distribution by a finite number of grid points. Again,

we have faster convergence with a smoother true density function and slower convergence with a higher dimensional  $\beta$ .

## 6.1 Breaking the Dimensionality Problem using Random Grids

We have obtained the approximation error (bias term) in Lemma 6.1 by exploiting the smoothness of the density function and the logit function where the degree of smoothness breaks the curse of dimensionality. Similarly to the arguments of Rust (1997) for dynamic programming, we can also handle the dimensionality problem using a random grids approach. Suppose the  $R$  number of grid points are drawn independently from the multivariate uniform distribution such that

$$\beta_{(k)}^r = \underline{\beta}_{(k)} + \left( \bar{\beta}_{(k)} - \underline{\beta}_{(k)} \right) u_{(k)}^r$$

for  $k = 1, \dots, K$  where  $u^r = (u_{(1)}^r, \dots, u_{(K)}^r)'$  follows the multivariate uniform distribution on  $[0, 1]^K$ . Further let

$$\theta^{*r} = \frac{(1/R)f(\beta^r)}{\sum_{r=1}^R (1/R)f(\beta^r)}. \quad (13)$$

Then we have

$$\mu_{\theta^*}(j, x) = \sum_{r=1}^R \theta^{*r} \gamma(r, j, x) = \frac{1}{(1/R) \sum_{r=1}^R f(\beta^r)} \frac{1}{R} \sum_{r=1}^R f(\beta^r) \gamma(r, j, x). \quad (14)$$

Note that applying the law of large numbers,

$$\begin{aligned} \frac{1}{R} \sum_{r=1}^R f(\beta^r) \gamma(r, j, x) &\xrightarrow{R \rightarrow \infty} \int_{\mathcal{B}} f(\underline{\beta} + (\bar{\beta} - \underline{\beta})u) \frac{\exp\left(x'_{j,i}(\underline{\beta} + (\bar{\beta} - \underline{\beta})u)\right)}{1 + \sum_{j'=1}^J \exp\left(x'_{j',i}(\underline{\beta} + (\bar{\beta} - \underline{\beta})u)\right)} du \\ &= \frac{1}{\prod_{k=1}^K (\bar{\beta}_{(k)} - \underline{\beta}_{(k)})} \int_{\mathcal{B}} f(\beta) \frac{\exp\left(x'_{j,i}\beta\right)}{1 + \sum_{j'=1}^J \exp\left(x'_{j',i}\beta\right)} d\beta, \end{aligned}$$

where the equality holds by applying the change of variables in the integral. Similarly we have

$$(1/R) \sum_{r=1}^R f(\beta^r) \xrightarrow{R \rightarrow \infty} \frac{1}{\prod_{k=1}^K (\bar{\beta}_{(k)} - \underline{\beta}_{(k)})} \int_{\mathcal{B}} f(\beta) d\beta = \frac{1}{\prod_{k=1}^K (\bar{\beta}_{(k)} - \underline{\beta}_{(k)})}.$$

Applying above two results to (14), we conclude

$$\mu_{\theta^*}(j, x) \xrightarrow{R \rightarrow \infty} \int_{\mathcal{B}} f(\beta) \frac{\exp(x'_{j,i}\beta)}{1 + \sum_{j'=1}^J \exp(x'_{j',i}\beta)} d\beta = P(j | x, f),$$

where we pick  $\theta^*$  as in (13). Finally we obtain the approximation error rate (bias term), applying the Lindeberg-Levy central limit theorem and the law of large numbers.

**Lemma 6.2** *For all  $x \in \mathcal{X}$  and for any  $P(j | x_i, f) \in \mathcal{H}^{(j)}$ , with  $\theta^*$  as in (13), we have that*

$$|\mu_{\theta^*}(j, x) - P(j | x, f)| = O_{\mathbb{P}}\left(R^{-1/2}\right).$$

and that  $\frac{1}{J} \sum_{j=1}^J \|\mu_{\theta^*}(j, x_i) - P(j | x_i, f)\|_{L_{2,N}}^2 = O_{\mathbb{P}}(R^{-1})$ .

The key conclusion is that these rates do not depend on  $K$ , the number of random coefficients. We can replace Lemma 6.1 with 6.2 and obtain alternative convergence rates for the conditional choice probability and the distribution of random coefficients.

## 7 Monte Carlo Experiments

We conduct a Monte Carlo experiment in order to study the finite sample properties of our estimator. We suppose that the true data generating process is indeed a random coefficients logit. In our Monte Carlo study,  $x_j$  is a  $2 \times 1$  vector. We generate  $x_{j,1}$  using the distribution  $x_{j,1} \sim N(1, 1)$ . Also,  $x_{j,2} \sim N(-0.8, 0.8^2) + 0.1x_{j,1}$ . There are  $J = 10$  products in each of our markets.  $J$  does not include the outside option. We use independent normal basis functions in order to approximate a density function rather than a CDF, as section 2.5 discusses. The sample size is  $J \cdot T$ , the number of products times the number of markets. We use a sample size of 20,000, corresponding to 2000 markets with 10 products each. The number of basis points is  $R = \frac{T \cdot J}{40} = 500$ . There is little measurement error in shares; we calculate market shares for aggregate data generated by 1 million consumers.<sup>10</sup>

Rather than a criterion such as integrated mean squared error, we prefer to test the structural use of our estimates. For each run, after we compute the estimate  $\hat{f}(\beta)$ , we evaluate its predictive power by drawing new product characteristics (from the same distributions) and predicting

---

<sup>10</sup>We choose the grid of points using a Halton sequence over the relevant support. We compute numerical integrals using 200 simulated draws.

shares. We compare our results to those using the true  $f^0(\beta)$ . This tests the structural use of discrete choice models to predict the demand for new goods.

We generate data using three alternative distributions of the random coefficients  $f(\beta)$ . In the first design, the tastes for characteristics are generated from a mixture of two normals,

$$0.4 \cdot N \left( \begin{bmatrix} 3 \\ -1 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{bmatrix} \right) + 0.6 \cdot N \left( \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix} \right).$$

All distributions of random coefficients will have a non-trivial variance matrix, while our basis functions are independent distributions of normal random coefficients. In the second design, the true coefficients are generated by a mixture of four normals,

$$\begin{aligned} &0.2 \cdot N \left( \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{bmatrix} \right) + 0.4 \cdot N \left( \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{bmatrix} \right) \\ &+ 0.3 \cdot N \left( \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix} \right) + 0.1 \cdot N \left( \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix} \right). \end{aligned}$$

In the third design and final design, the true coefficients are generated by the mixture of six normals,

$$\begin{aligned} &0.1 \cdot N \left( \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{bmatrix} \right) + 0.2 \cdot N \left( \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{bmatrix} \right) \\ &+ 0.2 \cdot N \left( \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{bmatrix} \right) + 0.1 \cdot N \left( \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix} \right) \\ &+ 0.3 \cdot N \left( \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix} \right) + 0.1 \cdot N \left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix} \right). \end{aligned}$$

Note that we estimate the demand model for each of the designs only once; we do not perform a formal Monte Carlo study where we replicate our estimator with new datasets.

We perform 20 replications with different datasets in order to simulate the finite sample properties of our estimator. We summarize our results in Table 1. The first column states the distribution used to generate the random coefficients. The second column is the root mean squared error (RMSE) of our out of sample prediction for market shares. The RMSE is averaged over many new, counterfactual sets of products. It is not averaged over the statistical sampling error in the original estimator; we estimate for each design only once. The final column is the

number of basis functions that have positive weight.

The results suggest that our estimator of  $f(\beta)$  is excellent for predicting the market shares of new goods, with large samples.<sup>11</sup> To understand the units, with  $J = 10$  and a mean share of 0.10, a prediction error of 0.01 (10% of 0.1) on each product corresponds to a RMSE of  $\sqrt{10 \cdot 0.01^2 / 10} = 0.01$ . Observed prediction errors are small with this sample size: 0.0001. Note that we are able to fit the observed market shares well with a fairly small number of basis functions. The mean number of nonzero basis functions ranges from 32 to 99 in the results reported below. More complex true densities require more nonzero basis points for approximation. Even when we make  $R$  large, most of these basis functions will have zero probability. This result is consistent with the literature on mixtures, which demonstrates that even quite complicated distributions can be approximated with a surprisingly small number of mixture components.

We use the 20 replications to construct the finite-sample properties of the estimator of the marginal densities of both  $\beta_1$  and  $\beta_2$ . Figure 1 includes one plot for each of the three designs and each of the two random coefficients. The solid line is the true density as recorded above, while the dashed line is the fitted density for one of the twenty replications. We use the 20 replications to construct the 5th and 95th quantiles of the estimated densities at each point of evaluation. We see that the variability across replications is quite low at this sample size. Our estimator can recover the true marginal densities.

Figure 2 plots the true versus the estimated joint density of  $f(\beta)$ , for one out of our 20 replications. A visual inspection of the densities shows that we are able to generate excellent fits to the true densities of preferences using our estimator. Our Monte Carlo experiments demonstrate that it is possible to generate good estimates of market shares and the density of random coefficients with a large number of observations. In fact, we get the modes almost spot on, even for the challenging case of six modes.

## 8 Empirical Application to Dynamic Programming

As an illustration of our estimator, we apply it to the dynamic labor supply setting from Duflo, Hanna, and Ryan (2008), hereafter referred to as DHR. They consider the problem of incentivizing teachers to go to work, and estimate a dynamic labor supply model using the method of simulated moments. The model is a single agent, dynamic programming problem with a serially

---

<sup>11</sup>Although not shown, our estimator is quite good at fitting market share equations with much smaller samples. Larger samples are needed for excellent performance in density estimation, which is not surprising given the curse of dimensionality in estimating an infinite-dimensional object.

correlated, unobserved state variable. In order to accommodate unobserved heterogeneity across teachers, they estimate a parametric, two-type mixture model. We apply the present estimator to this setting, and show that our approach allows for a more flexible approximation of the underlying heterogeneity. Further, our new estimator is quicker to run and easier to implement.

Teacher absenteeism is a major problem in India, as nearly a quarter of all teachers are absent nationwide on any given day. The absenteeism rate is nearly 50 percent among non-formal education centers, NGO-run schools designed to provide education services to rural and poor communities. To address absenteeism, DHR ran a randomized field experiment where teachers were given a combination of monitoring and financial incentives to attend school. In a sample of 113 rural, single-teacher schools, 57 randomly selected teachers were given a camera and told to take two pictures of themselves with their students on days they attend work. On top of this monitoring incentive, teachers in the treatment group also received a strong financial incentive: for every day beyond 10 they worked in a month, they received a 50 rupee bonus on top of their baseline salary of 500 rupees a month. The typical work month in the sample was about 26 days, so teachers that went to school at every opportunity would receive greater wages than the control group, which was paid a flat amount of 1000 rupees per month. The program ran for 21 months, and complete work histories were collected for all teachers over this period.

DHR evaluates this program and find the combination of incentives was successful in reducing absenteeism from 42 percent to 21 percent. In order to disentangle the confounding effects of the monitoring and financial incentives, they exploit nonlinearities in the financial incentive to estimate the labor-supply function of the teachers.

A convenient aspect of the intervention is that the financial incentives reset each month. Therefore, the model focuses on the daily work decision of teachers within a month. Denote the number of days it is possible to work in each month as  $C$ . Let  $t$  denote the current day, and let the observed state variable  $d$  denote the number of days already worked in the month. Each day a teacher faces a choice of going to school or staying home. The payoff to going to school is zero. The payoff to staying home is equal to  $\mu_i + \epsilon_{i,t}$ , where  $\mu_i$  is specific to teacher  $i$  and  $\epsilon_{i,t}$  is a shock to payoffs. The shock  $\epsilon_{i,t}$  is serially correlated.

After the end of the month ( $C$ ), the teachers receive the following payoffs, denominated in rupees, which are a function of how many days that teacher worked in the month:

$$\pi(d) = \begin{cases} 500 + (10 - d) \cdot 50 & \text{if } d \geq 10, \\ 500 & \text{otherwise .} \end{cases} \quad (15)$$

Let  $r$  be the type of the teacher in our approximation. The choice decision facing each teacher in the form of a value function for periods  $t < C$  is

$$V^r(t, d_i, \epsilon_{i,t}) = \max \{E[V^r(t+1, d_i+1, \epsilon_{i,t+1}) | \epsilon_{i,t}], \mu^r + \epsilon_{i,t} + E[V^r(t+1, d_i, \epsilon_{i,t+1}) | \epsilon_{i,t}]\}. \quad (16)$$

At time  $C$ , the value function simplifies to

$$V^r(C, d_i, \epsilon_{i,C}) = \max \{\beta\pi(d_i+1), \mu^r + \epsilon_{i,C} + \beta\pi(d_i)\}, \quad (17)$$

where  $\beta$  is the marginal utility of an additional rupee. There is no continuation value in the right side of (17) as the stock of days worked reset to zero at the end of the month. These value functions illustrate the tradeoff teachers face early in each month between accruing days worked, in the hopes of receiving an extra payoff at the end of the month, and obtaining instant gratification by skipping work.<sup>12</sup>

DHR propose a simulated method of moments estimator that matches sequences of days worked at the beginning of each month. They find parameters which generate predicted probabilities for all the possible sequences of work histories in the first five days of each month as close as possible to their empirical counterparts.<sup>13</sup> This procedure results in  $2^5 - 1 = 31$  linearly independent moments to be matched in each month. They estimate several specifications of the above model using this approach; we focus on their preferred model, in which the shock to the outside option follows an AR(1) process with correlation parameter  $\rho$ , and the teacher-specific, deterministic component of the outside option  $\mu_i$  is drawn from a bimodal normal distribution. Note that this is a very computationally intensive problem, as the researcher must integrate out over both the distribution of outside options  $\mu_i$  and the serially correlated unobservable  $\epsilon_{i,t}$  in order to produce the probabilities of sequences of days worked. This is not a computationally trivial task, as the model requires several hundred thousand simulations in order to produce estimates of the probabilities of working sequences with low variance for each type. Using the parametric mixture, DHR estimate that in a given month 97.6 percent of the teachers have out-

<sup>12</sup>The day of the month  $t$  and stock of worked days  $d$  are naturally discrete state variables. For each combination of  $t$  and  $d$ , the continuous state  $\epsilon$  is discretized into 200 bins. For each simulated value of  $\epsilon$ , the closest bin value is taken to be the actual state for the purposes of dynamic programming. For numerical integration in the calculation of expectations such as  $E[V^r(t+1, d_i+1, \epsilon_{i,t+1}) | \epsilon_{i,t}]$ , the distribution of  $\epsilon_{i,t+1}$  is used to calculate the probability  $\epsilon_{i,t+1}$  lies in each of the 200 bins, and those probabilities weight  $V^r(t+1, d_i+1, \epsilon_{i,t+1})$  in the approximation to the expectation.

<sup>13</sup>Considering data on only the first five days in each month both simplifies the computational burden and breaks much of the statistical dependence across the beginning of months (as the correlation of  $\epsilon_{i,t}$  across months is low). We treat data from different months as statistically distinct from the viewpoint of the correlation in  $\epsilon_{i,t}$ .

side options drawn from a normal distribution with mean -0.428 and variance of 0.007, with the remaining 2.4 percent drawing from a normal distribution with mean 1.781 and variance 0.050. At the estimated parameters of this model, workers drawing the first distribution generally go to school every day, while workers drawing from second distribution are likely to never attend during a given month.

There are natural bounds on the level of the outside option, as low values of  $\mu_i$  lead to teachers always working and high values lead to teachers never working. The autocorrelation parameter is bounded between negative one and positive one. The beta parameter is also sharply bounded, as it has similar effects on work behavior as the outside option outside a narrow range.

We apply the present paper’s estimator to this setting, allowing for a nonparametric distribution of heterogeneity in the outside option. We hold the marginal utility of income and the autocorrelation parameter at their values estimated under the two-type parametric model estimated in DHR:  $\beta = 0.013$  and  $\rho = 0.449$ .<sup>14</sup> We estimate the model with a discrete approximation to the underlying distribution of heterogeneity in the outside option. We let the number of basis functions range between  $R = 5$  and  $R = 40$ , with the types uniformly distributed across the economic bounds on the outside option. At the lower extreme,  $\mu_i = -2.5$ , the teachers almost always go to work, and at the upper extreme,  $\mu_i = 4.0$ , teachers almost never go to work. We solve the model under each of those  $R$  draws for every month in the data set. In addition to the intra-month time series variation in the moments, our model is identified from variation in the number of days, the distribution of workdays (teachers receive one day off on the weekends), and the number of holidays, which count as a day worked in the teacher’s payoff function, across months. These exogenous sources of variation produce different probabilities of working even under the same set of model primitives.

For each month in the data, we solve the dynamic program for all  $R$  types, and then compute the probabilities of all possible sequences of days worked in the first five days of the month. We collate these probabilities together to produce a matrix with  $R$  columns and 31 rows, where each row corresponds to the probability of observing a specific work history for that month. We then stack these matrices across months to obtain a large matrix with  $R$  columns and  $31 \times 21 = 651$  rows corresponding to the  $31 = 2^5 - 1$  possible work histories and the 21 different months. We formed the corresponding vector of empirical probabilities as the vector of dependent variables. We then

---

<sup>14</sup>In a previous draft of the paper, we also experimented with allowing for unobserved heterogeneity in the other two parameters, the marginal utility of income ( $\beta$ ) and the degree of persistence in the AR(1) process ( $\rho$ ), with unrestricted covariance across the parameters. We found that the  $\beta$  and  $\rho$  parameters were always estimated tightly around their point estimates, and therefore we focus on the distribution of  $\mu_i$  in what follows.

assigned weights to each of the  $R$  types using the inequality constrained OLS estimator.<sup>15</sup> This estimator is quite similar to the specification in equation (8), except that we do not use panel data to construct sequences of choices for the same teacher across months, only within months.

The estimated distributions of types are shown in Figure 3. We use a discrete-type approximation to the distribution of  $\mu_i$ . The vertical axis is the weight of that type in the probability mass function that is an approximation to the true distribution of types. The squares represent the point estimates; the sum of these weights is always 1. The figures also show the 90% confidence intervals for the weight on each type. The confidence intervals are small because of our large dataset. In these and subsequent figures, we present four approximations, for  $R = 5$ ,  $R = 10$ ,  $R = 20$ , and  $R = 40$ . The  $R = 40$  estimates suggest a single-peaked distribution around a modal utility of staying home of around  $-0.2/\beta = -0.2/0.013 = -15$  rupees a day. This means that, at  $\epsilon_{i,t} = 0$ , a modal teacher will go to work for only a standard incentive like the threat of being fired. However, there is a positive probability of teachers with positive values of staying home,  $\mu_i$ .

Figure 4 shows the fit of the discrete approximation models to the distribution of days worked in the data. Keep in mind we only used the first five days of each month in estimation. The mean predicted distribution matches the observed distribution relatively well. In the specifications with  $R > 5$ , our model tends to underpredict the peak by one day and slightly overpredicts the proportion of days worked in the 15–20 range. We note that these results are particularly encouraging, as the model does not use the distribution of days worked directly in the estimation; as such, these results are a partial out-of-sample test.

A feature of the teacher data is that there is a completely independent second experiment which was conducted after the first intervention, in which the incentives facing teachers were changed. Instead of working 10 days to get a marginal per-day bonus of 50 rupees, teachers in the second experiment had to work 12 days to get a marginal per-day bonus of 70 rupees. Figure 5 shows the fits of the discrete approximation models in the out-of-sample test. These data are a bit noisier than the original data set, due to a smaller sample size, but the nonparametric model also does a fairly good job of matching the distribution of days worked. Our approach tends to underpredict the proportion of zero days worked and overpredict the number of days worked

---

<sup>15</sup>We use a penalized (for the constraints) Newton’s method to minimize the least squares objective function. We use the assumption that the true types are the types included to construct confidence intervals, as in section 4. Confidence intervals are constructed using the standard OLS formulas. An observation is a teacher / month. Five days are used for each teacher / month observation. The number of observations is 1123 teacher / months. The correlation in  $\epsilon_{i,t}$  across the first five days should be low, so we do not account for autocorrelation across teacher / months for the same teacher.

above 25.

While the advantage of a nonparametric approach is clear, it is also worth emphasizing that the computational burden of the present estimator is much lower than the parametric alternative. Even with good starting values, the parametric approach requires several thousand evaluations of the objective function in a typical Newton-based optimizer. For each guess of the parameters, the dynamic model requires several million forward simulations to produce reliable estimates of the predicted probabilities of work histories, which is a significant computational burden. Furthermore, specification testing in the parametric environment is costly, and the parameter estimates under one type were very different than under two types. One clear benefit of the present approach is that one is able to run literally dozens of different models in less time than it takes to optimize a single parametric specification.

## 9 Conclusion

In this paper, we have proposed a new method for estimating general mixtures model. Our method allows the researcher to drop standard parametric assumptions, such as independent normal random coefficients, that are commonly used in applied work. In terms of computer programming and execution time, our linear regression estimator is easier to work with than simulated likelihood or method of moments estimators for parametric models. Convergence of an optimization routine to the global optimum is guaranteed under linear regression with linear constraints, something that cannot be said for other statistical objective functions. Also, our estimator is much easier to program than alternatives such as the EM algorithm.

Our estimator is useful for estimating a wide variety of models. For example, in a dynamic programming setting, the estimator allows for a nonparametric distribution of random coefficients while simultaneously cutting the computational time compared to the no-random coefficients model. The computational savings arise because we must solve the dynamic program only once for each basis vector.

We explored the asymptotic properties of the linear regression estimator. We derived the asymptotic distribution for the case where the true model has a finite and known set of types. For more nonparametric results, we showed consistency in the function space of all distributions by viewing our estimator as a sieve estimator and applying techniques in Chen and Pouzo (2008). Under additional assumptions, we derive the pointwise rate of convergence of the estimated distribution function to the true distribution function. Many alternative estimators to ours lack asymptotic results in such generality.

We apply our estimator in an empirical example of estimating the distribution of agent preferences in a dynamic programming model with an unobserved, serially correlated state variable. Our estimator use dramatically less programming and execution time than a parametric alternative.

## A Proof of Consistency: Theorem 5.1

There are two cases. If the true distribution  $F_0 \in \mathcal{F}_{R(N)}$  is in some sieve space, then the estimator becomes parametric after some  $N$ , and the estimator is consistent. The remainder of the argument assumes  $F_0$  is not in any sieve space.

We verify the conditions of CP's Theorem 3.2. Assumption 3.1(i) says that  $\{x_i\}_{i=1}^n$  is a random sample with identical distributions, which we assume. Assumption 3.1(ii) says that  $\mathcal{F}$ , the space of distributions on  $\mathcal{B}$ , is separable under the Lévy-Prokhorov metric, which holds by Theorem 3.9 of Huber (2004). Assumption 3.1(iii) is identification, which holds by assumption.

Assumption 3.2(i) says that there exists a sequence of functions  $F_N \in \mathcal{F}_{R(N)}$  such that  $\|F_N - F_0\|_L \rightarrow 0$  as  $N \rightarrow \infty$ . First,  $\mathcal{B}^{R(N)}$  becomes dense in  $\mathcal{B}$  by assumption of the theorem. Second,  $\mathcal{F}_{R(N)}$  becomes dense in  $\mathcal{F}$  because the set of distributions on a dense subset  $\mathcal{B}^{R(N)} \subset \mathcal{B}$  is itself dense. To see this, remember that the class of all distributions with finite support is dense in the class of all distributions (Aliprantis and Border 2006, Theorem 15.10). Any distribution with finite support can be approximated using a finite support in a dense subset  $\mathcal{B}^{R(N)}$  (Huber 2004). Assumption 3.2(ii) is verified in a lemma below.

Let  $\hat{m}(x, F) = \hat{P}(x) - \int g(x, \beta) dF(\beta)$  and  $m(x, F) = P(x) - \int g(x, \beta) dF(\beta)$ . Assumption 3.3(i) is that  $\hat{m}(x, F)$  and  $\hat{P}(x)$  are measurable functions of the data for almost all  $x$  and for  $F \in \mathcal{F}_N$ , which for  $\hat{P}(x)$  we assume holds by the properties of the first stage. Remark B.1 in CP states that  $\hat{m}(x, F)$  is measurable if the sieve space  $\mathcal{F}_N$  is compact, which it is, because a finite set of points  $\mathcal{B}^{R(N)}$  is compact and so the class of distributions on that space is compact (Parthasarathy 1967, Theorem 6.4). Assumption 3.3(ii) is that the estimator  $\hat{F}_N$  is well-defined, which it is by the global convexity of the linear least-squares objective function and the use of fewer types or grid points  $R(N)$  than observations,  $N$ .

Assumption 3.4 is about an optional penalization function, which we do not use for simplicity. Assumptions 3.8 and 3.9 and 3.10 are largely about the first stage, and so hold by assumption.

**Lemma A.1** *Verification of CP Assumption 3.2(ii)* For a sequence  $\{F_N\}_{N \in \mathbb{N}}$  such that  $F_N \in \mathcal{F}_{R(N)}$  and  $\|F_N - F_0\|_L \rightarrow 0$  as  $N \rightarrow \infty$ ,

$$E_x \left[ m(x, F_N)^2 \right] = \int_{\mathcal{X}} \left[ \int_{\mathcal{B}} g(x, \beta) d(F_N - F_0)(\beta) \right]^2 dG(x) \rightarrow 0,$$

where  $G(x)$  is the distribution of  $x$ .

**Proof.** First,  $\left[ \int_{\mathcal{B}} g(x, \beta) d(F_N - F_0)(\beta) \right]^2 \rightarrow 0$  pointwise over  $\mathcal{X}$  by a basic property of weak

convergence and bounded and continuous functions  $g(x, \beta)$ . Because  $\mathcal{X} \times B$  is compact, we have that  $g(x, \beta)$  is uniformly continuous, so that  $g(x_2, \beta) \rightarrow g(x, \beta)$  uniformly as  $x_2 \rightarrow x$ . Hence  $\int_{\mathcal{B}} g(x, \beta) d(F_N - F_0)(\beta)$  is continuous in  $x$ .<sup>16</sup> Then  $[\int_{\mathcal{B}} g(x, \beta) d(F_N - F_0)(\beta)]^2$  is also continuous and measurable in  $x$ . Recalling that

$$\left[ \int_{\mathcal{B}} g(x, \beta) d(F_N - F_0)(\beta) \right]^2 \rightarrow 0$$

pointwise over  $\mathcal{X}$ , the bounded convergence theorem implies that

$$\int_{\mathcal{X}} \left[ \int_{\mathcal{B}} g(x, \beta) d(F_N - F_0)(\beta) \right]^2 dG(x) \rightarrow 0,$$

as  $N \rightarrow \infty$ . ■

**Lemma A.2 Verification of CP Assumption 3.12** *This assumption is satisfied for the sieve spaces  $\mathcal{F}_{R(N)}$  and the Lévy-Prokhorov metric.*

**Proof.** Consider  $E_x [m(x, F)^2]$ , with  $m(x, F)$  defined above, as a map from  $\mathcal{F}_{R(N)}$  to  $\mathbb{R}^+$ . As the true  $F_0$  is not in the sieve space  $\mathcal{F}_{R(N)}$ , then  $E_x [m(x, F)^2]$  takes on positive values for each  $F \in \mathcal{F}_{R(N)}$ , because the model is identified on a set  $\tilde{\mathcal{X}}$  with positive probability. But  $\mathcal{F}_{R(N)}$  is compact, so  $E_x [m(x, F)^2]$  attains some minimum positive value on  $\mathcal{F}_{R(N)}$  and we can let  $B(N)$  in CP's Assumption 3.12 be that value. Because the sieve spaces are increasing, we know that  $B(N)$  is decreasing in  $N$ . Let the function  $g_m(\cdot)$  in Assumption 3.12 be any function satisfying  $g_m(0) = 0$ ,  $g_m(a) \in (0, 1)$  for  $a > 0$ . ■

## B Proof of the Rate of Convergence

We let  $R = R(N)$ . We also let  $C, C_1, C_2, \dots$  denote generic positive constants. We let  $\xi_{\min}(R)$  denote the smallest eigenvalue of  $(E[\gamma(r, j, \cdot)\gamma(r', j, \cdot)])_{1 \leq r, r' \leq R}$ . 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.

<sup>16</sup>To see continuity in  $x$ , let  $\epsilon > 0$  be given. Let  $\int_{\mathcal{B}} d(F_N - F_0)(\beta) = A < \infty$ . Then  $\int_{\mathcal{B}} (g(x_2, \beta) - g(x, \beta)) d(F_N - F_0)(\beta) \leq A \cdot \epsilon$ , because  $g(x_2, \beta) \rightarrow g(x, \beta)$  uniformly as  $x_2 \rightarrow x$ .

We first prove preliminary lemmas that are useful to prove Theorem 6.1 and Theorem 6.2. We define

$$\Psi_{N,R} = \left( \frac{1}{N} \sum_{i=1}^N \gamma(r, j, x_i) \gamma(r', j, x_i) \right)_{1 \leq r, r' \leq R} \quad \text{and} \quad \Psi_R = (E [\gamma(r, j, x_i) \gamma(r', j, x_i)])_{1 \leq r, r' \leq R},$$

where we suppress  $\Psi_{N,R}$  and  $\Psi_R$ 's dependence on  $j$  for notational simplicity.

**Lemma B.1** *Suppose Assumption 6.1 and 6.2 hold. Then*

$$\min \left[ \Pr \left\{ \|\gamma(r, \cdot)\|_{L_{2,N}}^2 \leq 2 \|\gamma(r, \cdot)\|_{L_2}^2, \forall r \right\}, \Pr \left\{ \|\gamma(r, \cdot)\|_{L_2} \leq 2 \|\gamma(r, \cdot)\|_{L_{2,N}}, \forall r \right\} \right] \geq 1 - R \exp \left( -C_1 \frac{N c_0^4}{\zeta_0^4} \right).$$

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

Lemma B.1 suggests that as  $N$  increases,  $\|\gamma(r, \cdot)\|_{L_{2,N}}$  and  $\|\gamma(r, \cdot)\|_{L_2}$  get closer to each other with probability exponentially approaching to one. Lemma B.1 implies that  $\text{diag}(\Psi_{N,R}) \leq 2 \text{diag}(\Psi_R)$  holds with probability approaching to one because the  $\|\gamma(r, \cdot)\|_{L_{2,N}}^2$ 's are diagonal elements of  $\Psi_{N,R}$  and the  $\|\gamma(r, \cdot)\|_{L_2}^2$ 's are diagonal elements of  $\Psi_R$ .

The purpose of the following Lemma B.2 is to show that  $\Psi_{N,R} \geq \Psi_R/2$  holds with probability approaching to one. This in turn shows that  $\Psi_{N,R}$  is nonsingular with probability approaching to one.

**Lemma B.2** *Let  $\mathcal{H}_R^j$ , defined in (11), be the linear space spanned by some functions  $\gamma(1, j, \cdot), \dots, \gamma(R, j, \cdot)$ . Then*

$$\Pr \left\{ \sup_{\mu \in \mathcal{H}_R^j \setminus \{0\}} \frac{\|\mu\|_{L_2}^2}{\|\mu\|_{L_{2,N}}^2} > 2 \right\} \leq R^2 \exp \left( -C_2 \frac{N}{\zeta_0^4 R^2} \right)$$

for some constant  $C$ .

**Proof.** Let  $\phi_1, \dots, \phi_M$  be an orthonormal basis of  $\mathcal{H}_R^j$  in  $L_2$  with  $M \leq R$ . Also let  $\bar{\rho}(D)$  denote the following quantity for a symmetric matrix  $D$ :

$$\bar{\rho}(D) = \sup_l \sum_l |a_l| \sum_{l'} |a_{l'}| |D_{l,l'}|,$$

where the sup is taken over sequences  $\{a_l\}_{l=1}^M$  with  $\sum_{l=1}^M a_l^2 = 1$ . Then, from Lemma 5.2 in Baraud (2002),

$$\Pr \left\{ \sup_{\mu \in \mathcal{H}_R^j \setminus \{0\}} \frac{\|\mu\|_{L_2}^2}{\|\mu\|_{L_{2,N}}^2} > c \right\} \leq M^2 \exp \left( -N \frac{(\varpi_0 - c^{-1})^2}{4\varpi_1 \max\{\bar{\rho}^2(A), \bar{\rho}(B)\}} \right) \quad (18)$$

where  $A_{l,l'} = \sqrt{E[\phi_l^2 \phi_{l'}^2]}$  and  $B_{l,l'} = \|\phi_l \phi_{l'}\|_\infty$  for  $l, l' = 1, \dots, M$  and  $\varpi_0$  and  $\varpi_1$  denote the lower bound and upper bound of the density of  $X$ , respectively. We find  $|A_{l,l'}| \leq \zeta_0^2$  and  $|B_{l,l'}| \leq \zeta_0^2$ . It follows that

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

where  $(\sum_l |a_l|)^2 \leq M \sum_l |a_l|^2$  holds by the Cauchy-Schwarz inequality (and note  $\sum_l |a_l|^2 = 1$  in our construction). Similarly we have  $\bar{\rho}(B) \leq \zeta_0^2 R$ . Noting  $M \leq R$ , from (18), we conclude

$$\Pr \left\{ \sup_{\mu \in \mathcal{H}_R^j \setminus \{0\}} \frac{\|\mu\|_{L_2}^2}{\|\mu\|_{L_{2,N}}^2} > 2 \right\} \leq R^2 \exp \left( -C_2 \frac{N}{\zeta_0^4 R^2} \right).$$

■

Next combining Lemma B.1 and Lemma B.2, we obtain a bound for the probability that the matrix  $\Psi_{N,R}$  is not smaller than a diagonal matrix in the semi-positive definite sense.

**Lemma B.3** *Suppose Assumption 6.1 and 6.2 hold. Then,*

$$\Pr \left\{ \Psi_{N,R} - \frac{\xi_{\min}(R)}{4\zeta_0^2} \text{diag}(\Psi_{N,R}) \geq 0 \right\} \geq 1 - R \exp \left( -C_1 \frac{N c_0^4}{\zeta_0^4} \right) - R^2 \exp \left( -C_2 \frac{N}{\zeta_0^4 R^2} \right).$$

**Proof.** First, note that  $\Psi_R - \frac{\xi_{\min}(R)}{\zeta_0^2} \text{diag}(\Psi_R) \geq 0$  (positive semi-definite) because  $\Psi_R$  is a positive definite matrix by Assumption 6.2 (ii) and a well-known result involving eigenvalues. Now let  $\mathcal{H}_R^j$  be the linear space spanned by  $\gamma(1, j, \cdot), \dots, \gamma(R, j, \cdot)$ . Now note that under the event  $\|\gamma(r, \cdot)\|_{L_{2,N}}^2 \leq 2 \|\gamma(r, \cdot)\|_{L_2}^2 \forall r = 1, \dots, R$ , we have  $\text{diag}(\Psi_{N,R}) \leq 2 \text{diag}(\Psi_R)$  and note that under the event  $\left\{ \sup_{\mu \in \mathcal{H}_R^j \setminus \{0\}} \frac{\|\mu\|_{L_2}^2}{\|\mu\|_{L_{2,N}}^2} > 2 \right\}$ , we have  $\Psi_{N,R} \geq \Psi_R/2$ . Therefore, under the

event  $\left\{ \|\gamma(r, \cdot)\|_{L_{2,N}}^2 \leq 2 \|\gamma(r, \cdot)\|_{L_2}^2, \forall r = 1, \dots, R \right\} \cap \left\{ \sup_{\mu \in \mathcal{H}_R^j \setminus \{0\}} \frac{\|\mu\|_{L_2}^2}{\|\mu\|_{L_{2,N}}^2} > 2 \right\}$ , we have

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

since  $\Psi_R - \frac{\xi_{\min}(R)}{\zeta_0^2} \text{diag}(\Psi_R) \geq 0$ . It follows that

$$\begin{aligned} & 1 - \Pr \left\{ \Psi_{N,R} - \frac{\xi_{\min}(R)}{4\zeta_0^2} \text{diag}(\Psi_{N,R}) \geq 0 \right\} \leq \\ & 1 - \Pr \left\{ \|\gamma(r, \cdot)\|_{L_{2,N}}^2 \leq 2 \|\gamma(r, \cdot)\|_{L_2}^2, \forall r = 1, \dots, R \right\} + 1 - \Pr \left\{ \sup_{\mu \in \mathcal{H}_R^j \setminus \{0\}} \frac{\|\mu\|_{L_2}^2}{\|\mu\|_{L_{2,N}}^2} > 2 \right\} \\ & \leq R \cdot \exp \left( -C_1 \frac{N\zeta_0^4}{\zeta_0^4} \right) + R^2 \cdot \exp \left( -C_2 \frac{N}{\zeta_0^4 R^2} \right) \end{aligned}$$

where the last inequality is obtained from Lemma B.1 and Lemma B.2. ■

**Lemma B.4** *Suppose Assumptions 6.1 and 6.2 hold. Then, for given positive sequence  $\eta_N$*

$$\begin{aligned} & \Pr \left\{ \left| \frac{1}{N} \sum_{i=1}^N e_{j,i} \gamma(r, j, x_i) \right| \leq \eta_N \|\gamma(r, j, \cdot)\|_{L_{2,N}} \text{ for all } r = 1, \dots, R \right\} \\ & \geq 1 - R \cdot \exp \left( -\frac{N\eta_N^2 \zeta_0^2}{8\zeta_0^2} \right) - R \cdot \exp \left( -C_1 \frac{N\zeta_0^4}{\zeta_0^4} \right). \end{aligned}$$

**Proof.** Hoeffding (1963)'s inequality implies that

$$\begin{aligned} & E_X \left[ \Pr \left\{ \left| \frac{1}{N} \sum_i e_{j,i} \gamma(r, j, X_i) \right| \geq \eta_N \|\gamma(r, j, \cdot)\|_{L_{2,N}}, \forall r \leq R \mid X_1, \dots, X_N \right\} \right] \quad (19) \\ & \leq E_X \left[ \sum_{r=1}^R \exp \left( -2N\eta_N^2 \|\gamma(r, j, \cdot)\|_{L_{2,N}}^2 / 4\zeta_0^2 \right) \right] \end{aligned}$$

because  $E[e_{j,i} \gamma(r, j, X_i) \mid X_1, \dots, X_N] = 0$  and because choice probabilities lie in  $[0, 1]$ ,  $-\zeta_0 \leq e_{j,i} \gamma(r, j, X_i) \leq \zeta_0$  uniformly.

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

$$\begin{aligned} \sum_{r=1}^R \exp\left(-N\eta_N^2 \|\gamma(r, \cdot)\|_{L_{2,N}}^2 / 2\zeta_0^2\right) &\leq \sum_{r=1}^R \exp\left(-N\eta_N^2 \|\gamma(r, \cdot)\|_{L_2}^2 / 8\zeta_0^2\right) \\ &\leq \sum_{r=1}^R \exp\left(-N\eta_N^2 c_0^2 / 8\zeta_0^2\right) = R \exp\left(-N\eta_N^2 c_0^2 / 8\zeta_0^2\right). \end{aligned} \quad (20)$$

From (19)-(20), combining the bound for  $\Pr\left\{ \|\gamma(r, \cdot)\|_{L_2} > 2 \|\gamma(r, \cdot)\|_{L_{2,N}}, \forall r = 1, \dots, R \right\} = R \exp\left(-C_1 \frac{Nc_0^4}{\zeta_0^4}\right)$  in Lemma B.1, the claim follows.  $\blacksquare$

The following Lemma B.5 decomposes the approximation error into a bias term and a variance term.

**Lemma B.5** *Suppose Assumptions 6.1 and 6.2 hold. Then, for any  $N \geq 1$ ,  $R \geq 2$ , and  $a > 1$ , we have for all  $\theta \in \Delta_N$ ,*

$$\frac{1}{J} \sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 \leq \frac{a+1}{a-1} \frac{1}{J} \sum_{j=1}^J \left\| \mu_{\theta}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 + \frac{\zeta_0^2}{\xi_{\min}(R)} \frac{8a^2}{a-1} \eta_N^2 R,$$

where the inequality holds with probability greater than  $1 - p_{N,R}$ ,

$$p_{N,R} = RJ \exp\left(-\frac{N\eta_N^2 c_0^2}{8\zeta_0^2}\right) + 2RJ \exp\left(-C_1 \frac{Nc_0^4}{\zeta_0^4}\right) + R^2 J \exp\left(-C_2 \frac{N}{\zeta_0^2 R^2}\right).$$

**Proof.** Because  $\mu_{\hat{\theta}}(\cdot)$  (i.e.,  $\hat{\theta}$ ) is the solution of the minimization problem in (10), we have

$$\sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - \hat{P}(j | x_i) \right\|_{L_{2,N}}^2 \leq \sum_{j=1}^J \left\| \mu_{\theta}(j, x_i) - \hat{P}(j | x_i) \right\|_{L_{2,N}}^2 \quad (21)$$

for any  $\theta \in \Delta_{R(N)}$ . Now note that

$$\begin{aligned}
& \sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - \hat{P}(j | x_i) \right\|_{L_{2,N}}^2 = \sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - P(j | x_i, f_0) + P(j | x_i, f_0) - \hat{P}(j | x_i) \right\|_{L_{2,N}}^2 \\
& = \sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 - \sum_{j=1}^J \frac{2}{N} \sum_i e_{j,i} (\mu_{\hat{\theta}}(j, x_i) - P(j | x_i, f_0)) + \sum_{j=1}^J \|e_{j,i}\|_{L_{2,N}}^2,
\end{aligned} \tag{22}$$

where we use the definition  $e_{j,i} = \hat{P}(j | x_i) - P(j | x_i, f_0)$ . Similarly we obtain

$$\begin{aligned}
& \sum_{j=1}^J \left\| \mu_{\theta}(j, x_i) - \hat{P}(j | x_i) \right\|_{L_{2,N}}^2 = \\
& \sum_{j=1}^J \left\| \mu_{\theta}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 - \sum_{j=1}^J \frac{2}{N} \sum_i e_{j,i} (\mu_{\theta}(j, x_i) - P(j | x_i, f_0)) + \sum_{j=1}^J \|e_{j,i}\|_{L_{2,N}}^2.
\end{aligned} \tag{23}$$

Subtracting (23) from (22) and noting (23)  $\geq$  (22) from (21), we obtain

$$\begin{aligned}
& \sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 \\
& \leq \sum_{j=1}^J \left\| \mu_{\theta}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 + \sum_{j=1}^J \frac{2}{N} \sum_i e_{j,i} (\mu_{\hat{\theta}}(j, x_i) - \mu_{\theta}(j, x_i)).
\end{aligned}$$

Now let  $V_{N,r}(j) = \frac{1}{N} \sum_{i=1}^N e_{j,i} \gamma(r, j, x_i)$ . Then, we have

$$\frac{1}{N} \sum_i e_{j,i} (\mu_{\hat{\theta}}(j, x_i) - \mu_{\theta}(j, x_i)) = \sum_{r=1}^R V_{N,r}(j) (\hat{\theta}^r - \theta^r)$$

by construction of  $\mu_{\hat{\theta}}$  and  $\mu_{\theta}$  and we obtain by the triangle inequality,

$$\sum_{j=1}^J \left\| \mu_{\hat{\theta}}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 \leq \sum_{j=1}^J \left\| \mu_{\theta}(j, x_i) - P(j | x_i, f_0) \right\|_{L_{2,N}}^2 + 2 \sum_{j=1}^J \sum_{r=1}^R |V_{N,r}(j)| \left| \hat{\theta}^r - \theta^r \right|. \tag{24}$$

Define two events  $E_{0,j} = \bigcap_{r=1}^R \left\{ |V_{N,r}(j)| \leq \eta_N \|\gamma(r, \cdot)\|_{L_{2,N}} \right\}$  for some positive sequence  $\eta_N$  and  $E_{1,j} = \left\{ \Psi_{N,R} - \frac{\xi_{\min}(R)}{4\zeta_0^2} \text{diag}(\Psi_{N,R}) \geq 0 \right\}$ . Then, under  $E_{0,j} \cap E_{1,j}$ , we note

$$\begin{aligned}
\sum_{r=1}^R V_{N,r}^2(j) (\hat{\theta}^r - \theta^r)^2 &\leq \eta_N^2 \sum_{r=1}^R \|\gamma(r, \cdot)\|_{L_{2,N}}^2 (\hat{\theta}^r - \theta^r)^2 \\
&= \eta_N^2 \frac{1}{N} \sum_{i=1}^N \sum_{r=1}^R (\hat{\theta}^r - \theta^r)^2 \gamma(r, j, x_i)^2 \\
&= \eta_N^2 (\hat{\theta} - \theta)' \text{diag}(\Psi_{N,R}(j)) (\hat{\theta} - \theta) \\
&\leq \eta_N^2 (\xi_{\min}(R)/4\zeta_0^2)^{-1} (\hat{\theta} - \theta)' \Psi_{N,R}(j) (\hat{\theta} - \theta) \\
&= \eta_N^2 (\xi_{\min}(R)/4\zeta_0^2)^{-1} \|\mu_{\hat{\theta}}(j, x_i) - \mu_{\theta}(j, x_i)\|_{L_{2,N}}^2.
\end{aligned} \tag{25}$$

The above uses Lemma B.3. From (24) and (25), it follows that on  $\bigcap_{j=1}^J (E_{0,j} \cap E_{1,j})$

$$\begin{aligned}
&\sum_{j=1}^J \|\mu_{\hat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 \\
&\leq \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 + 2 \sum_{j=1}^J \sum_{r=1}^R |V_{N,r}(j)| |\hat{\theta}^r - \theta^r| \\
&\leq \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 + 2 \sum_{j=1}^J \sqrt{R} \sqrt{\sum_{r=1}^R V_{N,r}^2(j) (\hat{\theta}^r - \theta^r)^2} \\
&\leq \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 + 2 \sum_{j=1}^J \eta_N \sqrt{R \left( \frac{\xi_{\min}(R)}{4\zeta_0^2} \right)^{-1}} \|\mu_{\hat{\theta}}(j, x_i) - \mu_{\theta}(j, x_i)\|_{L_{2,N}} \\
&\leq \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 \\
&+ 2 \sum_{j=1}^J \eta_N \sqrt{R \left( \frac{\xi_{\min}(R)}{4\zeta_0^2} \right)^{-1}} \left( \|\mu_{\hat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}} + \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}} \right)
\end{aligned}$$

by the Cauchy-Schwarz inequality, RCS, and the triangle inequality. Applying the inequality  $2xy \leq x^2a + y^2/a$  (any  $x, y, a > 0$ ) to  $x = \eta_N \sqrt{R (\xi_{\min}(R)/4\zeta_0^2)^{-1}}$  and  $y = \|\mu_{\hat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}$  and to  $x = \eta_N \sqrt{R (\xi_{\min}(R)/4\zeta_0^2)^{-1}}$  and  $y = \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}$ , respectively, we obtain under  $\bigcap_{j=1}^J (E_{0,j} \cap E_{1,j})$ ,

$$\begin{aligned}
&\sum_{j=1}^J \|\mu_{\hat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 \\
&\leq \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 + \sum_{j=1}^J \|\mu_{\hat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 / a \\
&+ \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 / a + 2a\eta_N^2 R (\xi_{\min}(R)/4\zeta_0^2)^{-1} \cdot J
\end{aligned}$$

It follows that under  $\bigcap_{j=1}^J (E_{0,j} \cap E_{1,j})$ , for all  $a > 1$ ,

$$\sum_{j=1}^J \|\mu_{\hat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 \leq \frac{a+1}{a-1} \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 + \frac{\zeta_0^2}{\xi_{\min}(R)} \frac{8a^2}{a-1} \eta_N^2 R J.$$

The conclusion follows from

$$\Pr \left\{ \left( \bigcap_{j=1}^J E_{0,j} \right)^C \right\} \leq \sum_{j=1}^J \Pr \{ E_{0,j}^C \} = J \left\{ R \exp \left( -\frac{N \eta_N^2 c_0^2}{8 \zeta_0^2} \right) + R \exp \left( -C_1 \frac{N c_0^4}{\zeta_0^4} \right) \right\}$$

by Lemma B.4 and  $\Pr \left\{ \left( \bigcap_{j=1}^J E_{1,j} \right)^C \right\} = J \left\{ R \exp \left( -C_1 \frac{N c_0^4}{\zeta_0^4} \right) + R^2 \exp \left( -C_2 \frac{N}{\zeta_0^4 R^2} \right) \right\}$  by Lemma B.3. ■

## B.1 Proof of Theorem 6.1

From Lemma B.5, we find that the best convergence rate will be obtained when the order of  $\eta_N$  is as small as possible while keeping the convergence of  $p_{N,R}$  to zero. By inspecting  $p_{N,R} = J \left\{ R \exp \left( -\frac{N \eta_N^2 c_0^2}{8 \zeta_0^2} \right) + 2R \exp \left( -C_1 \frac{N c_0^4}{\zeta_0^4} \right) + R^2 \exp \left( -C_2 \frac{N}{\zeta_0^4 R^2} \right) \right\}$ , we note that the optimal rate is obtained when  $\eta_N = O \left( \sqrt{\frac{\log(R(N)N)}{N}} \right)$  since the first term in  $p_{N,R}$  dominates the second term in  $p_{N,R}$  when  $\eta_N$  is small enough and  $p_{N,R} \rightarrow 0$  with  $\eta_N = O \left( \sqrt{\frac{\log(R(N)N)}{N}} \right)$ . The inspection of the third term in  $p_{N,R}$  reveals that we also require  $R = R(N)$  should satisfy  $\frac{R(N)^2 \log(R(N))}{N} \rightarrow 0$  so that  $p_{N,R} \rightarrow 0$ .

The result of Theorem 6.1 follows from these requirements and Lemma B.5 as

$$\frac{1}{J} \sum_{j=1}^J \|\mu_{\hat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}}^2 = O_{\mathbb{P}} \left( \max \left\{ \frac{R(N) \log(R(N)N)}{N}, \frac{1}{J} \sum_{j=1}^J \|\mu_{\theta}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}} \right\} \right)$$

since  $\frac{\zeta_0^2}{\xi_{\min}(R)} \frac{8a^2}{a-1} \eta_N^2 R(N) = O \left( \frac{\log(R(N)N)}{N} R(N) \right)$  under  $\eta_N = O \left( \sqrt{\frac{\log(R(N)N)}{N}} \right)$  and because Lemma B.5 holds for any  $\theta \in \Delta_{R(N)}$ . Combining the rate conditions on  $R(N) = C \cdot N^{\rho}$  with  $\frac{K}{2\bar{s}+K} < \rho$  and  $\frac{R(N)^2 \log R(N)}{N} \rightarrow 0$ , we obtain  $\frac{K}{2\bar{s}+K} < \rho < 1/2$ . Also see that  $\bar{s} > K/2$  is required

for  $\rho$  to exist.

## B.2 Proof of Lemma 6.1

First we construct approximating power series as tensor products of higher order polynomials of  $\beta$  with the length equal to  $L$

$$\{\varphi_1(\beta), \dots, \varphi_l(\beta), \dots, \varphi_L(\beta)\}$$

where  $\varphi_l(\beta)$  is the  $l^{\text{th}}$  tensor product polynomials. For example, we can let  $\varphi_1(\beta) = 1$ ,  $\varphi_2(\beta) = \beta_{(1)}$ ,  $\dots$ , and  $\varphi_l(\beta) = \beta_{(1)}^{l_1} \beta_{(2)}^{l_2} \dots \beta_{(K)}^{l_K}$ .

Now note that  $\left( \frac{\exp(x'_j \beta)}{1 + \sum_{j'=1}^J \exp(x'_{j'} \beta)} \right) f(\beta)$  is a member of the Hölder class ( $\bar{s}$ -smooth) of functions since it is uniformly bounded and all of its own and partial derivatives up to the order of  $\bar{s}$  are also uniformly bounded by our restriction on  $f$  in  $\mathcal{H}^{(j)}$  and the logit specification assumption. Therefore, we can approximate  $\left( \frac{\exp(x'_j \beta)}{1 + \sum_{j'=1}^J \exp(x'_{j'} \beta)} \right) f(\beta)$  well using power series (see Chen, 2006) and obtain the approximation error rate due to Timan (1963) as

$$\sup_{\beta \in \mathcal{B}} \left| \left( \frac{\exp(x'_j \beta)}{1 + \sum_{j'=1}^J \exp(x'_{j'} \beta)} \right) f(\beta) - \sum_{l=1}^L a_l(x) \varphi_l(\beta) \right| = O(L^{-\bar{s}/K}) \quad (26)$$

all  $x \in \mathcal{X}$ . Now define

$$\underline{\beta}_{(k)} = b_{(k),1} < b_{(k),2} < \dots < b_{(k),r_j+1} = \bar{\beta}_{(k)}, \quad k = 1, \dots, K,$$

be partitions of the intervals  $[\underline{\beta}_{(k)}, \bar{\beta}_{(k)}]$ ,  $k = 1, \dots, K$ , into  $r_1, \dots, r_K$  subintervals, respectively. Then we can define the  $R = r_1 r_2 \dots r_K$  number of subcubes

$$C_{\iota_1, \dots, \iota_K} = \prod_{k=1}^K [b_{(k),\iota_k}, b_{(k),\iota_k+1}], \quad \iota_k = 1, 2, \dots, r_k,$$

which become a partition  $P(\mathcal{B})$  of  $\mathcal{B}$ . For any choice of  $R$  points

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

(one  $b_{\iota_1, \dots, \iota_K}$  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

$$\{c(\iota_1, \dots, \iota_K) \equiv c(b_{\iota_1, \dots, \iota_K}) \mid \iota_k = 1, \dots, r_k, k = 1, \dots, K\}$$

such that

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

where  $\mathcal{R}(\delta_R)$  denotes a remainder term with  $\delta_R = \max\{\text{diam}(C_{\iota_1, \dots, \iota_K}) : C_{\iota_1, \dots, \iota_K} \in P(\mathcal{B})\}$ . 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  $\beta_{(k)}$ '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  $\varphi_l(\beta) = \prod_{k=1}^K \varphi_{l,k}(\beta_{(k)})$ . If we approximate  $\int \varphi_{l,k}(\beta_{(k)}) d\beta_{(k)}$  using a univariate quadrature with weights  $\{c_k(1), \dots, c_k(r_k)\}$ , 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 remainder term. Then we can make the univariate quadrature 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  $p \leq r_k$ .

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  $\beta_{(k)}$ '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

$$\begin{aligned} \int \varphi_l(\beta) d\beta &= \prod_{k=1}^K \int \varphi_{l,k}(\beta_{(k)}) d\beta_{(k)} \\ &= \prod_{k=1}^K \left( \sum_{\iota_k=1}^{r_1} c_k(\iota_k) \varphi_{l,k}(b_{(k),\iota_k}) \right) \end{aligned} \quad (27)$$

for  $l = 1, \dots, L$ . By adding and subtracting terms, it follows that

$$\begin{aligned} P(j|x, f) - \mu_{\theta^*}(j, x) &= \\ &= \int f(\beta) \left( \frac{\exp(x'_j \beta)}{1 + \sum_{j'=1}^J \exp(x'_{j'} \beta)} \right) d\beta - \int \sum_{l=1}^L a_l(x) \varphi_l(\beta) d\beta \end{aligned} \quad (28)$$

$$+ \sum_{l=1}^L a_l(x) \int \varphi_l(\beta) d\beta - \sum_{l=1}^L a_l(x) \prod_{k=1}^K \left( \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{(k),\iota_k}) \right) \quad (29)$$

$$+ \sum_{l=1}^L a_l(x) \prod_{k=1}^K \left( \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{(k),\iota_k}) \right) - \mu_{\theta^*}(j, x). \quad (30)$$

We first bound the terms in (28) as

$$\begin{aligned} & \left| \int f(\beta) \left( \frac{\exp(x'_j \beta)}{1 + \sum_{j'=1}^J \exp(x'_{j'} \beta)} \right) d\beta - \int \sum_{l=1}^L a_l(x) \varphi_l(\beta) d\beta \right| \\ & \leq \int \left| f(\beta) \left( \frac{\exp(x'_j \beta)}{1 + \sum_{j'=1}^J \exp(x'_{j'} \beta)} \right) - \sum_{l=1}^L a_l(x) \varphi_l(\beta) \right| d\beta = O\left(L^{-\bar{s}/K} \cdot \text{vol}(\mathcal{B})\right) \end{aligned}$$

from (26). Second note that (29) becomes zero due to (27).

Now we construct  $\tilde{\varphi}_l(b^r)$   $r = 1, \dots, R$  such that

$$\sum_{r=1}^R \tilde{\varphi}_l(b^r) = \prod_{k=1}^K \left\{ \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{(k),\iota_k}) \right\}$$

with  $R = r_1 \cdots r_K$  and  $b^r = (b_{(1)}^r, b_{(2)}^r, \dots, b_{(K)}^r)$  in  $\mathbf{b} \equiv \{b : b = (b_{(1),\iota_1}, \dots, b_{(K),\iota_K}), \iota_1 =$

$1, \dots, r_1, \dots, \iota_K = 1, \dots, r_K\}$ . Then, we have

$$\tilde{\varphi}_l(b^r) = c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) \varphi_l(b^r)$$

for any  $b^r = (b_{(1),\iota_1}, \dots, b_{(K),\iota_K})$  in  $\mathbf{b}$ ,  $r = 1, \dots, R = r_1 r_2 \cdots r_K$ . Define

$$\theta^{*r} = \frac{c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)}{\sum_{r=1}^R c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)}.$$

It follows that

$$\begin{aligned} & \left| \sum_{l=1}^L a_l(x) \tilde{\varphi}_l(b^r) - \theta^{*r} \gamma(r, j, x_j) \right| \tag{31} \\ &= \left| \frac{\sum_{l=1}^L a_l(x) c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) \varphi_l(b^r)}{c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)} - \frac{\sum_{r=1}^R c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)}{\sum_{r=1}^R c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)} \gamma(r, j, x_j) \right| \\ &= \left| c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) \right| \left| \sum_{l=1}^L a_l(x) \varphi_l(b^r) - \frac{f(b^r)}{\sum_{r=1}^R c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)} \gamma(r, j, x_j) \right| \\ &\leq \left| c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) \right| \left| \sum_{l=1}^L a_l(x) \varphi_l(b^r) - f(b^r) \gamma(r, j, x_j) \right| \\ &+ \left| c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) \right| \left| \frac{\sum_{r=1}^R c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r) - 1}{\sum_{r=1}^R c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)} \right| |f(b^r) \gamma(r, j, x_j)| \\ &= O(R^{-1} L^{-\bar{s}/K}), \end{aligned}$$

First note that  $\left| c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) \right|$  is at most  $O(R^{-1})$ , which is obtained if one uses the uniform weights, i.e.,  $c_k(b_{(k),\iota_k}^r) = \dots = c_k(b_{(k),\iota_k}^R)$  for  $k = 1, \dots, K$ . Second note that  $\left| \sum_{l=1}^L a_l(x) \varphi_l(b^r) - f(b^r) \gamma(r, j, x_j) \right| = O(L^{-\bar{s}/K})$  due to the bound in (26). Third note that  $\sum_{r=1}^R c_1(b_{(1),\iota_1}^r) \cdots c_K(b_{(K),\iota_K}^r) f(b^r)$  is another quadrature approximation of integral  $\int_{\mathcal{B}} f(\beta) d\beta = 1$ . Since  $f(\beta)$  itself belongs to a Hölder class, the approximation error rate of this integral becomes  $O(L^{-\bar{s}/K})$ . Combining these results, our conclusion of the bound in (31) follows.

Now by letting  $\theta^* = (\theta^{*1}, \dots, \theta^{*R})$ , we have

$$\begin{aligned}
& \left| \sum_{l=1}^L a_l(x) \prod_{k=1}^K \left\{ \sum_{\iota_k=1}^{r_k} c_k(\iota_k) \varphi_{l,k}(b_{(k),\iota_k}) \right\} - \mu_{\theta^*}(j, x) \right| \\
&= \left| \sum_{r=1}^R \sum_{l=1}^L a_l(x) \tilde{\varphi}_l(b^r) - \sum_{r=1}^R \theta^{*r} \gamma(r, j, x_j) \right| \\
&\leq \sum_{r=1}^R \left| \sum_{l=1}^L a_l(x) \tilde{\varphi}_l(b^r) - \theta^{*r} \gamma(r, j, x_j) \right| \\
&= O(RR^{-1}L^{-\bar{s}/K}) = O(L^{-\bar{s}/K})
\end{aligned}$$

where the second equality holds by (31). From this, we bound (30) as  $O(L^{-\bar{s}/K})$ .

Combining the bounds, we conclude

$$\begin{aligned}
|P(j|x, f) - \mu_{\theta^*}(j, x)| &= O\left(L^{-\bar{s}/K} \cdot \text{vol}(\mathcal{B})\right) + 0 + O(L^{-\bar{s}/K}) = O(L^{-\bar{s}/K}) \\
&= O\left(\left(\left(R^{1/K} + 1\right)^K\right)^{-\bar{s}/K}\right) = O\left(\left(R^{1/K} + 1\right)^{-\bar{s}}\right) \leq O(R^{-\bar{s}/K}).
\end{aligned}$$

The second conclusion in the lemma is trivial since the above holds for all  $x \in \mathcal{X}$ .

### B.3 Proof of Lemma 6.2

Define

$$\begin{aligned}
\hat{A}(x) &= \frac{1}{R} \sum_{r=1}^R f(\beta^r) \gamma(r, j, x), \quad A(x) = \frac{1}{\prod_{k=1}^K (\bar{\beta}_{(k)} - \underline{\beta}_{(k)})} \int_B f(\beta) \frac{\exp(x'_{j,i} \beta)}{1 + \sum_{j'=1}^J \exp(x'_{j',i} \beta)} d\beta \\
\hat{B} &= (1/R) \sum_{r=1}^R f(\beta^r), \quad B = \frac{1}{\prod_{k=1}^K (\bar{\beta}_{(k)} - \underline{\beta}_{(k)})} \int_B f(\beta) d\beta.
\end{aligned}$$

Then, by plugging in definition of terms and adding and subtracting terms, we have

$$\begin{aligned}\sqrt{R}(\mu_{\theta^*}(j, x) - P(j | x, f)) &= \sqrt{R} \left( \hat{A}(x)/\hat{B} - A(x)/B \right) \\ &= \sqrt{R} \frac{\left( \hat{A}(x) - A(x) \right)}{\hat{B}} - \sqrt{R} \frac{A(x)}{B\hat{B}} (\hat{B} - B).\end{aligned}\tag{32}$$

Applying the Lindeberg-Levy central limit theorem and the law of large numbers, we further obtain

$$\begin{aligned}\sqrt{R}(\hat{A}(x) - A(x)) &= O_{\mathbb{P}}(1) \\ \hat{A}(x) &= A(x) + o_{\mathbb{P}}(1) \\ \hat{B} &= B + o_{\mathbb{P}}(1) \\ \sqrt{R}(\hat{B} - B) &= O_{\mathbb{P}}(1)\end{aligned}\tag{33}$$

uniformly over  $x \in \mathcal{X}$  since  $\frac{\exp(x'_{j,i}\beta)}{1 + \sum_{j'=1}^J \exp(x'_{j',i}\beta)}$  is uniformly bounded. Combining (32) and (33), we obtain  $\sqrt{R}(\mu_{\theta^*}(j, x) - P(j | x, f)) = O_{\mathbb{P}}(1)$ .

The second claim in Lemma 6.2 follows by applying the dominated convergence theorem, noting that  $\hat{A}(x)$  is uniformly bounded by  $M \geq \sup_{\beta \in \mathcal{B}} |f(\beta)|$ .

#### B.4 Proof of Theorem 6.2

The first result of Theorem 6.2 follows from Theorem 6.1 combined with Lemma 6.1 with  $R(N)$  satisfying (12) so that the variance term and the bias term are balanced. Now we show the second claim. Note that with probability approaching to one, we have  $2 \|\gamma(r, j, x_i)\|_{L_{2,N}} \geq \|\gamma(r, j, x)\|_{L_2} \geq c_0 > 0$  for all  $r = 1, \dots, R$  and  $j = 1, \dots, J$  by Assumption 6.2(ii) and Lemma

B.1. It follows that

$$\begin{aligned}
& \frac{c_0}{2} \sum_{r=1}^R \left| \widehat{\theta}^r - \theta_0^{*r} \right| \\
\leq & \frac{1}{J} \sum_{j=1}^J \sum_{r=1}^R \|\gamma(r, j, x_i)\|_{L_{2,N}} \left| \widehat{\theta}^r - \theta_0^{*r} \right| \leq \frac{1}{J} \sum_{j=1}^J \sqrt{R} \sqrt{\sum_{r=1}^R \|\gamma(r, j, x_i)\|_{L_{2,N}}^2} \left( \widehat{\theta}^r - \theta_0^{*r} \right)^2 \\
\leq & \sqrt{\frac{\xi_{\min}(R)}{4\zeta_0^2}} \sqrt{R} \frac{1}{J} \sum_{j=1}^J \|\mu_{\widehat{\theta}}(j, x_i) - \mu_{\theta_0^*}(j, x_i)\|_{L_{2,N}} \\
\leq & \sqrt{\frac{\xi_{\min}(R)}{4\zeta_0^2}} \sqrt{R} \frac{1}{J} \sum_{j=1}^J \left( \|\mu_{\widehat{\theta}}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}} + \|\mu_{\theta_0^*}(j, x_i) - P(j|x_i, f_0)\|_{L_{2,N}} \right) \\
\leq & \sqrt{\frac{\xi_{\min}(R)}{4\zeta_0^2}} \sqrt{R} \left( \sqrt{\frac{1}{J} \sum_{j=1}^J \|\mu_{\widehat{\theta}}(j, \cdot) - P(j|\cdot, f_0)\|_{L_{2,N}}^2} + \sqrt{\frac{1}{J} \sum_{j=1}^J \|\mu_{\theta_0^*}(j, \cdot) - P(j|\cdot, f_0)\|_{L_{2,N}}^2} \right)
\end{aligned}$$

where the second inequality holds by RCS, the third inequality holds similarly with (25), the fourth inequality holds by the triangle inequality, and the last inequality is due to the Cauchy-Schwarz inequality. Therefore, from Lemma 6.1 and the first result of Theorem 6.2 with our choice of  $R(N)$  and  $\eta_N = O\left(\sqrt{\frac{\log(R(N)N)}{N}}\right)$  (so that the variance term balances with the bias term), it follows that

$$\sum_{r=1}^R \left| \widehat{\theta}^r - \theta_0^{*r} \right| = O\left(\sqrt{R(N)} \sqrt{\frac{R(N) \cdot \log(R(N)N)}{N}}\right) = O\left(R(N) \sqrt{\frac{\log(R(N)N)}{N}}\right).$$

### B.5 Proof of Theorem 6.3

Define a pseudo true distribution such that  $F_R(\beta) = \sum_{r=1}^R \theta_0^{*r} 1[\beta^r \leq \beta]$ . It is not difficult to see that

$$\begin{aligned}
\sup_{\beta \in \mathcal{B}_0} \left| \widehat{F}(\beta) - F_R(\beta) \right| &= \sup_{\beta \in \mathcal{B}_0} \left| \sum_{r=1}^R \widehat{\theta}^r 1[\beta^r \leq \beta] - \sum_{r=1}^R \theta_0^{*r} 1[\beta^r \leq \beta] \right| \\
&= \sup_{\beta \in \mathcal{B}_0} \left| \sum_{r=1}^R \left( \widehat{\theta}^r - \theta_0^{*r} \right) 1[\beta^r \leq \beta] \right| \\
&\leq \sum_{r=1}^R \left| \widehat{\theta}^r - \theta_0^{*r} \right| = O_{\mathbb{P}}\left(R(N) \sqrt{\frac{\log(R(N)N)}{N}}\right)
\end{aligned}$$

where the last inequality holds by the triangle inequality and the last equality holds by Theorem 6.2. Note that

$$F_0(\beta) = \int f_0(b) 1[b \leq \beta] db$$

and  $F_R(\beta)$  becomes a quadrature approximation. We use a similar strategy with the proof of Lemma 6.1 where we approximate  $f(b)$  with  $\sum_{l=1}^L a_{f,l} \varphi_l(b)$  such that  $\sup_{\beta \in \mathcal{B}_0} \left| f(b) - \sum_{l=1}^L a_{f,l} \varphi_l(b) \right| = O(L^{-\bar{s}/K})$  due to Timan (1963). We find

$$F_0(\beta) - F_R(\beta) = O\left(R^{-\bar{s}/K}\right) \text{ a.e. } \beta \in \mathcal{B}_0.$$

We conclude

$$\begin{aligned} \left| \widehat{F}_N(\beta) - F_0(\beta) \right| &\leq \left| \widehat{F}_N(\beta) - F_R(\beta) \right| + |F_R(\beta) - F_0(\beta)| \\ &= O_{\mathbb{P}}\left(R(N) \sqrt{\frac{\log(R(N)N)}{N}}\right) + O\left(R^{-\bar{s}/K}\right) \text{ a.e. } \beta \in \mathcal{B}_0. \end{aligned}$$

## References

- [1] Akerberg, Daniel A. (2009). "A New Use of Importance Sampling to Reduce Computational Burden in Simulation Estimation." *Quantitative Marketing and Economics*.
- [2] Akhiezer, N.I. (1965), *The classical moment problem and some related questions in analysis*, Oliver & Boyd.
- [3] Aliprantis, C.D. and K.C. Border (2006), *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer.
- [4] Amemiya, T. (1983), "Non-Linear Regression Models," *Handbook of Econometrics I*, edited by Z. Griliches and M.D. Intrilligator, 333-389.
- [5] Andrews, D.K.W. (1994), "Empirical Process Methods in Econometrics," *Handbook of Econometrics IV*, edited by R.F. Engle and D.L. McFadden, 2247-2294.
- [6] Andrews, D.K.W. (2000), "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space," *Econometrica*, 68, 399-405.
- [7] Andrews D.K.W. (1999), "Estimation When a Parameter is on a Boundary," *Econometrica* 67, 1341-1383.
- [8] Andrews D.K.W. (2002), "Generalized Method of Moments Estimation When a Parameter is on a Boundary," *Journal of Business & Economics Statistics* 20-4, 530-544.
- [9] Andrews D.K.W and P. Guggenberger (2009a), "Hybrid and Size-corrected Subsample Methods," *Econometrica*, 77.
- [10] Andrews, D.K.W and P. Guggenberger (2009b), "Applications of Subsampling, Hybrid, and Size-Correction Methods," working paper.
- [11] Andrews, D.K.W and P. Guggenberger (2010), "Asymptotic Size and a Problem with Subsampling and the m Out of n Bootstrap", *Econometric Theory*, 26.
- [bfkr0] Bajari, P, J. Fox, K. Kim, and S. Ryan (2009), "The Random Coefficient Logit Model is Identified" working paper.
- [12] Baraud, Y. (2002), "Model Selection for Regression on a Random Design", *ESAIM Probability & Statistics* 7, 127-146

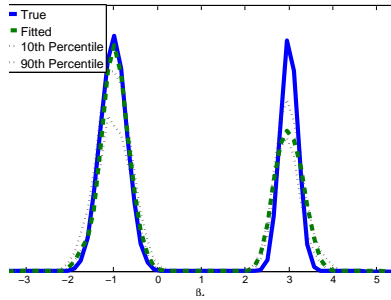
- [13] Berry, S and Haile, P. (2008), “Nonparametric Identification of Multinomial Choice Models with Heterogeneous Consumers and Endogeneity”, working paper.
- [14] Berry, S., J. Levinsohn, and A. Pakes (1995), “Automobile Price in Market Equilibrium”, *Econometrica* (63), July 1995.
- [15] Billingsley, Patrick (1995), *Probability and Measure*, 3rd Edition.
- [16] Böhning, D. “Convergence of Simar’s Algorithm for Finding the Maximum Likelihood Estimate of a Compound Poisson Process”, *The Annals of Statistics*, 10(3), 1006–1008. 1982.
- [17] Briesch, R.A., P.K. Chintagunta, and R.L. Matzkin (2007), “Nonparametric Discrete Choice Models with Unobserved Heterogeneity”, Working paper.
- [18] Chen, X. (2006), “Large Sample Sieve Estimation of Semi-Nonparametric Models,” *Handbook of Econometrics* VI, Elsevier.
- [19] Chen, X. and D. Pouzo (2008), "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Moments", Yale working paper.
- [20] Chen, X. and D. Pouzo (2008b), "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals", *Journal of Econometrics*, forthcoming.
- [21] Cramér, H. and H. Wold (1936), “Some Theorems on Distribution Functions”, *Journal of the London Mathematical Society*, s1-11(4), 290–294.
- [22] Dempster, A.P., N.M. Laird and D.B. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39, 1, 1-38.
- [23] D’Haultfoeuille, Xavier (2009), “On the Completeness Condition in Nonparametric Instrumental Problems”, *Econometric Theory*.
- [24] Devroye, Luc, and Laszlo Györfi (1985), *Nonparametric Density Estimation, The  $L_1$  View*, New York, Wiley.
- [25] Duflo, Esther, Rema Hanna and Stephen P. Ryan (2008), "Monitoring Works: Getting Teachers to Come to School", MIT working paper.
- [26] Fox, Jeremy T. and Amit Gandhi (2009a), “Identifying Heterogeneity in Economic Choice Models”, working paper.

- [27] Fox, Jeremy T. and Amit Gandhi (2009b), “Full Identification in the Generalized Selection Model”, working paper.
- [28] Gautier, Eric and Kitamura, Yuichi. (2008), “Nonparametric Estimation in Random Coefficients Binary Choice Models”, working paper.
- [29] Geweke, J. (1986), “Exact Inference in the Inequality Constrained Normal Linear Regression Model,” *Journal of Applied Econometrics*, Vol. 1, No. 2, pp. 127-141.
- [30] Heckman, J. and Singer, B. “Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data”, *Econometrica* 52(2), 271-320 .
- [31] Hoderlein, Stefan, Jussi Klemelä and Enno Mammen (2008), “Analyzing the Random Coefficient Model Nonparametrically”, working paper.
- [32] Hoeffding, W. (1963), “Probability Inequalities for Sums of Independent Random Variables”, *Journal of the American Statistical Association*, 58, 13-30.
- [33] Huber, J. (2004) *Robust Statistics*, Wiley.
- [34] Ichimura, Hidehiko and T. Scott Thompson (1998), “Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution,” *Journal of Econometrics*, 86(2), 269–295.
- [35] Judd, K. L. (1998), *Numerical Methods in Economics*. MIT Press.
- [36] Judge, G.G. and Takayama, T. (1966), “Inequality Restrictions in Regression Analysis”, *Journal of the American Statistical Association*, Vol. 61, No. 313, pp. 166-181.
- [37] Kamakura, W.A. (1991), “Estimating flexible distributions of ideal-points with external analysis of preferences”, *Psychometrika*, 56, 3, 419-431.
- [38] Krein, M.G. and A.A. Nudel'man (1973, translation 1977), *The Markov moment problem and extremal problems: ideas and problems of P. L. Čebyšev and A. A. Markov and their further development*. Translations of mathematical monographs v. 50, American Mathematical Society, Providence.
- [39] Krommer, A.R. and C.W. Ueberhuber (1998), *Computation Integration*, SIAM.

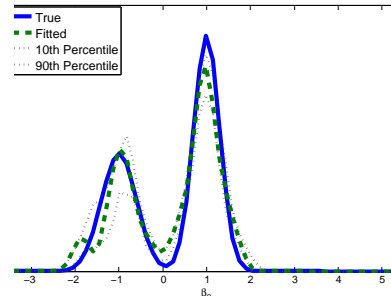
- [40] Laird, Nan (1978), “Nonparametric Maximum Likelihood Estimation of a Mixing Distribution”, *Journal of the American Statistical Association*, Vol. 73, No. 364, pp. 805–811.
- [41] Lehmann, E.L. and J.P. Romano (2005), *Testing Statistical Hypotheses*, 3rd Ed. Springer.
- [42] Lewbel, Arthur. (2000) “Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables”, *Journal of Econometrics*, 97, 1, 145–177.
- [43] Li, Jonathan Q. and Andrew R. Barron (2000), “Mixture density estimation”, *Advances in Neural Information Processing Systems*, Vol. 12, pp. 279–285.
- [44] Liew, C.K. (1976), “Inequality Constrained Least-Squares Estimation”, *Journal of the American Statistical Association*, Vol. 71, No. 355, pp. 746–751.
- [45] Lindsay, B.G. (1983) “The Geometry of Mixture Likelihoods: A General Theory”, *The Annals of Statistics*, 11(1), 86–94.
- [46] Manski, Charles F. (1975) “Maximum Score Estimation of the Stochastic Model of Choice”, *Journal of Econometrics*, 3(3), 205–228.
- [47] McFadden, Daniel and Kenneth Train (2000), “Mixed MNL models for discrete response”, *Journal of Applied Econometrics*, 15(5): 447–470.
- [48] Nevo, Aviv. 2001, “Measuring Market Power in the Ready-to-Eat Cereal Industry”, *Econometrica*, 69(2): 307–342.
- [49] Newey, W.K. (1997), “Convergence Rates and Asymptotic Normality for Series Estimators”, *Journal of Econometrics* 79, 147–168.
- [50] Parthasarathy, K.R. (1967), *Probability Measures on Metric Spaces*, Academic Press.
- [51] Petrin, Amil (2002), “Quantifying the Benefits of New Products: The Case of the Minivan”, *Journal of Political Economy*, 110:705–729, 2002.
- [52] Petrin, Amil and Kenneth Train (2009), “Control Function Corrections for Omitted Attributes in Differentiated Products Markets”, *Journal of Marketing Research*.
- [53] Quandt, R.E. and Ramsey, J.B. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, 73, 364, 730–738.

- [54] Rossi, Peter E., Greg M. Allenby, and Robert McCulloch (2005), *Bayesian Statistics and Marketing*. West Sussex: John Wiley & Sons.
- [55] Rust, John (1987), "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher", *Econometrica*, 55(5): 999–1033.
- [56] Rust, John (1994), "Structural Estimation of Markov Decision Processes", in *Handbook of Econometrics*, vol. 4, edited by Robert F. Engle and Daniel L. McFadden. Amsterdam: North-Holland.
- [57] Rust, John (1997), "Using randomization to break the curse of dimensionality", *Econometrica*, 65, 3, 487–516.
- [58] Shohat, J.A. and Tamarkin, J.D. (1943), *The Problem of Moments*, American Mathematics Society, Providence, RI.
- [59] Teicher, H. (1963), "Identifiability of Finite Mixtures", *Annals of Mathematical Statistics*, 34, 1265-1269.
- [60] Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso", *Journal of the Royal Statistical Society, Series B*. 58, 267-288.
- [61] Timan, A.F. (1963), *Theory of Approximation of Functions of a Real Variable*, MacMilan, New York.
- [62] Verbeek, J.J., N. Vlassis and B. Kröse (2003), "Efficient Greedy Learning of Gaussian Mixture Models", *Neural Computation*, Vol. 15, pp 469–485.
- [63] Wolak, F. "Testing inequality constraints in linear econometric models", *Journal of Econometrics*, Elsevier, vol. 41(2), pages 205-235, June 1989.
- [64] Yakowitz, S.J. and Spragins, J.D. (1968) "On the identifiability of finite mixtures", *The Annals of Mathematical Statistics*, pages 209-214.

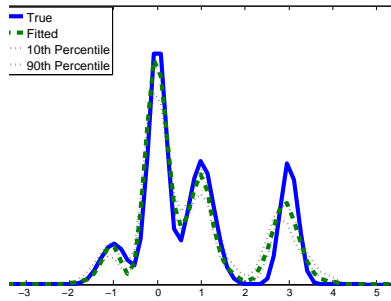
Figure 1: True, Estimated, and 90% Sampling Intervals for Marginal Densities from the Monte Carlo Experiment



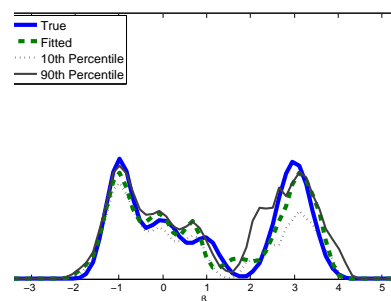
Density of  $\beta_1$ : Two Normal Mixtures



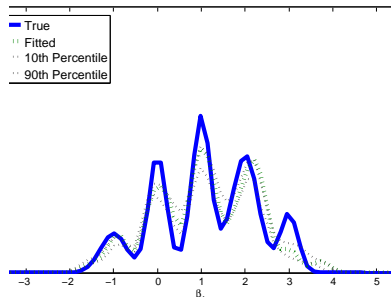
Density of  $\beta_2$ : Two Normal Mixtures



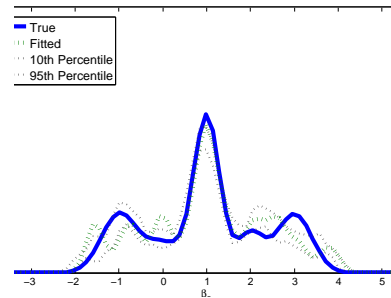
Density of  $\beta_1$ : Four Normal Mixtures



Density of  $\beta_2$ : Four Normal Mixtures

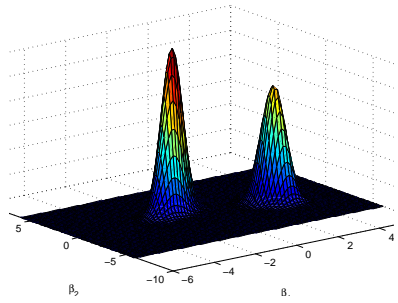


Density of  $\beta_1$ : Six Normal Mixtures

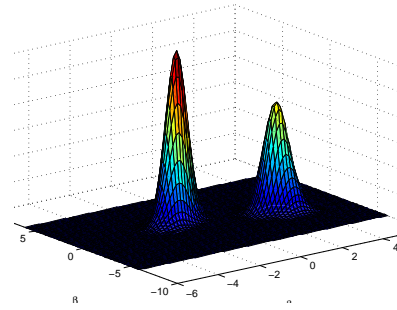


Density of  $\beta_2$ : Six Normal Mixtures

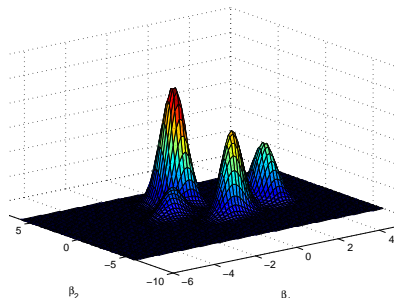
Figure 2: True and Estimated Joint Densities from the Monte Carlo Experiment



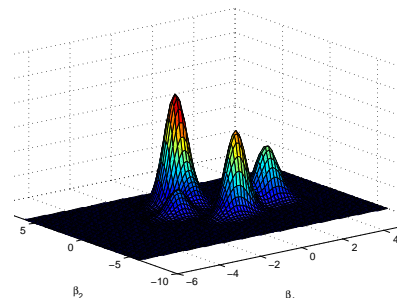
Two Normal Mixtures: Truth



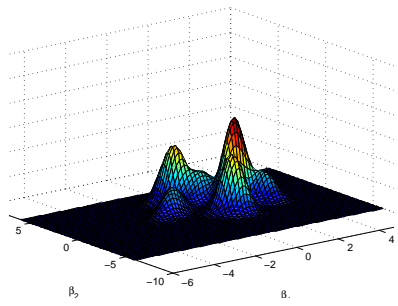
Two Normal Mixtures: Estimated



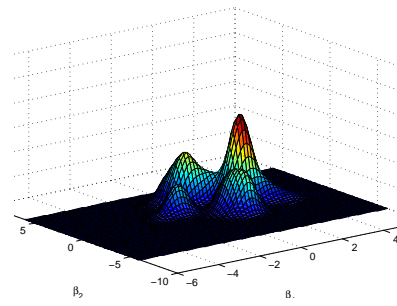
Four Normal Mixtures: Truth



Four Normal Mixtures: Estimated



Six Normal Mixtures: Truth



Six Normal Mixtures: Estimated

Figure 3: Estimated Distributions of Heterogeneity for the Benefits of Staying Home

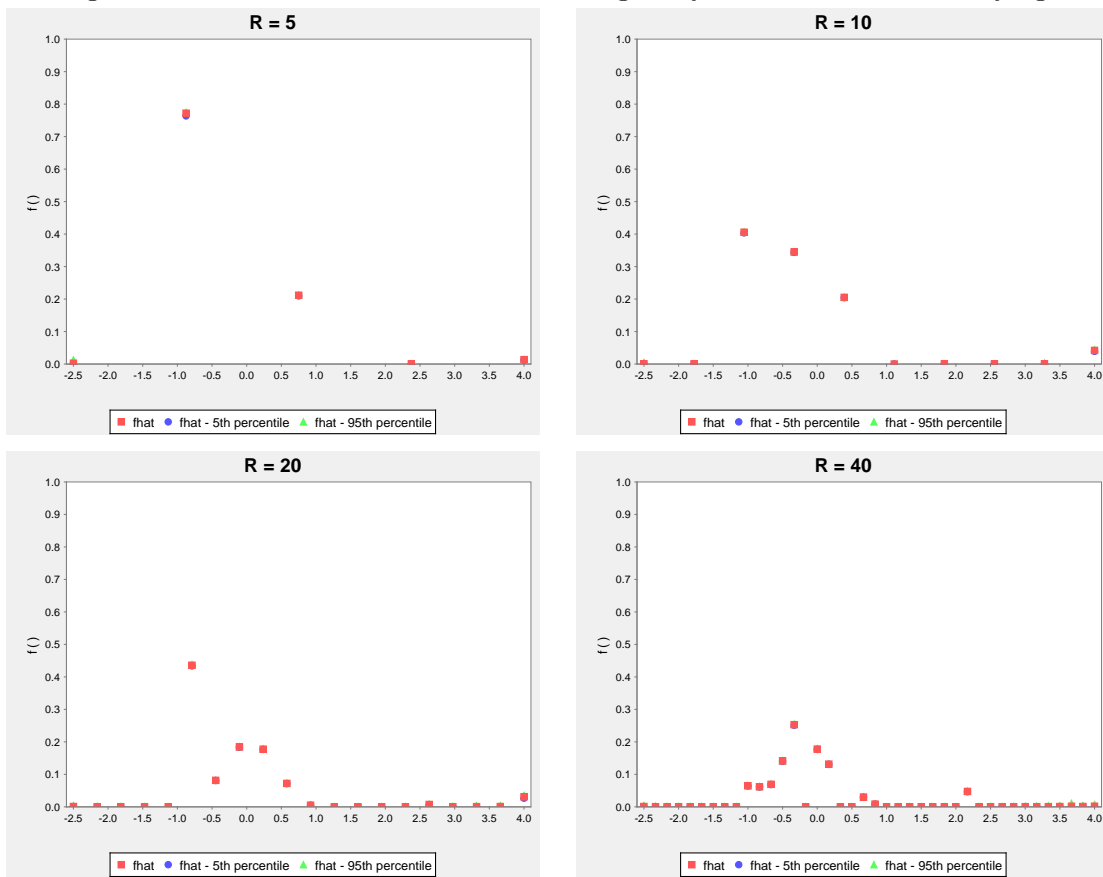


Figure 4: Predicted Distribution of Days Worked in Months Where First Five Days Used for Estimation

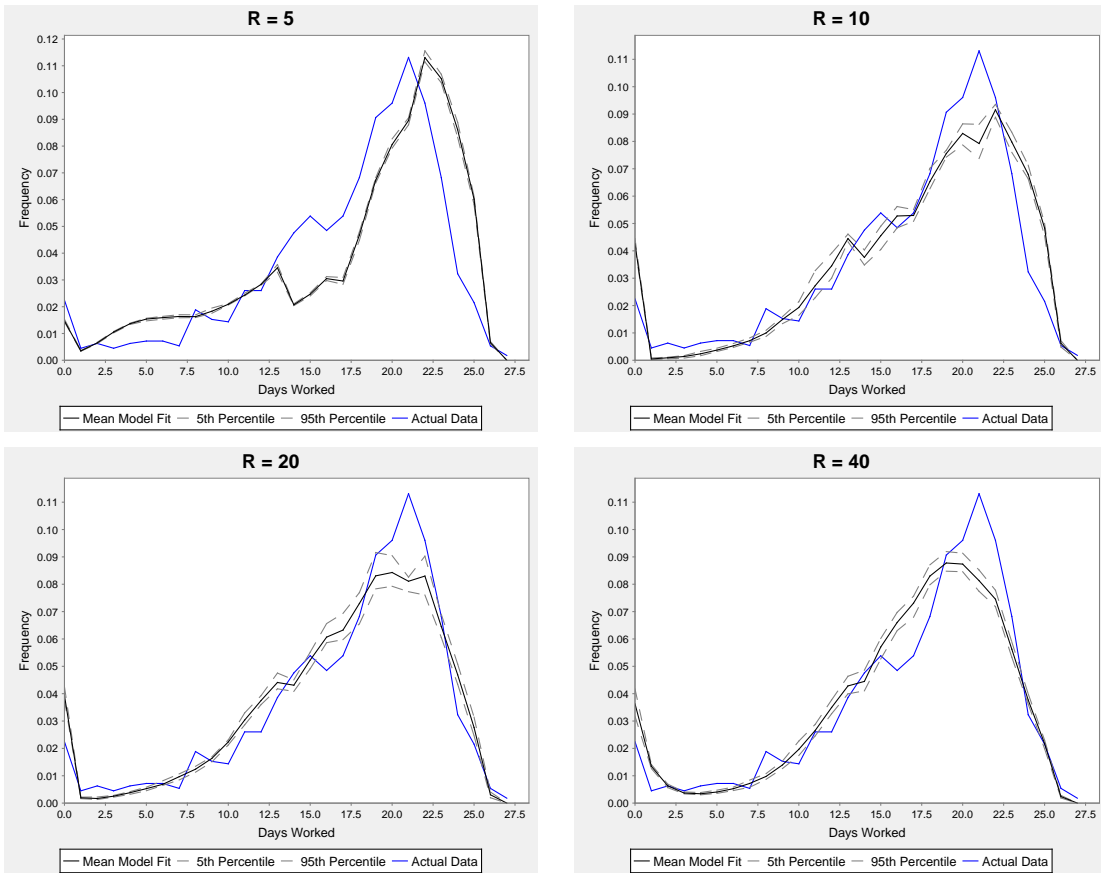


Figure 5: Predicted Distribution of Days Worked Under Out-of-Sample Compensation Scheme

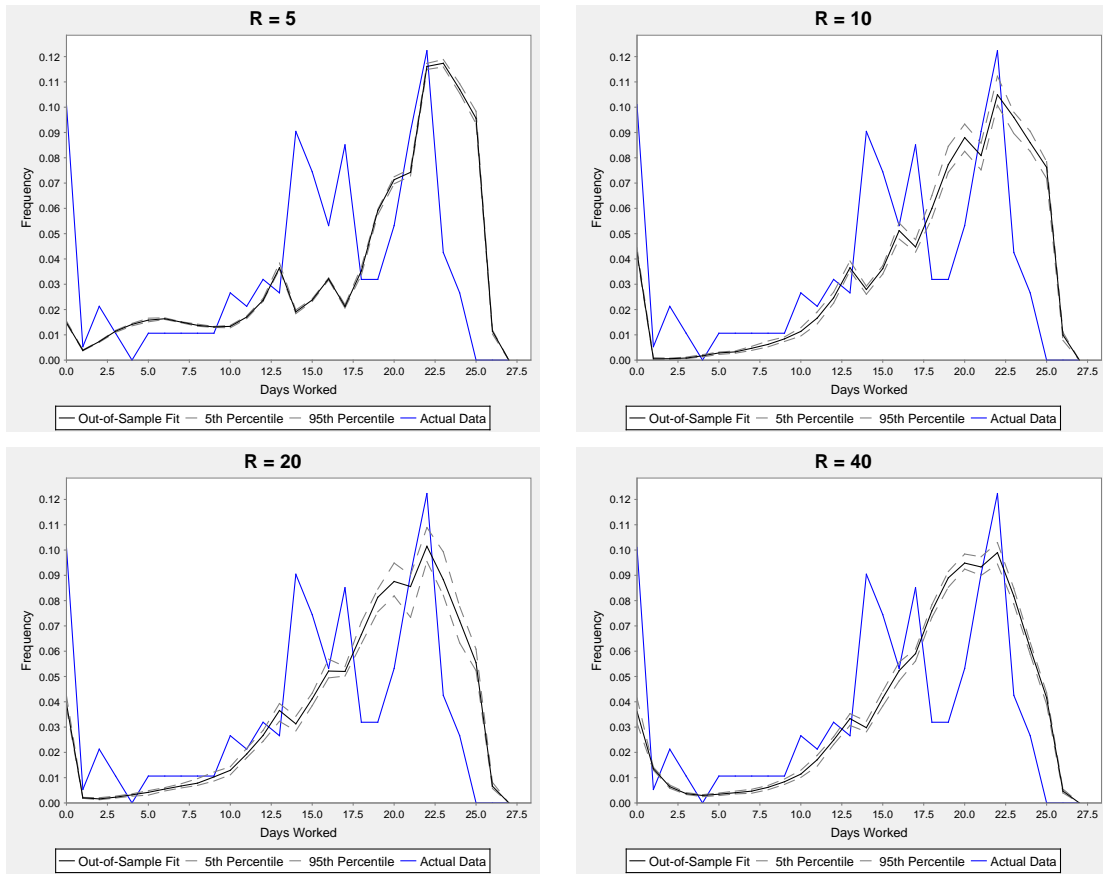


Table 1: Market-Share Approximations and the Number of Positive Weights, for Monte Carlo

| <b>Design</b>    | <b>RMSE</b> |          |          | <b># of positive weights</b> |     |     |
|------------------|-------------|----------|----------|------------------------------|-----|-----|
|                  | Mean        | Min      | Max      | Mean                         | Min | Max |
| Normal Mixture 2 | 0.000096    | 0.000042 | 0.000185 | 32                           | 17  | 73  |
| Normal Mixture 4 | 0.000123    | 0.000045 | 0.000227 | 70                           | 36  | 125 |
| Normal Mixture 6 | 0.000080    | 0.000043 | 0.000130 | 99                           | 62  | 154 |