A simple estimator for the distribution of random coefficients

Jeremy T. Fox
University of Michigan and NBER

Kyoo il Kim
University of Minnesota

Stephen P. Ryan
MIT and NBER

Patrick Bajari
University of Minnesota and NBER

We propose a simple 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 random coefficient models. For complex structural models, one does not need to nest a solution to the economic model during optimization. We present a Monte Carlo study and an empirical application to dynamic programming discrete choice with a serially correlated unobserved state variable.

Keywords. Random coefficients, mixtures, demand, logit, mixed logit, dynamic programming, teacher labor supply.

JEL classification. C25, C14, L.

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 hetero-

Jeremy T. Fox: jtfox@umich.edu
Kyoo il Kim: kyookim@umn.edu
Stephen P. Ryan: sryan@mit.edu
Patrick Bajari: bajari@econ.umn.edu

Fox thanks the National Science Foundation, the Olin Foundation, and the Stigler Center for generous funding. Bajari thanks the National Science Foundation, Grant SES-0720463, for generous research support. Thanks to helpful comments from seminar participants at the AEA meetings, Chicago, Chicago GSB, UC Davis, European Commission Antitrust, the Far East Econometric Society meetings, LSE, Mannheim, MIT, Northwestern, Paris I, Quantitative Marketing and Economics, Queens, Rochester, Rutgers, Stanford, Stony Brook, Toronto, UCL, USC, UT Dallas, Virginia, and Yale. Thanks to comments from Xiaohong Chen, Andrew Chesher, Philippe Fevrier, Amit Gandhi, Han Hong, David Margolis, Andrés Musalem, Peter Reiss, Jean-Marc Robin, Andrés Santos, Azeem Shaikh, and Harald Uhlig. Thanks to Chenchuan Li for research assistance.

DOI: 10.3982/QE49

geneity in preferences. A classic example is that consumers may have heterogeneous preferences over a set of product characteristics in an industry with differentiated products. A growing econometric literature has addressed heterogeneity by providing estimators that allow the coefficients of the economic model to vary across agents. In this paper, we describe a general method for estimating such *random coefficient* models that is easy to compute. Our estimator exploits a reparametrization of the underlying model so that the parameters enter linearly. This is in contrast to previous approaches in the literature, such as the expectation-maximization (EM) algorithm, Markov chain Monte Carlo (MCMC), simulated maximum likelihood, simulated method of moments, and minimum distance, which are highly nonlinear. Linearity simplifies the computation of the parameters. Our approach also avoids nested solutions to economic models in optimization, which gives a strong computational advantage for complex structural models, as we will discuss.

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 θ^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

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

The key insight of our approach is that the dependent variable in our model, $\Pr(y = 1 \mid x)$, is linearly related to the model parameters θ^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. We demonstrate that the θ^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.

Many alternative estimators require computationally expensive nonlinear optimization, and as a result researchers frequently use tightly specified distributions of heterogeneity 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 a parsimonious family of distributions.

We show how to extend this simple intuition to a more general framework. 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 the model parameters. Also, 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 location and the scale of the parameter space. Further, we introduce a cross-validation method for picking the grid points.

Importantly, we discuss how our approach can be extended to more complex, structural economic choice models. As an example, we show how to estimate a distribution of random coefficients in a dynamic programming, discrete choice model such as Rust (1987). The computational issues in traditional simulation estimators applied to static models are more severe for dynamic models when the dynamic programming problem must be solved for each realization of the random coefficients in each call of the objective function by the optimization routine. If there are S simulation draws in a traditional simulation estimator and the optimization routine requires L evaluations of the objective function to achieve convergence, the dynamic programming problem must be solved $S \cdot L$ times in the traditional approach. In our estimator, the dynamic program must be solved only R times, when R is the number of support points for the random coefficients. In our dynamic programming empirical application to teacher labor supply in India, we demonstrate that our approach can dramatically reduce the computational burden of these models.

Our estimator is a general mixtures estimator. A frequent application of mixtures estimators in the literature has been demand estimation. In the paper, we discuss some strengths and weaknesses of our approach relative to some other approaches proposed in the demand estimation literature. First, if the researcher has aggregate data on market shares with price endogeneity, the approach of Berry, Levinsohn, and Pakes (1995) will impose fewer assumptions on the supply side of the model. Our approach can accommodate aggregate data, even if market shares are measured with error, which Berry et al. do not allow. However, we need to specify a reduced-form pricing function analogous to the pricing functions in Kim and Petrin (2010) and especially Fox and Gandhi (2010) in order to account for price endogeneity (or adapt the arguments in Berry and Haile (2010b) and Fox and Gandhi (2011b)). It is not possible to prove the existence of the particular pricing function that we use from general supply-side assumptions. Second, in small samples, Bayesian methods such as Rossi, Allenby, and McCulloch (2005) and Burda, Harding, and Hausman (2008) that use prior information will likely have better finite-sample performance, at the computational expense of having to evaluate the objective function many times. Our methods are intended for applications where the researcher has access to large sample sizes.

The most common frequentist, mixtures 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 as well as the difficulty in diagnosing convergence (Seidel, Mosler, and Alker (2000), Verbeek, Vlassis, and Kröse (2003), Biernacki, Celeux, and Govaert (2003), Karlis and Xekalaki (2003)).[1] Further, the EM algorithm has a slow rate of convergence even when it does converge to a global solution (Pilla and Lindsay (2001)). Li and Barron (2000) introduced another alternative, but again our approach is computationally simpler. Our estimator is also computationally simpler than the minimum distance estimator of Beran and Millar (1994), which in our experience often has an objective function with an intractably large number of local minima. The discrete-grid idea (called the histogram approach) is found outside of economics in Kamakura (1991), who used a discrete grid to estimate an ideal-point model; he did not discuss any of our extensions. Of course, mixtures themselves have a long history in economics, such as Quandt and Ramsey (1978).

The outline of our paper is as follows. In Section 2, we describe our model and baseline estimator. We focus on the motivating example of the multinomial logit with individual data. Section 3 discusses picking the grid of points and introduces our cross-validation approach. Section 4 provides various extensions to the baseline estimator. Section 5 provides examples of discrete choice with aggregate data, dynamic programming discrete choice, mixed continuous and discrete choice (selection), and discrete choice models with endogenous regressors. We include a large number of examples so as to motivate our estimator for applied readers. In Section 6, we discuss how to conduct inference. 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 Duflo, Hanna, and Ryan (forthcoming). The dynamic programming problem has a serially correlated, unobserved state variable.

2. The baseline estimator

2.1 *General notation*

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

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

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

a modeling primitive. In our simple example in the Introduction, g_A corresponds to the logit choice probability for a given coefficient β. The cumulative distribution function (CDF) of the random coefficients is denoted as $F(\beta)$. We restrict our discussion of consistency and asymptotic inference to models with a finite number R of support points β^r. The weight on the support point β^r is $\theta^r \geq 0$, where $\sum_{r=1}^R \theta_r = 1$. Given these definitions, it follows that

$$P_A(x) = \sum_{r=1}^{R} \theta^r g_A(x, \beta^r). \tag{1}$$

On the right hand side of the above equation, $g_A(x, \beta)$ gives the probability of A conditional on x and β. We average over the distribution of β using the CDF $F(\beta) = \sum_{r=1}^R \theta^r 1[\beta^r \leq \beta]$, where $1[\beta^r \leq \beta] = 1$ when $\beta^r \leq \beta$, to arrive at $P_A(x)$, the population probability of the event A conditional on x.

In our framework, the object the econometrician wishes to estimate is the tuple $(\theta^1, \dots, \theta^R)$, the weights on the support points. For this paper, we assume the support points are known (or known up to location and scale parameters in Section 4.3) and focus on the computational advantages of our approach.

2.2 *The multinomial logit model*

As we discussed in the Introduction, one motivating example for our paper is the logit with random coefficients. We begin by discussing this example in detail. In Section 5, we show how our approach extends to other random coefficient models, including dynamic programming discrete choice and demand models where the dependent variable is represented by a vector of discrete and continuous variables. The estimation method that we propose can, in principle, be applied to any model that can be written in the form (1). Indeed, in discrete choice with large samples, we can completely do away with the logit errors and allow the distribution of the choice- and consumer-specific shocks to be estimated, which is a special case of the identification results on multinomial choice in Fox and Gandhi (2010). Our examples start with the logit structure only for expositional clarity, in part because the logit is commonly used in empirical work.

In the logit model, agents $i = 1, \dots, N$ can choose between $j = 1, \dots, J$ mutually exclusive alternatives and one outside good (good 0). 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 nonprice product characteristics, the price of good j, and the demographics of agent i. We let $x_i = (x_{i,1}', \dots, x_{i,J}')$ denote the stacked vector of the J $x_{i,j}$'s, the observable characteristics.

In the model, there are $r = 1, \dots, R$ types of agents. The unobservable preference parameters of type r are equal to the $K \times 1$ vector β^r. We discuss how to choose the types β^r below. For the moment, assume that the β^r are fixed and exogenously specified. As in our example in the Introduction, it might be helpful to think of the β^r being defined using a fixed grid on a compact set. The random variable β is distributed independently of x. The probability of type r in the population is θ^r. Let $\theta = (\theta^1, \dots, \theta^R)'$ denote the corresponding vector. The θ must lie on the unit simplex, or

$$\sum_{r=1}^{R} \theta^r = 1, \tag{2}$$

$$\theta^r \geq 0. \tag{3}$$

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

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

There is an outside good with utility $u_{i,0} = \varepsilon_{i,0}$. Assume that $\varepsilon_{i,j}$ is distributed as Type I extreme value and is independent of x_i and β^r. Agents in the model are assumed to be utility maximizers. The observable dependent variable $y_{i,j}$ is generated as

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

Let y_i be the vector $(y_{i,1}, \dots, y_{i,J})$. The probability of type r picking choice j at x_i is (using $A = \{j\}$ in the general notation above)

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

It then follows that the conditional probability of observing that agent i selects choice j is

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

2.3 *Linear regression*

We study the estimation problem of recovering θ given the researcher has $i = 1, \dots, N$ observations on (x_i, y_i). A simple method for estimating the parameters θ 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

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

Define the $R \times 1$ vector $z_{i,j} = (z_{i,j,1}, \dots, z_{i,j,R})'$ with individual elements $z_{i,j,r} = g_j(x_i, \beta^r)$. Recall that β^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 fixed regressor. Next, by the definition of a choice probability,

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

This implies that the following ordinary least-squares problem produces a consistent estimator of the θ^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 $\hat{\theta} = (Z'Z)^{-1}Z'Y$.

Equation (5) implies that the standard mean independence condition is satisfied for least squares. The least-squares estimator has a unique solution as long as Z has rank R. The estimator is consistent under standard assumptions for least squares.

2.4 *Inequality constrained linear 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 0 and 1. Also, this may be important if the economist wishes to interpret the distribution of β as a structural parameter. Our baseline estimator estimates θ using inequality constrained least squares,

$$\begin{aligned} &\hat{\theta} = \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} (y_{i,j} - z'_{i,j}\theta)^2 \\ &\quad \text{subject to (2) and (3).} \end{aligned} \tag{6}$$

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

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

where $1[\beta^r \leq \beta] = 1$ when $\beta^r \leq \beta$. Thus, we have a structural estimator for a distribution of random parameters in addition to a method for approximating choice probabilities. We call the inequality-constrained least-squares estimator our baseline estimator to distinguish it from several extensions below. This baseline estimator is consistent under standard regularity conditions (see, e.g., Newey and McFadden (1994) and Andrews (2002)).[2]

[2] Linear inequality constraints restrict only the parameter space of θ. Whether the estimator $\hat{\theta}$ is on a boundary or is in the interior of the parameter space does not create any complications for establishing consistency (Andrews (2002)). Standard sufficient conditions for the consistency of extremum estimators that include our estimators are available in the literature.

3. Picking the grid of basis points

The approach just presented is computationally simple: we can always find a global optimum and we avoid many evaluations of complex structural models, as we discuss in Section 5.1. The main disadvantage of the baseline approach is that the estimates may be sensitive to the choice of the grid of points $\mathcal{B}_R = (\beta^1, \dots, \beta^R)$. A natural first question to ask about the baseline estimator is how to pick the number of grid points R as a function of the sample. While our consistency arguments in this paper apply to a fixed R, in practice we still need to choose R. Conditional on picking R, how should one pick the exact grid of points $\mathcal{B}_R$? This section addresses these questions. First we present advice based on our experiences using the estimator in Monte Carlo experiments. Second, we adopt a data-driven cross-validation approach. In addition, Section 4.3 introduces an alternative to cross-validation: a location and scale model for when the support of the random coefficients is not known. We focus on the baseline estimator here.

3.1 *Advice from Monte Carlos*

We have run many Monte Carlo studies in the process of developing our estimator. In Monte Carlo experiments in Section 7, we set $R = C^{\dim(x_j)}$ with $\dim(x_j) = 2$ for some constants C. Although we find that typically more of the R weights are estimated to be nonzero when the true F looks, roughly speaking, to be more complicated, our Monte Carlo experiments show that a relatively small R can give a good approximation to $F(\beta)$, when N is large.

A drawback of the baseline estimator is that the researcher has to have some sense of the support of the random coefficients. In the logit, the econometrician may use a preliminary estimate of a logit model with fixed coefficients to determine a region of support for the random coefficients. For example, the econometrician may center the grid for the β^r at the initial logit estimates and take some multiple of their confidence intervals as the support or he or she can let the support for estimation grow as the sample size grows.

A second, related approach to picking a support is that the econometrician may experiment with alternative sets of grid points to see how the choice of grid points influences estimation results. A limitation of this approach is that it introduces a pretest bias and the standard errors in Section 6 will need to be adjusted. For example, a procedure where a researcher uses a diffuse grid with a wide support and then reestimates the model after increasing the detail of the grid where mass appears to be located, is, in its full statistical structure, a two-step estimator. Rigorous standard errors should account for both of the estimation steps. This approach is, however, acceptable when a researcher focuses on the approximation of the random coefficients distribution but not its inference.

Given a choice of support, we experimented with Halton and Weyl sequences; they tend to improve Monte Carlo performance over evenly spaced grids. We report in Section 7 that random grids did not work as well as evenly spaced grids in our investigations.

3.2 *Cross-validation*

Cross-validation is a technique often employed in econometrics when overfitting may be an issue. The idea is to minimize the mean squared error of the final estimator by comparing the out-of-sample fit of parameter estimates using candidate grid choices on training data sets to holdout samples. The mean squared error accounts for both bias (not being able to capture the true distribution F_0 because it is not in the approximating space) and variance (statistical sampling error from a finite sample N of data). Our estimator, like most others, has a bias/variance trade-off: choosing a higher R makes the grid $\mathcal{B}_R$ more flexible and so lowers the bias, but there are more parameters to estimate and variance increases. Cross-validation seeks to balance bias and variance in a particular data set. We introduce the method here, but in the interests of conciseness do not develop the statistical properties formally. Unfortunately, there are no theoretical results available for using cross-validation for estimators of distributions, other than a few exceptions using kernels (Hansen (2004)).

Let $\tilde{\mathcal{B}}$ be a hypothetical choice of a grid of points. This notation includes the number of points, the support of the points, and the method for picking the points within the support. Given a grid $\tilde{\mathcal{B}}$, the mean squared error (MSE) of the fit to choice probabilities and market shares for a problem with J discrete outcomes is

$$\mathrm{MSE}_N(\tilde{\mathcal{B}}) \equiv E\left[\sum_{j=1}^{J}\bigl\{P_j(x,F_0)-P_j(x,\hat{F}^N_{\tilde{\mathcal{B}}})\bigr\}^2\right], \tag{7}$$

where $\hat{F}^N_{\tilde{\mathcal{B}}}$ is the estimator of the distribution function with N data points and the grid choice $\tilde{\mathcal{B}}$. This criterion focuses on fitting choice probabilities.[3]

We need to approximate the MSE in (7) so as to implement cross-validation. We adopt a variant of cross-validation called *leave one partition out* or I-fold cross-validation. We first partition the data into I subsets N_ι, so that $N_\iota \cap N_{\iota'} = \emptyset$ for $\iota \neq \iota'$ and $\bigcup_{\iota=1}^{I} N_\iota = \{1, 2, \dots, N\}$. To evaluate the performance of a grid $\tilde{\mathcal{B}}$, we estimate the distribution of random coefficients using each of the I samples where one of the partitions is left out; in other words, we use each of the samples $\{1, 2, \dots, N\}\backslash N_\iota$ as training samples and use the corresponding N_ι as holdout samples. Let $\hat{F}^{-\iota,N}_{\tilde{\mathcal{B}}}$ be the estimate leaving out the ιth sample N_ι; there are I such estimates.

For $i \in N_\iota$, the estimate $\hat{F}^{-\iota,N}_{\tilde{\mathcal{B}}}$ not using the observations in N_ι, and the assumption that the observations are independent and identically distributed (i.i.d.),

$$\begin{aligned}
&E\left[\sum_{j=1}^{J}\bigl\{P_j(x_i,F_0)-P_j(x_i,\hat{F}^{-\iota,N}_{\tilde{\mathcal{B}}})\bigr\}^2\right]\\
&\quad=\sum_{j=1}^{J}E\bigl[\{P_j(x_i,F_0)-y_{i,j}\}^2+2y_{i,j}(P_j(x_i,F_0)-y_{i,j})
\end{aligned}$$

[3] In practice, we can also apply a cross-validation approach based on the distance between F_0 and $\hat{F}_N$. We do not present this version for conciseness. In our Monte Carlo in Section 7, we assess estimation error in part by the root mean integrated squared error of $\hat{F}_N$.

$$
\begin{aligned}
&\quad - 2(P_j(x_i, F_0) - y_{i,j})P_j(x_i, \hat{F}_{\tilde{\mathcal{B}}}^{-\iota,N})\big] \\
&\quad + \sum_{j=1}^{J} E\big[\{y_{i,j} - P_j(x_i, \hat{F}_{\tilde{\mathcal{B}}}^{-\iota,N})\}^2\big] \qquad (8)\\
&= \sum_{j=1}^{J} E\big[\{P_j(x_i, F_0) - y_{i,j}\}^2 + 2y_{i,j}(P_j(x_i, F_0) - y_{i,j})\big] \\
&\quad + \sum_{j=1}^{J} E\big[\{y_{i,j} - P_j(x_i, \hat{F}_{\tilde{\mathcal{B}}}^{-\iota,N})\}^2\big]
\end{aligned}
$$

where we have $E[2(P_j(x_i, F_0) - y_{i,j})P_j(x_i, \hat{F}_{\tilde{\mathcal{B}}}^{-\iota,N})] = 0$ by the law of iterated expectations, because $P_j(x_i, \hat{F}_{\tilde{\mathcal{B}}}^{-\iota,N})$ does not depend on $y_{i,j}$, $i \in N_\iota$, and because $E[P_j(x_i, F_0) - y_{i,j} \mid x_i] = 0$ by the definition of a choice probability. Also note that the first term in (8) does not depend on $\tilde{\mathcal{B}}$ and that we can consistently estimate the second term in (8) using the sample mean of the observations in the holdout sample. Therefore, the optimal choice of grid $\tilde{\mathcal{B}}$ from cross-validation is

$$
\hat{\mathcal{B}}_N = \arg\min_{\tilde{\mathcal{B}}} \frac{1}{N} \sum_{\iota=1}^{I} \sum_{i \in N_\iota} \sum_{j=1}^{J} \{y_{i,j} - P_j(x_i, \hat{F}_{\tilde{\mathcal{B}}}^{-\iota,N})\}^2, \qquad (9)
$$

because any $\hat{\mathcal{B}}_N$ that minimizes the above criterion also approximately minimizes the MSE in (7).

For a given candidate grid $\tilde{\mathcal{B}}$, the computational cost of our cross-validation procedure is that the estimator must be computed I times, one for each partition. In empirical work using cross-validation, researchers often pick a small I, such as $I = 10$. Cross-validation using our estimator is more computationally attractive than cross-validation using the other estimators in the Introduction simply because the run time of our estimator is low and because particular routines are guaranteed to find the global minimum to the problem (6). Many estimators have bias/variance trade-offs, but cross-validation is only practical in methods where the estimator is fast and reliable.

In practice, a grid $\tilde{\mathcal{B}}$ is a high dimensional object. It may make more sense to optimize (9) over only the number of support points R and the boundaries of the support. Some other scheme, such as Weyl or Halton sequences, can be used to pick points once the number of points and the support have been chosen. Also, instead of performing a formal optimization over grids, it may make computational sense to select only finite L candidates $\tilde{\mathcal{B}}_1, \dots, \tilde{\mathcal{B}}_L$ and pick the one that gives the lowest value for the criterion in (9).

4. Extensions to the estimator

In this section, we introduce four extensions of our baseline estimator that may be of some practical use to applied researchers.

4.1 *Generalized least squares*

We discuss confidence regions for our baseline estimator in Section 6. Our estimator will be more efficient if we use generalized least squares and model both the heteroskedasticity in the linear probability model and the correlation across multiple regression observations corresponding to the same statistical observation. Consider the logit example where $\Pr(y_{i,j} = 1 \mid x_i) = \sum_{r=1}^{R} \theta^r g_j(x_i, \beta^r)$. Using the formulation in Section 2.3, there are J regression observations for each statistical observation i, one for each choice. Based on the properties of the multinomial distribution, the conditional variance of $y_{i,j}$ is $\psi_i^j = \Pr(y_{i,j} = 1 \mid x_i) \cdot (1 - \Pr(y_{i,j} = 1 \mid x_i))$ and the covariance between $y_{i,j}$ and $y_{i,k}$, $k \neq j$, is $\psi_i^{j,k} = -\Pr(y_{i,j} = 1 \mid x_i) \cdot \Pr(y_{i,k} = 1 \mid x_i)$. Given this, the optimal generalized least-squares weighting matrix is the inverse of the block diagonal matrix

$$\Psi = \begin{bmatrix} \psi_1^1 & \psi_1^{1,2} & \cdots & \psi_1^{1,J} & 0 & 0 & 0 & \cdots & 0 \\ \psi_1^{1,2} & \psi_1^2 & \cdots & \psi_1^{2,J} & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & 0 & 0 & \cdots & 0 \\ \psi_1^{1,J} & \psi_1^{2,J} & \cdots & \psi_1^J & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \psi_N^1 & \psi_N^{1,2} & \cdots & \psi_N^{1,J} \\ 0 & 0 & 0 & 0 & 0 & \psi_N^{1,2} & \psi_N^2 & \cdots & \psi_N^{2,J} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \psi_N^{1,J} & \psi_N^{2,J} & \cdots & \psi_N^J \end{bmatrix}.$$

In feasible generalized least squares, one first estimates the θ^r's using the baseline estimator and then uses the first-stage estimates to estimate $\Pr(y_{i,j} = 1 \mid x_i)$, which itself is used to estimate Ψ. The generalized least-squares estimator uses the stacked matrices from Section 2.3 to minimize

$$(Y - Z\theta)' \Psi^{-1} (Y - Z\theta)$$

subject to the constraints (2) and (3).

4.2 *Smooth basis densities*

A limitation of the baseline estimator is that the CDF of the random parameters will be a step function. In applied work, it is often attractive to have a smooth distribution of random parameters. In this subsection, we describe one approach to estimating a density instead of a CDF. Instead of modeling the distribution of the random parameters as a mixture of point masses, we model the density as a mixture of normal densities.

Let a basis r be a normal distribution with mean the $K \times 1$ vector μ^r and standard deviation the $K \times 1$ vector σ^r. Let $N(\beta_k \mid \mu_k^r, \sigma_k^r)$ denote the normal density of the kth random parameter. Under normal basis functions, the joint density for a given r is just the product of the marginals or

$$N(\beta \mid \mu^r, \sigma^r) = \prod_{k=1}^{K} N(\beta_k \mid \mu_k^r, \sigma_k^r).$$

Let θ^r denote the probability weight given to the rth basis, $N(\beta \mid \mu^r, \sigma^r)$. As in the baseline estimator, it is desirable to constrain θ to lie in the unit simplex, that is, (2) and (3).

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

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

We use the $\approx$ to emphasize the simulation approximation to the integrals. We can then estimate θ 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 linear least squares, a globally convex optimization problem with an easily computed unique solution. The resulting density estimate is $\hat{f}(\beta) = \sum_{r=1}^{R} \hat{\theta}^r N(\beta \mid \mu^r, \sigma^r)$. If the true distribution has a smooth 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.[4]

4.3 *Choice of support and location scale model*

In many applications, the econometrician may not have good prior knowledge about the support region where most of the random coefficients β^r lie. This is particularly true in models where the covariates are high dimensional. We discussed cross-validation as a data-driven response to this problem earlier.

Another approach is to introduce location and scale parameters. To illustrate the idea, 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 the unit interval. We include a set of location and scale parameters a_k

[4] The consistency of this estimator has not been established, but it works well in Monte Carlo studies reported in an earlier draft.

and b_k, $k = 1, \dots, K$, and define the rth random coefficient for the kth characteristic as $a_k + b_k \beta_k^r$.

In numerical optimization, we now search over $R + 2K$ parameters corresponding to θ, and $a = (a_1, \dots, a_K)'$ and $b = (b_1, \dots, b_K)'$. The choice probabilities predictions for type r are

$$g_j(x_i, a + b\beta^r) = \frac{\exp\left(\sum_{k=1}^{K} x_{k,i,j}(a_k + b_k \beta_k^r)\right)}{1 + \sum_{j'=1}^{J} \exp\left(\sum_{k=1}^{K} x_{k,i,j'}(a_k + b_k \beta_k^r)\right)},$$

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

$$\min_{a,b,\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R} \theta^r g_j(x_i, a + b\beta^r) \right)^2 \qquad (10)$$
$$\text{subject to (2) and (3).}$$

The appropriate MATLAB routine is lsqnonlin and there is no theorem that states that this nonlinear least-squares routine converges to a global minimum. Note that the model has $2K$ nonlinear parameters and the remaining R parameters still enter linearly. Also, a and b do not enter the constraints, which are still linear. Using exact, rather than numerical derivatives improves the speed and convergence of the estimator. The derivatives to our objective function can be expressed in closed form up to any order, at least for the logit example. One can also use a two-step profiling approach. In the first step, given trial values of a and b, obtain the profiled estimates of $\theta(a, b)$ as the $\hat{\theta}(a, b)$ that solves (10) for values of a and b. Then in the second step, obtain the estimates of a and b that minimize the objective function in (10) after replacing θ with $\hat{\theta}(a, b)$.

In numerical experimentation with the logit example, we have found the convergence of the estimator to be robust. However, finding the global optimum of the location and scale model likely becomes more challenging as the complexity of the underlying economic model increases. This contrasts with the estimator using a fixed grid, where the computational complexity of optimization is not related to the complexity of the underlying economic model.

In many applied problems, the number of characteristics K may be large and it may not be practical to estimate a K-dimensional distribution of random coefficients. Therefore, it may be desirable to only let a subset of the most important characteristics have random coefficients. Allowing for homogeneous parameters is possible by a trivial modification of the nonlinear least-squares formulation in (10).

4.4 *Imposing independence across random coefficients*

Often a researcher wants to impose independence across two sets of random coefficients to increase statistical precision. Estimating a K-dimensional joint distribution function requires more data than estimating K one-dimensional marginal distribution functions. As suggested by Ram Rao in a private conversation, one can decompose $\beta = (\beta^1, \beta^2)$ and estimate a set of weights $\theta^{r^1,1}$, $r^1 = 1, \dots, R^1$, on β^1 and $\theta^{r^2,2}$, $r^2 = 1, \dots, R^2$, on β^2. Hence, the predicted choice probability

$$\Pr(y_{i,j} = 1 \mid x_i) = \sum_{r^1=1}^{R^1} \sum_{r^2=1}^{R^2} \theta^{r^1,1} \theta^{r^2,2} g_j\bigl(x_i, \bigl(\beta^{r^1}, \beta^{r^2}\bigr)\bigr)$$

enters the minimization problem (6). As with the location and scale model, the parameters $\theta^{r^1,1}$ and $\theta^{r^2,2}$ now enter the model nonlinearly and nonlinear least squares will have to be used.

5. Additional examples of applications

Our estimator can, in principle, be applied to any setting where the model can be written in the form (1). Below we discuss some additional examples that fit our framework that may be of interest to applied researchers. The baseline estimator (including cross-validation) can be applied to these examples.

5.1 *Dynamic programming models*

Our approach can be applied to dynamic discrete choice models as in Rust (1987, 1994). We generalize the framework he considered by allowing for a distribution of random coefficients. Suppose that the flow utility of agent i in period t from choosing action j is

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

The error term $\varepsilon_{i,j,t}$ is a preference shock for agent i's utility to choice j at time period t. For simplicity, the error term is i.i.d. Type I extreme value across agents, choices, and time periods. Agent i's decision problem is dynamic because there is a link between the current and the future values of $x_{i,t} = (x'_{i,1,t}, \dots, x'_{i,J,t})$ through current decisions. Let $\pi(x_{i,t+1} \mid x_{i,t}, j_{i,t})$ denote the transition probability for the state variable $x_{i,t}$ as a function of the action of the agent, $j_{i,t}$. Here the transition rule does not involve random coefficients and we assume that it can be estimated in a first stage. If the transition rule were to include random coefficients, we could estimate the entire model in a one-step procedure, a slight generalization of the approach below. We would estimate a joint distribution of the random coefficients in the transition rule and in the choice model. We could estimate a distribution of random coefficients in the transition rule in a first stage if we are willing to assume that the random coefficients in the transition rule and choice model are independent.

The goal is to estimate $F(\beta)$, the distribution of the random coefficients. Again we pick R basis vectors β^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

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

where the scalar $\delta \in [0, 1)$ is a discount factor fixed by the researcher before estimation, as is usually done in empirical practice. Solving for the value function $V^r(x_{i,t})$ involves solving the Bellman equations as a system of equations, one for each discrete state $x_{i,t}$:

$$V^r(x_{i,t}) = \log\left(\sum_{j=1}^{J} \exp\bigl(x'_{i,j,t}\beta^r + \delta E[V^r(x_{i,t+1}) \mid x_{i,t}, j]\bigr)\right). \tag{11}$$

Solving this system of nonlinear equations can be done with a contraction mapping or other scheme for solving dynamic programs.

We use panel data on N panels of length T each. 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. The minimization problem when panel data are used for extra information on heterogeneity is

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{NW} \sum_{i=1}^{N} \sum_{w=1}^{W} \left(y_{i,w} - \sum_{r=1}^{R} \theta^r \left(\prod_{t=1}^{T} g_{w_t}(x_{i,t}, \beta^r) \right) \right)^2 \quad \text{subject to (2) and (3).} \tag{12}$$

With panel data, the estimator matches sequences of choices. Here $\prod_{t=1}^{T} g_{w_t}(x_{i,t}, \beta^r)$ is the product of the dynamic logit choice probabilities, which are calculated using the assumption that the choice-specific errors $\varepsilon_{i,j,t}$ are independent across t. As in the static logit example, we could use normal density functions as basis functions to smooth the estimates of the distribution of random coefficients. However, it is not computationally desirable to use the location and scale model, because modifying these parameters would require us to re-solve the model.[5]

Indeed, a major computational advantage of our approach is that we need to solve the Bellman equations (11) only once for each of the R types. By comparison, most other methods for general, potentially nonlinear mixtures require the researcher to evaluate the $g_j(x_{i,t}, \beta^r)$ function as part of optimizing some statistical objective function. If the

[5] Like other estimators for dynamic programming models, there may be an initial conditions problem that may attenuate with increasing panel length (Heckman (1981)).

function $g_j(x_{i,t}, \beta)$ must be evaluated S times in a standard simulation estimator for each evaluation of the objective function in the simulation estimator, the dynamic program will have to be computed $S \cdot L$ times, when L is the number of times the objective function is called by the optimization routine. By contract, our method requires only R solutions to dynamic programs. We view the usefulness of our estimator for reducing the computational burden of estimating complex structural models to be one of our estimator's key selling points.

The idea of presolving a complex economic model for only R types before optimization commences is also found in Ackerberg (2009). Ackerberg cannot estimate his model using linear regression, although an advantage is the ease of allowing homogeneous parameters that enter into a linear index that also has an additive, heterogeneous error, as in $x'\gamma + v_i$, where γ is not a heterogeneous parameter but the additive error v_i is. This is because only the value of the sum $x'\gamma + v_i$ is necessary to solve the dynamic program, and that sum is a heterogeneous value because of the error v_i. In our method we impose independence between regressors and errors, so if we add homogeneous parameters to our model, we have to resolve the dynamic programming problem every time the optimization routine changes the guess of the homogeneous parameters. We lose part of our method's computational simplicity with homogeneous parameters.

5.2 *Joint discrete and continuous demand*

We can also estimate a model with a joint discrete and continuous choice using our methods. Fox and Gandhi (2011a) introduced identification results for this type of model. Inspired by the classic application of Dubin and McFadden (1984), suppose that 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. Fox and Gandhi studied identification using heterogeneous functions rather than a heterogeneous vector of parameters β. Heterogeneous functions constitute a greater generality than is often used in estimation. We proceed with a simple example that is mostly a special case of the identification analysis in Fox and Gandhi. Let the notation for the discrete choice be the same as before. The electricity usage equation of consumer i of type r for air conditioner j is

$$a^r_{i,j} = w'_{i,j}\gamma^r_j + \eta^r_j,$$

where $w_{i,j}$ is a vector of observable characteristics that affect electricity demand. There can be overlap between the elements of $x_{i,j}$ and $w_{i,j}$. The parameter γ^r_j is a potentially choice-specific random coefficient vector for type r in the outcome equation. The scalar η^r_j is a choice-specific error term. Let $w_i = (w'_{i,1}, \dots, w'_{i,J})$, $\gamma = (\gamma_1, \dots, \gamma_J)$, and $\eta = (\eta_1, \dots, \eta_J)$.

Because the dependent variable includes a continuous element, we need to exploit the general model in (1) and work with a set $A = [a^l_j, a^{l+1}_j)$ for the real-valued dependent variable. In a finite sample, the researcher must discretize the continuous outcome variable a_j by choosing L bins: $[a^0_j, a^1_j)$, $[a^1_j, a^2_j)$, through $[a^{L-1}_j, a^L_j)$. A higher L increases the computational burden and the closeness of the approximation to the continuous outcome model. Potentially, discretization could affect identification and in that case L should be increased with the sample size N.

Let $y_{i,j}^{l} = 1$ 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

$$\begin{aligned}\Pr(y_{i,j}^{l} = 1 \mid x_i, w_i; \beta^r, \gamma^r, \eta^r) &= g_{j,l}(x_i, w_i, \beta^r, \gamma^r, \eta^r) \\ &= \frac{\exp(x_{i,j}'\beta^r)}{1 + \sum_{j'=1}^{J} \exp(x_{i,j'}'\beta^r)} 1[a_j^l \le w_{i,j}'\gamma_j^r + \eta_{i,j}^r < a_j^{l+1}].\end{aligned}$$

The unknown object of interest is the joint distribution $F(\beta, \gamma, \eta)$ of the multinomial choice random coefficients β, the electricity usage random coefficients γ, and the additive errors in utility usage η. In this case one can choose a grid of taste parameters $(\beta^r, \gamma^r, \eta^r)_{r=1}^{R}$. A statistical observation is $(j_i, a_{i,j}, x_i, w_i)$, which can be transformed into $((y_{i,j}^{l})_{1\le j\le J, 0\le l<L}, x_i, w_i)$ by a change of variables. Data on $a_{i,k}$ for $k \ne j_i$ are not needed for this transformation. The estimate of θ minimizes the objective function

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{l=1}^{L} \left(y_{i,j}^{l} - \sum_{r=1}^{R} \theta^r g_{j,l}(x_i, w_i, \beta^r, \gamma^r, \eta^r) \right)^2 \quad \text{subject to (2) and (3).} \tag{13}$$

Estimating θ provides an estimator of $F(\beta, \gamma, \eta)$,

$$\hat{F}(\beta, \gamma, \eta) = \sum_{r=1}^{R} \hat{\theta}^r 1[\beta^r \le \beta, \gamma^r \le \gamma, \eta^r \le \eta]. \tag{14}$$

This distribution completely determines the model.

This joint continuous and discrete demand model is an example of a selection model. In this example, the continuous outcome $a_{i,j}$ is selected because the researcher only observes the electricity usage for air conditioner j for those individuals who purchase air conditioner j, that is, $j_i = j$. Regressing $a_{i,j}$ on $w_{i,j}$ for those individuals who choose j will not consistently estimate $E[\gamma_j]$ if γ_j is not statistically independent of β. Those agents who choose air conditioner j, $j_i = j$, have certain β preferences, and the correlation of the preferences for air conditioners β with the usage random coefficients γ and η and the correlation of the characteristics of air conditioners relevant for purchase x with the characteristics of air conditioners relevant for usage w induce correlation between $w_{i,j}$ and both η_i and γ_i in the sample of those who pick a particular air conditioner j. Jointly modeling both the choice of air conditioner and the 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. Commonly, selection models focus on allowing random coefficients in only the outcome equation.

5.3 *Omitted variable bias from endogenous regressors*

In some cases, the regressors may be correlated with omitted factors. The most common case in demand estimation is when price is correlated with omitted factors. Here, we outline an approach to correcting omitted variable bias in models with random coefficients. It is not a special case of the control function and inversion approaches emphasized in the prior literature on demand estimation because the unobservables correlated with price need not be scalars (like omitted product characteristics) and we do not identify the values of the unobservables.[6]

For expositional simplicity and generality, we let price and the corresponding instruments vary at the individual level, but price, the demand errors correlated with price, and the instruments for price could vary at the market level instead. If the endogenous regressors vary at the market level, the standard assumption is that the market-level unobservables are statistically independent of the individual-level unobservables. In estimation, we could impose independence between the market-level and the individual-level errors following the approach in Section 4.4, where in that section's notation β^1 refers to individual-level random coefficients and β^2 is the realization of the demand errors correlated with price.

Consider the multinomial choice model where the random coefficients $\beta_{i,j}$ can now vary with product identity j to allow for unobserved product characteristics, such as the term $\xi_{j,t}$ from the literature following Berry, Levinsohn, and Pakes (1995). For example, there can be an intercept term $\beta_{i,j,K}$ reflecting how product quality affects demand. We let $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,J})$ be the stacked vector of random coefficients, including random intercepts, for all J choices. In our approach, all elements of β_i, not just the intercept, may be correlated with prices. Fox and Gandhi (2010) addressed the correlation of $x_{i,j}$ with $\beta_{i,j}$ 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. Fox and Gandhi discussed identification using random functions; here we specialize to random coefficients for more practical estimation. Fox and Gandhi (2011b) showed identification results where a more complex pricing equation is derived from the equilibrium to a Bertrand–Nash pricing game, as in the model of Berry, Levinsohn, and Pakes.

If price $p_{i,j}$ is an endogenous regressor and $w_{i,j}$ are the instruments for price, then the auxiliary equation for type r is

$$p^r_{i,j} = w'_{i,j}\gamma^r_j + \eta^r_j,$$

where γ^r_j is a vector of random coefficients showing how price is affected by instruments and η^r_j is an additive error in price. The difference between the previous selection case

[6] In the control function approach, some generalized residual from a first-stage estimation is inserted into the second stage to control for the (typically scalar) omitted factor. This is discussed in Kim and Petrin (2010). The other approach is, during optimization over the structural parameters, to uniquely invert the omitted factor from data on market shares (or choice probabilities) and an assumption that there is a continuum of consumers in each market, so that the market shares or choice probabilities lack sampling and measurement error. The latter inversion approach is introduced in Berry, Levinsohn, and Pakes (1995) and developed in terms of identification in Berry and Haile (2010a, 2010b).

and the case of omitted variable bias here is that price $p_{i,j}$ is observed for all J choices for each agent. There is no selection problem and this model is not a special case of the previous one. Instead, there is a traditional omitted variables problem where price $p_{i,j}$ enters the logit demand model as a component of $x_{i,j}$ and the random variable $p_{i,j}$ is not independent of $\beta_{i,j}$, the vector of random coefficients for choice j in the logit model.

Estimation works with the reduced form of the model. Let $x_{i,j,1} = p_{i,j}$, or the endogenous regressor price is the first characteristic in the vector of product characteristics $x_{i,j}$. Let $\beta_{1,j}$ reflect the random coefficient on price in the discrete choice utility for product j. Let $\tilde{\beta}_{i,j}$ be the $K-1$ other, nonprice random coefficients, including possibly a random intercept representing unobserved product characteristics, and let $\tilde{x}_{i,j}$ be the $K-1$ other, nonprice product characteristics for product j. As price takes on real values, we bin each of the j prices $p_{i,j}$ into L intervals $[a_j^l, a_j^{l+1})$, as we did for the continuous outcome in the selection model. With this discretization, the transformed binary outcome $y_{i,j}^{l_1,\dots,l_J}$ is

$$y_{i,j}^{l_1,\dots,l_J} = \begin{cases} 1, & j \text{ picked and } p_{i,j'} \in [a_{j'}^{l_{j'}}, a_{j'}^{l_{j'}+1})\ \forall j' = 1, \dots, J, \\ 0, & \text{otherwise.} \end{cases}$$

The outcome $y_{i,j}^{l_1,\dots,l_J}$ is indexed both by the product picked j and the value of all J prices. The probability that $y_{i,j}^{l_1,\dots,l_J} = 1$ and the consumer has random coefficients of type r is

$$\begin{aligned} &\Pr(y_{i,j}^{l_1,\dots,l_J} = 1 \mid x_i, w_i; \beta^r, \gamma^r, \eta^r) \\ &\quad = g_{j,l_1,\dots,l_J}(x_i, w_i, \beta^r, \gamma^r, \eta^r) \qquad (15) \\ &\quad = \frac{\exp(\tilde{x}_{i,j}'\tilde{\beta}_j^r + \beta_{1,j}^r(w_{i,j}'\gamma_j^r + \eta_j^r))}{1 + \sum_{j'=1}^{J} \exp(\tilde{x}_{i,j'}'\tilde{\beta}_j^r + \beta_{1,j}^r(w_{i,j'}'\gamma_{j'}^r + \eta_{j'}^r))} \prod_{j'=1}^{J} 1[a_{j'}^l \le w_{i,j'}'\gamma_{j'}^r + \eta_{j'}^r < a_{j'}^{l+1}]. \end{aligned}$$

The probability is equal to the logit probability of choice j given the prices predicted by the random coefficients in the pricing equation times the event that all predicted prices for type r are in the correct interval.

The key idea that allows us to correct for omitted variable bias is that we work with the reduced form of the model: we replace the actual J prices in the data $p_{i,j}$ with the J predicted prices $w_{i,j}'\gamma_j^r + \eta_j^r$ from the auxiliary pricing equation. The actual data $p_{i,j}$ that are allowed to be statistically dependent with β_i do not appear in (15). Only data on $\tilde{x}_{i,j}$, the exogenous regressors, and $w_{i,j}$, the instruments that enter the pricing equation, are used. We assume that $(\tilde{x}_i, w_i)$, where $\tilde{x}_i = (\tilde{x}_{i,1}', \dots, \tilde{x}_{i,J}')$ and $w_i = (w_{i,1}', \dots, w_{i,J}')$, is independent of $(\beta_i, \gamma_i, \eta_i)$, the heterogeneity realizations. Estimation proceeds analogously to the selection case (13), except $g_{j,l_1,\dots,l_J}(x_i, w_i, \beta^r, \gamma^r, \eta^r)$ replaces $g_{j,l}(x_i, w_i, \beta^r, \gamma^r, \eta^r)$ in (13). As in the selection example, the omitted variables example has as its object of interest $F(\beta, \gamma, \eta)$. The estimator of $F(\beta, \gamma, \eta)$ is (14), with the estimates $\hat{\theta}$ coming from the omitted variable bias and not the selection model. Note that this framework allows random coefficients both in the pricing and in discrete choice decisions, and those random coefficients have an unrestricted joint distribution $F(\beta, \gamma, \eta)$.

5.4 *Aggregate data and measurement error in shares*

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

$$u^r_{i,j} = x'_{j,t}\beta^r + \varepsilon_{i,j}.$$

In this model, the utility of individuals is a function of product- and market-varying attributes $x_{j,t}$. In applied work, characteristics 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 all the $x_{j,t}$. Suppose that the economist only observes a noisy measure $\hat{s}_{j,t}$ 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 $\hat{s}_{j,t}$. Simple algebra implies that

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

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

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{JT} \sum_{t=1}^{T} \sum_{j=1}^{J} \left(\hat{s}_{j,t} - \sum_{r=1}^{R} \theta^r g_j(x_t, \beta^r)\right)^2$$
$$\text{subject to (2) and (3).}$$

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. Indeed, Berry, Linton, and Pakes (2004) showed that small amounts of sampling error in shares can result in large biases in parameter estimates in the framework of Berry, Levinsohn, and Pakes.

6. Estimating confidence regions with finite types

We now discuss computing standard errors under the assumption that the set of R types used in estimation is the true set of R types that generates the data, allowing for irrelevant types (types of zero mass where $\theta^r = 0$). In other words, the grid of R points $\mathcal{B}_R$ is a superset of the true grid that takes on positive support. Under this assumption, one can use the common ordinary least-squares (OLS) heteroskedasticity-consistent standard errors for the unknown weights $\theta^1, \dots, \theta^R$. The traditional OLS confidence intervals are computed using the unconstrained point estimates in Section 2.3 instead of the point estimates with constraints from Section 2.4. However, the confidence regions so constructed give correct coverage (more than 95%) for the estimates both with and without the constraints (Andrews and Guggenberger (2010a, footnote 7)). We use the common standard error formulas in our empirical example below.

We need to use heteroskedasticity-consistent standard errors (SE) because the errors in a linear probability model such as (6) are heteroskedastic. We 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 $\hat{\theta}^r$, remove infeasible values by reporting, for a 95% two-sided confidence interval,

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

where $\mathrm{SE}(\hat{\theta}^r)$ 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 θ but rather for functions $m(\theta; X)$, where X denotes some arbitrary data. For example, researchers may wish to construct confidence intervals for the distribution $\hat{F}(\beta) = m(\hat{\theta}; X) = \sum_{r=1}^{R} \hat{\theta}^r 1[\beta^r \le \beta]$ evaluated at a particular value of β (X does not enter m here). To construct standard errors for $m(\hat{\theta}; X)$, first construct the distribution of $\hat{\theta}^r - \theta_0^r$ as above and then use the delta method. A 95% confidence interval is then

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

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

The common heteroskedasticity-consistent standard errors give more than 95% coverage but, based on our Monte Carlo evidence, are often quite conservative in that the coverage is much more than 95%.[7] In empirical work, we often find that many of the included R types have estimated weights of $\hat{\theta}^r = 0$. Thus, in principle, we can construct less conservative confidence intervals by recognizing that the parameters on the boundary of the parameter space cannot have an asymptotically normal distribution.[8] While

[7] We performed Monte Carlo studies using Tikhonov regularization/ridge regression (and Golub, Heath, and Wahba's (1979) generalized cross-validation method to pick the perturbation value) to compute standard errors. Tikhonov regularization reduced the size of the confidence regions some, but the coverage was still much more than 95%.

[8] Statistical inference for linear regression subject to a set of inequality constraints was studied by Judge and Takayama (1966), Liew (1976), Geweke (1986), and Wolak (1989).

Andrews (1999, 2002) and Andrews and Guggenberger (2010b) studied related cases, the reality is that this recent literature has not developed general-enough results that could allow us to estimate confidence intervals for our problem in a way that gives asymptotically correct coverage as defined by Andrews and Guggenberger. Indeed, Andrews and Guggenberger studied only the case of a regression with one inequality constraint and i.i.d. observations. Traditional OLS confidence intervals using fixed critical values and based on point estimates imposing one constraint are recommended by those authors, but there is no suggestion that traditional OLS methods with fixed critical values and based on point estimates imposing the constraints give asymptotically correct coverage if there are two or more inequality constraints, as in our estimator.

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) showed that the standard bootstrap is inconsistent but that subsampling and the m-out-of-n bootstrap are pointwise consistent. Andrews and Guggenberger (2010b) showed that the latter two resampling procedures are not uniformly consistent and may have poor finite-sample coverage. Andrews and Guggenberger (2009, 2010a) discussed 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 θ. The hybrid subsampling method has correct asymptotic coverage under the definition of Andrews and Guggenberger for the special case of one inequality constraint, but as Andrews and Guggenberger (2010a) showed, so does one of its ingredients, the traditional fixed critical value method that Andrews and Guggenberger recommend. Subsampling by itself does not have correct asymptotic coverage.[9]

In an appendix available on request, we show consistency and derive the sampling distribution of the inequality-constrained nonlinear least-squares estimator of the location and scale model. The distribution for nonlinear least squares 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)).

7. Monte Carlo

We conduct a Monte Carlo experiment to study the finite-sample properties of our estimator. We use individual-level discrete choice data, where the true data generating process is the random coefficients logit. In our Monte Carlo study, x_j is a 2×1 vector. Each covariate is drawn independently from $\mathcal{N}(0, 1.5^2)$. Each agent makes a single choice from a menu of $J = 10$ products in each of our individual-specific choice sets, with the option of making no purchase. We vary the number of individuals N from 2000 to 5000 to 10,000, and use $M = 50$ replications for each sample size.

We use discrete basis functions to approximate the underlying CDF, which in this section in truth has continuous support. The number of basis points is $R = t^2$, $t = 3, \dots, 9$. The basis functions are uniformly distributed over the support $[-3, 5] \times [-3, 5]$,

[9] In an unreported Monte Carlo study, we also found that subsampling could undercover the true parameters: it has coverage less than 95%.

which is sufficient to cover the support of the true distributions used below with coverage probabilities being close to 1. For each run, after we compute the estimate $\hat{F}(\beta)$, we evaluate its squared difference from the true distribution function $F_0(\beta)$ at $S = 10{,}000$ points (β_s's below) uniformly spaced over $[-6, 6] \times [-6, 6]$. We use root mean integrated squared error (RMISE) to assess performance of our estimator. Our definition of RMISE for an estimator $\hat{F}$ is

$$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{S}\sum_{s=1}^{S}\left(\hat{F}_m(\beta_s) - F_0(\beta_s)\right)^2\right]},$$

where again we use $M = 50$ replications, each with a new fake data set. We also report the integrated absolute error (IAE), which for a given replication m is

$$\frac{1}{S}\sum_{s=1}^{S}\left|\hat{F}_m(\beta_s) - F_0(\beta_s)\right|.$$

This is a measure of the mean absolute value of the estimation error, taken across the points of evaluation for a given replication. We compute the mean, minimum, and maximum IAE's across the M replications.

We generate data using three alternative distributions $F(\beta)$ for the random coefficients. Let the variance matrix $\Sigma_1 = \left[\begin{smallmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{smallmatrix}\right]$ and likewise let $\Sigma_2 = \left[\begin{smallmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{smallmatrix}\right]$. In the first design, the tastes for characteristics are generated from a mixture of two normals:

$$0.4 \cdot \mathcal{N}([3, -1], \Sigma_1) + 0.6 \cdot \mathcal{N}([-1, 1], \Sigma_2).$$

In the second design, the true coefficients are generated by a mixture of four normals:

$$\begin{aligned} &0.2 \cdot \mathcal{N}([3, 0], \Sigma_1) + 0.4 \cdot \mathcal{N}([0, 3], \Sigma_1) \\ &\quad + 0.3 \cdot \mathcal{N}([1, -1], \Sigma_2) + 0.1 \cdot \mathcal{N}([-1, 1], \Sigma_2). \end{aligned}$$

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

$$\begin{aligned} &0.1 \cdot \mathcal{N}([3, 0], \Sigma_1) + 0.2 \cdot \mathcal{N}([0, 3], \Sigma_1) + 0.2 \cdot \mathcal{N}([1, -1], \Sigma_1) \\ &\quad + 0.1 \cdot \mathcal{N}([-1, 1], \Sigma_2) + 0.3 \cdot \mathcal{N}([2, 1], \Sigma_2) + 0.1 \cdot \mathcal{N}([1, 2], \Sigma_2). \end{aligned}$$

We summarize our results in Table 1, where the true distribution has two or four components, and in Table 2, where the true distribution has six components. The first two columns report the sample size N and the number R of basis points used in the estimation. The next column reports the RMISE of the estimated distribution functions. The following three columns reports the mean, minimum, and maximum of the IAE. The final three columns report the mean, minimum, and maximum of the number of basis functions that have positive weight.

Table 1. Monte Carlo results: Truth is two or four mixtures.

			Integrated Absolute Error			No. of Positive Weights		
N	*R*	RMISE	Mean	Minimum	Maximum	Mean	Minimum	Maximum
				Number of Mixtures: 2				
2000	9	0.035	0.016	0.011	0.033	4.3	3	6
2000	16	0.037	0.019	0.01	0.035	5.6	4	8
2000	25	0.04	0.021	0.011	0.054	6.2	4	8
2000	36	0.045	0.023	0.011	0.07	6.7	4	10
2000	49	0.048	0.024	0.011	0.077	7	4	10
2000	64	0.053	0.025	0.012	0.087	7.4	4	11
2000	81	0.053	0.024	0.012	0.081	7.6	5	10
5000	9	0.034	0.013	0.0096	0.019	4.5	3	6
5000	16	0.035	0.015	0.01	0.022	5.5	3	9
5000	25	0.035	0.016	0.01	0.022	6.6	4	10
5000	36	0.035	0.016	0.011	0.027	7.3	4	12
5000	49	0.036	0.016	0.011	0.03	7.7	5	11
5000	64	0.037	0.016	0.011	0.034	8	5	12
5000	81	0.037	0.016	0.011	0.034	8.2	5	15
10,000	9	0.034	0.012	0.0097	0.017	4.6	3	8
10,000	16	0.034	0.013	0.0096	0.019	5.6	3	8
10,000	25	0.034	0.013	0.0096	0.022	6.9	3	11
10,000	36	0.035	0.014	0.0099	0.023	7.7	4	11
10,000	49	0.035	0.014	0.01	0.023	8.3	4	12
10,000	64	0.035	0.014	0.011	0.026	8.8	5	15
10,000	81	0.035	0.014	0.011	0.024	9	5	14
				Number of Mixtures: 4				
2000	9	0.14	0.091	0.09	0.095	6.1	4	8
2000	16	0.11	0.076	0.065	0.089	6.8	4	10
2000	25	0.081	0.05	0.029	0.074	7.7	5	11
2000	36	0.069	0.041	0.014	0.072	8.6	5	12
2000	49	0.09	0.054	0.024	0.081	9.1	6	14
2000	64	0.096	0.059	0.032	0.076	12	5	64
2000	81	0.1	0.061	0.038	0.081	11	6	17
5000	9	0.14	0.091	0.089	0.092	6.7	5	9
5000	16	0.11	0.074	0.067	0.081	6.7	5	10
5000	25	0.074	0.044	0.026	0.071	7.9	5	11
5000	36	0.056	0.033	0.015	0.067	9.5	6	12
5000	49	0.082	0.047	0.026	0.073	11	7	13
5000	64	0.09	0.054	0.036	0.072	11	7	17
5000	81	0.091	0.054	0.03	0.07	13	7	20
10,000	9	0.14	0.09	0.089	0.092	6.4	5	8
10,000	16	0.1	0.073	0.069	0.076	6.5	4	10
10,000	25	0.067	0.041	0.029	0.053	8.3	5	13
10,000	36	0.041	0.024	0.012	0.044	9.9	5	15
10,000	49	0.073	0.04	0.024	0.065	11	7	16
10,000	64	0.089	0.051	0.029	0.068	12	8	18
10,000	81	0.094	0.055	0.029	0.069	14	7	20

Table 2. Monte Carlo results: Truth is six mixtures.

			Integrated Absolute Error			No. of Positive Weights		
N	R	RMISE	Mean	Minimum	Maximum	Mean	Minimum	Maximum
				Number of Mixtures: 6				
2000	9	0.2	0.15	0.14	0.16	6.1	5	7
2000	16	0.11	0.07	0.064	0.093	9.7	7	12
2000	25	0.12	0.087	0.064	0.099	10	7	23
2000	36	0.057	0.04	0.016	0.063	12	8	15
2000	49	0.092	0.065	0.054	0.079	13	9	17
2000	64	0.082	0.059	0.043	0.085	14	10	19
2000	81	0.072	0.052	0.029	0.083	15	11	23
5000	9	0.19	0.15	0.14	0.15	6.2	5	7
5000	16	0.11	0.069	0.065	0.088	10	7	12
5000	25	0.12	0.09	0.082	0.096	9.7	7	11
5000	36	0.045	0.03	0.018	0.053	13	8	16
5000	49	0.091	0.063	0.052	0.075	13	9	16
5000	64	0.075	0.053	0.04	0.074	15	11	19
5000	81	0.068	0.05	0.037	0.068	17	9	23
10,000	9	0.19	0.15	0.15	0.15	6.2	5	7
10,000	16	0.11	0.069	0.064	0.075	10	7	13
10,000	25	0.12	0.09	0.079	0.098	10	8	14
10,000	36	0.043	0.028	0.014	0.057	13	8	18
10,000	49	0.091	0.062	0.055	0.07	14	9	18
10,000	64	0.073	0.051	0.044	0.066	15	11	23
10,000	81	0.067	0.05	0.039	0.067	17	11	24

The results in Tables 1 and 2 suggest that our estimator of $F(\beta)$ exhibits excellent performance, even in relatively small samples. By and large, RMISE and IAE are relatively low and decrease with N and R. Note that we are able to fit the underlying distribution functions well with a fairly small number of basis functions. While performance generally increases with the number of basis functions, it is worth noting that the fit can decrease with increases in R, as the grids do not nest each other for marginal increases in R. This is apparent in this sample for $N = 2000$, where smaller R's have lower RMISE's. The mean number of nonzero basis functions ranges from 4 to 17, with more complex true distributions requiring more nonzero basis points for approximation. This result is consistent with the literature on mixtures, which demonstrates that complicated distributions can be approximated with a surprisingly small number of mixture components.

We plot the estimated marginal distributions of both β_1 and β_2. Figure 1 includes one plot for each of the three designs and each of the two random coefficients. For comparison, we also include a plot of a bivariate normal fitted to the same data sets using maximum likelihood, along with the 5th and 95th quantiles of the estimated distributions at each point of evaluation. The general fits of our current estimator are excellent; the discrete distribution tracks the CDF of the underlying model uniformly well. The variability across replications is quite low at this sample size, and the true CDF lies within the 90 percent confidence bands with the small exception of boundary values

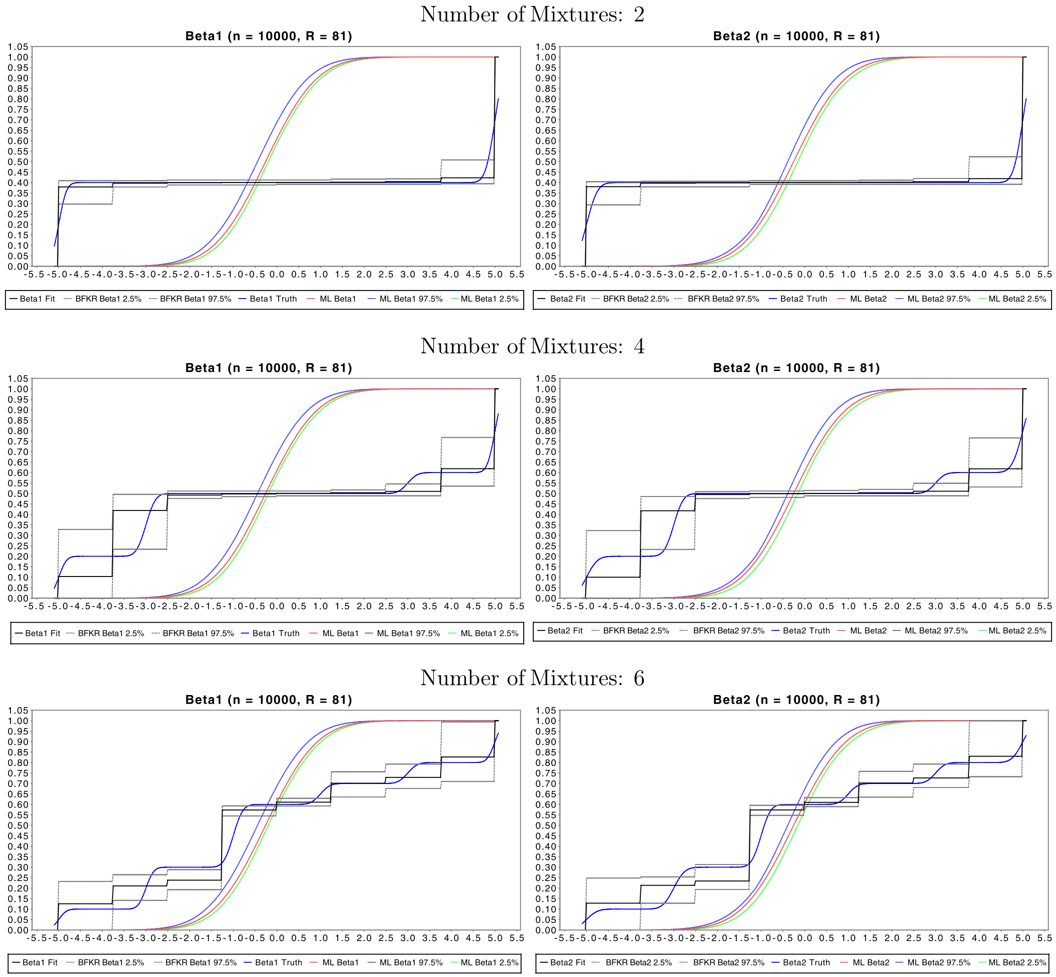

FIGURE 1. Marginal distribution of β_1 and β_2 with $N = 10{,}000$ and $R = 81$.

where the discrete CDF jumps. On balance, our estimator consistently recovers the true marginal distributions of the two coefficients.

Figures 2 and 3 plot the true versus the estimated joint distribution (using one of our replications) of the model with six mixtures. Here $R = 81$ and $N = 10{,}000$. A visual inspection of the distributions shows that we are able to generate excellent fits to the true distribution functions of preferences using our estimator. For comparison, Figure 4 is a plot of the estimation error. The sawtooth nature of the error plot is driven by the fact that the discrete approximation to the CDF results in step functions, which then imply that the errors have discontinuities at the basis points. Our estimator produces a good approximation to the underlying distribution function even with limited data and when the number of basis points is small.

In Table 3, we report the results for a bivariate normal distribution with a nonzero correlation parameter run on the same data sets. The bivariate normal has a RMISE and an IAE noticeably larger than our values. The fit does not improve with an increase in

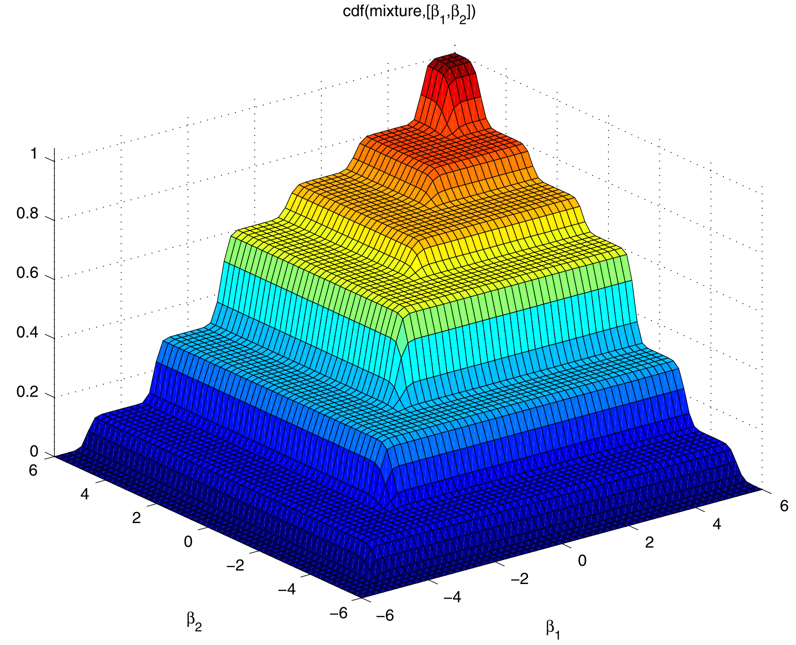

Figure 2. True distributions for six mixtures.

the number of observations. As can be seen in the graphs of the marginal densities in Figure 1, the bivariate normal distribution simply cannot match the shape of the underlying true CDF. Interestingly, it has a better fit on the true models with more mixture

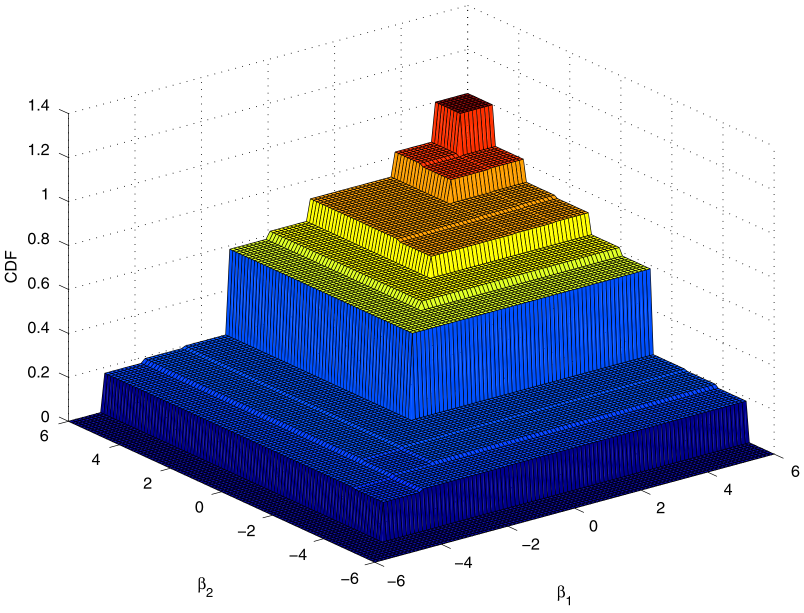

Figure 3. Fitted distributions for six mixtures with $N = 10{,}000$ and $R = 81$.

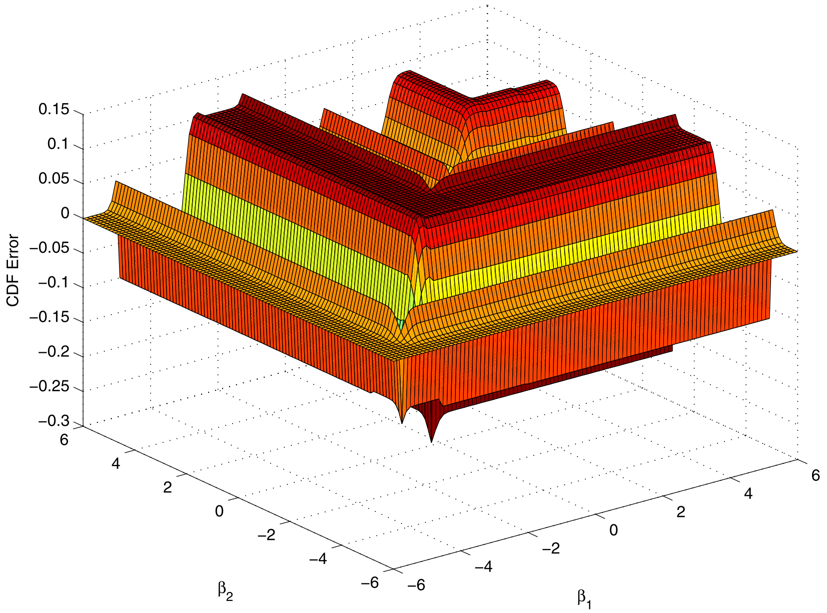

FIGURE 4. Error in fitted distribution for six mixtures with $N = 10{,}000$ and $R = 81$.

components, as the underlying CDF has more features and on average is closer in shape to the bivariate normal. In all cases, however, our estimator dominates the performance of the bivariate normal.

We also experimented with using a random grid to populate the basis points.[10] The random grid exhibited much poorer convergence properties than the uniform grid, pri-

TABLE 3. Monte Carlo results: Bivariate normal maximum likelihood.

		Integrated Absolute Error		
N	RMISE	Mean	Minimum	Maximum
		Number of Mixtures: 2		
2000	0.29	0.24	0.19	0.31
5000	0.3	0.25	0.2	0.31
10,000	0.3	0.25	0.2	0.3
		Number of Mixtures: 4		
2000	0.15	0.082	0.066	0.12
5000	0.15	0.078	0.07	0.088
10,000	0.16	0.074	0.07	0.086
		Number of Mixtures: 6		
2000	0.18	0.094	0.09	0.1
5000	0.18	0.094	0.088	0.1
10,000	0.19	0.095	0.089	0.11

[10] Results available from authors on request.

marily due to the fact that there are no guarantees in small samples (with correspondingly small numbers of basis points) that the grid will have good coverage over the relevant areas of the parameter space. For this reason, we suggest the use of a uniform grid when data and computational limits allow it.

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 (forthcoming), hereafter referred to as DHR. They considered the problem of incentivizing teachers to go to work and estimated a dynamic labor supply model using the method of simulated moments. The model is a single agent, dynamic programming problem with a serially correlated, unobserved state variable. To accommodate unobserved heterogeneity across teachers, they estimated a 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 nonformal education centers, nongovernment-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 who went to school at every opportunity received greater wages than the control group, who were 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 evaluated this program and found the combination of incentives was successful in reducing absenteeism from 42 percent to 21 percent. To disentangle the confounding effects of the monitoring and financial incentives, they exploited 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 focused on the daily work decisions 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 $\mu_i + \varepsilon_{i,t}$, where μ_i is specific to teacher i and $\varepsilon_{i,t}$ is a shock to payoffs. The shock $\varepsilon_{i,t}$ 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 d that teacher worked in the month:

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

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

$$\begin{aligned} V^r(t, d_i, \varepsilon_{i,t}) = \max\bigl\{&E\bigl[V^r(t+1, d_i+1, \varepsilon_{i,t+1}) \mid \varepsilon_{i,t}\bigr], \\ &\mu^r + \varepsilon_{i,t} + E\bigl[V^r(t+1, d_i, \varepsilon_{i,t+1}) \mid \varepsilon_{i,t}\bigr]\bigr\}. \end{aligned}$$

At time C, the value function simplifies to

$$V^r(C, d_i, \varepsilon_{i,C}) = \max\bigl\{\beta\pi(d_i+1), \mu^r + \varepsilon_{i,C} + \beta\pi(d_i)\bigr\}, \tag{16}$$

where β is the marginal utility of an additional rupee. There is no continuation value in the right side of (16) as the stock of days worked resets to zero at the end of the month. These value functions illustrate the trade-off teachers face early in each month between accruing days worked, in the hope of receiving an extra payoff at the end of the month, and obtaining instant gratification by skipping work.[11]

DHR proposed a simulated method of moments estimator that matches sequences of days worked at the beginning of each month. They found parameters that 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.[12] This procedure resulted in $2^5 - 1 = 31$ linearly independent moments to be matched in each month. They estimated 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 ρ and the teacher-specific deterministic component of the outside option μ_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 μ_i and the serially correlated unobservable $\varepsilon_{i,t}$ to produce the probabilities of sequences of days worked. The model requires several hundred thousand simulations to produce estimates of the probabilities of working sequences with low variance for each type. Using a two-type mixture estimated using

[11] 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 ε is discretized into 200 bins. For each simulated value of ε, 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 $E[V^r(t+1, d_i+1, \varepsilon_{i,t+1}) \mid \varepsilon_{i,t}]$, the distribution of $\varepsilon_{i,t+1}$ is used to calculate the probability $\varepsilon_{i,t+1}$ lies in each of the 200 bins, and those probabilities weight $V^r(t+1, d_i+1, \varepsilon_{i,t+1})$ in the approximation to the expectation.

[12] 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 $\varepsilon_{i,t}$ across months is low). We treat data from different months as statistically distinct from the viewpoint of the correlation in $\varepsilon_{i,t}$.

the method of simulated moments, DHR estimated that in a given month 97.6 percent of the teachers have outside options drawn from a normal distribution with a mean of -0.428 and a variance of 0.007, and the remaining 2.4 percent of teachers have outside options drawn from a normal distribution with mean 1.781 and variance 0.050. At the estimated parameters of this model, workers who draw the first distribution generally go to school every day, while workers who draw from second distribution are likely to never attend school during a given month.

There are natural bounds on the level of the outside option, as low values of μ_i lead to teachers always working and high values lead to teachers never working. The autocorrelation parameter is bounded between -1 and $+1$. The β parameter is also sharply bounded, as its effects on work behavior are similar to the outside option, outside a narrow range.

We apply the present paper's estimator to this setting, allowing for a more flexible 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 model in DHR: $\beta = 0.013$ and $\rho = 0.449$.[13]

We estimate the model with a discrete approximation to the 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 intramonth 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 $31 = 2^5 - 1$ 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 $31 * 21 = 651$ rows corresponding to the 31 possible work histories and the 21 different months. We formed the corresponding vector of empirical probabilities as the vector of dependent variables. We then estimated weights for each of the R types using the inequality constrained least-squares estimator.[14] This estimator is similar to the specification in equation (12), except that we do not use panel

[13] 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 (β) and the degree of persistence in the AR(1) process (ρ), with unrestricted covariance across the parameters. We found that the β and ρ parameters were always estimated tightly around their point estimates and, therefore, we focus on the distribution of μ_i in what follows.

[14] We use a penalized (for the constraints) Newton 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 6. Confidence intervals are constructed using the standard OLS formulas. An observation is

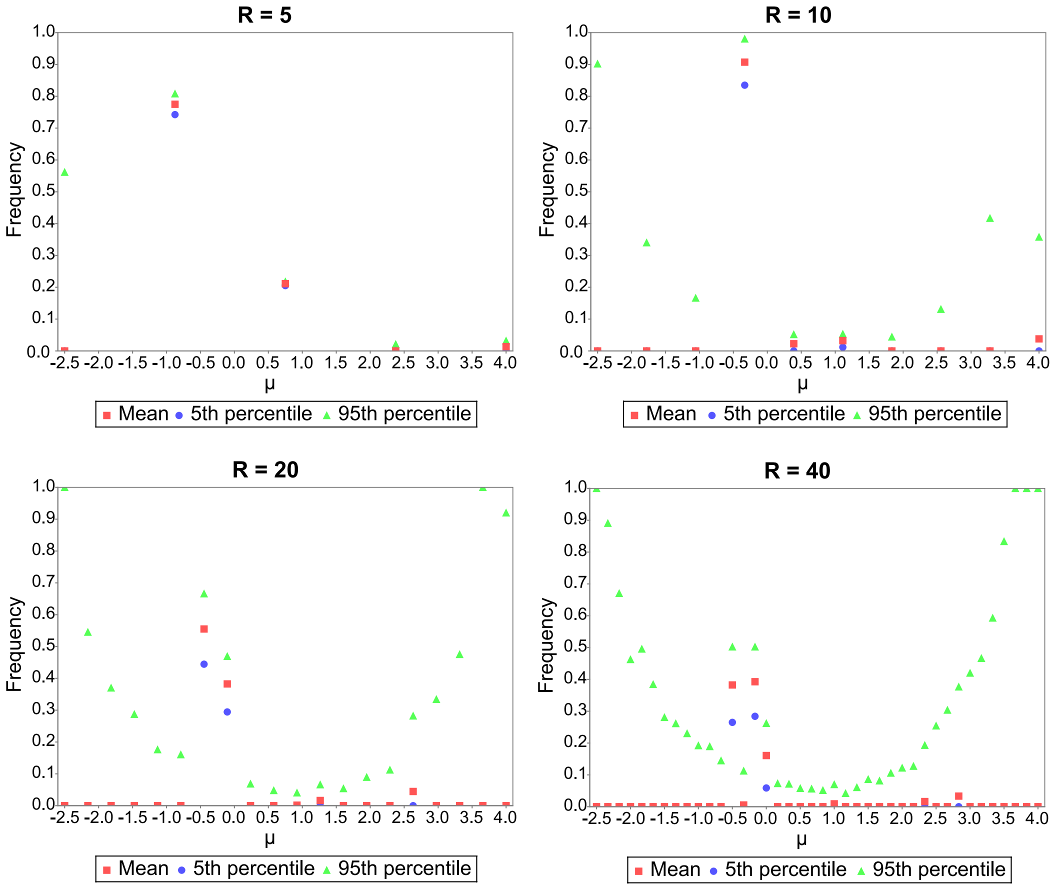

FIGURE 5. Estimated distributions of heterogeneity for the benefits of staying home.

data to construct sequences of choices for the same teacher across months, only within months.

The estimated distributions of types are shown in Figure 5. 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 smaller in the center of the distribution. Keep in mind that these are univariate confidence regions for individual weights; the confidence regions for functions of all of the weights, such as $\hat{F}(\beta)$ at a particular β, may be relatively narrower. In these and subsequent figures, we present four approximations: $R = 5$, 10, 20, and 40. The $R = 40$ estimates suggest a double-peaked distribution of a utility of staying home in the range from -0.5 to 0.0. Three basis points in the range -0.5–0.0 are given substantial weight. The right tail is thin: most weights are 0, but the

a teacher/month. Five days are used for each teacher/month observation. The number of observations is 1123 teacher/months. The correlation in $\varepsilon_{i,t}$ across the first five days should be low, so we do not account for autocorrelation across teacher/months for the same teacher.

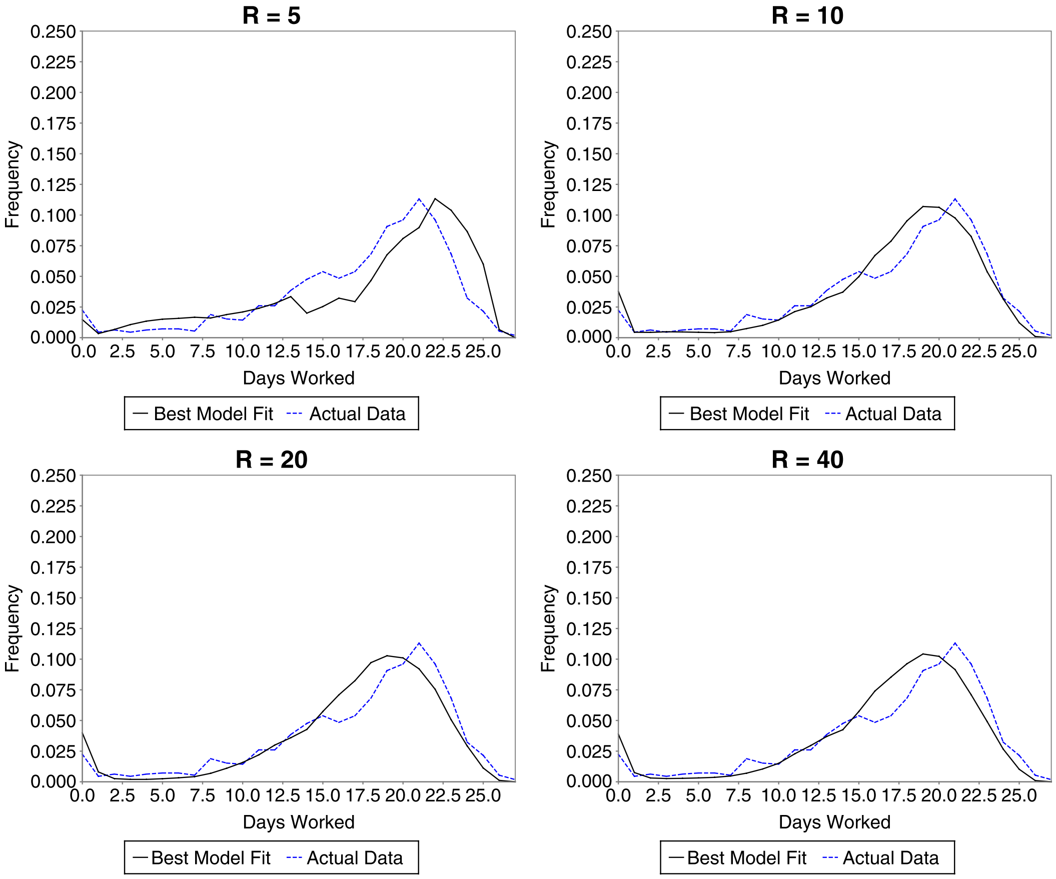

Figure 6. Predicted distribution of days worked in months where the first the five days are used for estimation.

type $\mu = 2.83$ has an estimated frequency of 3%. The middle of this support, -0.25, gives a value of staying home of $-0.25/\beta = -0.25/0.013 = -19$ rupees a day. This means that at $\varepsilon_{i,t} = 0$, a modal teacher will go to work for only a standard incentive like the threat of being fired or simply an attitude of professionalism. However, there is a positive probability of teachers with positive values of staying home, μ_i. Our estimates do not exactly match those in DHR, but they capture the same finding that most teachers are between -0.5 and 0.0, with a small fraction of teachers who always prefer to skip work in the right tail.

Figure 6 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 our 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 overpredict the proportion of days worked in the 15–20 range. We note that these fit 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.

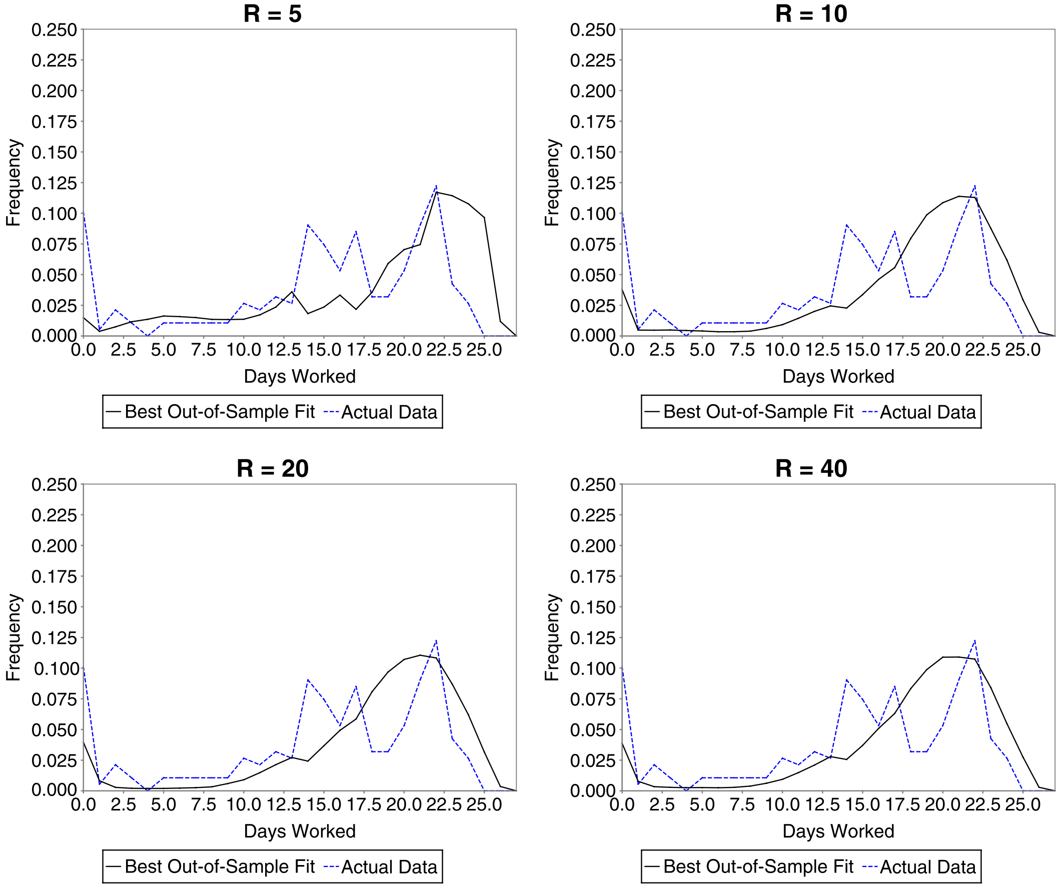

Figure 7. Predicted distribution of days worked under the out-of-sample compensation scheme.

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 7 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 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 above 18.

It is worth emphasizing that the computational burden of the present estimator is much lower than the alternative used in DHR. We ran several timing tests to evaluate the performance of our estimator against alternatives used in DHR. We tabulated the time it took our estimator to run from start to finish while varying the number of points in the discrete support, $R \in \{5, 10, 20, 40\}$, and the number of simulations used to simulate the conditional moments (here probabilities of labor supply histories) in both our estimator and the generalized method of moments (GMM) estimator used in DHR,

Table 4. Run times in seconds of our estimator and the two-type GMM estimator used in DHR.[a]

Number of Simulations in Conditional Labor Supply History Probabilities	Two-Type GMM Starting Values		Our Estimator Number of Basis Points R			
	DHR Optimum	No Heterogeneity	5	10	20	40
$S = 2000$	282	640	11.5	35.4	44.1	64.8
$S = 20{,}000$	1010	2360	33	67	120	242
$S = 200{,}000$	87,800	110,000	280	555	978	1960

[a] The objective functions for both estimators were minimized using the nonlinear solver KNITRO.

$S \in \{2000, 20{,}000, 200{,}000\}$. In the GMM estimator, we let the distribution of marginal utility be drawn from a single normal distribution. We set the number of draws from that distribution to be equal to 128 when computing the conditional moments. We started the GMM alternative with two sets of starting values. The first starting value was set at the parameter values from a model without heterogeneity on the coefficient of staying home, where we set the variance of the distribution on the value of staying home to be equal to 1. In the second experiment, we set the starting value to be the optimized value in DHR. In the second case, the optimizer simply has to verify that the starting value is a locally optimal solution and should reflect the lowest possible burden in computing the solution to the model.

Table 4 reports the run times for the two-type GMM and our estimators. For $S = 2000$ and $S = 20{,}000$, our estimator with $R = 40$ is around four times faster than the GMM estimator started at the converged parameter vector from DHR. It is over 40 times faster when $S = 200{,}000$. When the GMM estimator used in DHR is not started at the final vector, the $R = 40$ speed advantage of our estimator is between 10 and 55 times faster than the alternative. An interesting feature to highlight is that adding complexity to our model is essentially a linear operation; doubling the number of basis points roughly doubles the execution time. Also, as all of our run time is spent evaluating probabilities for different consumer types β, our approach is easy to parallelize across multiple processors.

9. Conclusion

In this paper, we have proposed a new method for estimating general mixtures models. 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. Convergence of an optimization routine to the global optimum is guaranteed under linear regression with linear inequality constraints, something that cannot be said for other statistical objective functions. Also, our estimator is easier to program and to use than alternatives such as the EM algorithm.

Our estimator is useful for estimating a wide variety of models. In a dynamic programming setting, the estimator allows for a distribution of random coefficients while simultaneously cutting the computational time even compared to the model without random coefficients. Our approach has dramatic computational savings compared to other estimators for a dynamic programming model with random coefficients. The computational savings arise because we must solve the dynamic program only once for each basis vector.

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 uses less programming and execution time than an alternative simulation estimator.

References

Ackerberg, D. A. (2009), "A new use of importance sampling to reduce computational burden in simulation estimation." *Quantitative Marketing and Economics*, 7 (4), 343–376. [396]

Andrews, D. K. W. (1999), "Estimation when a parameter is on a boundary." *Econometrica*, 67, 1341–1383. [402]

Andrews, D. K. W. (2000), "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space." *Econometrica*, 68, 399–405. [402]

Andrews, D. K. W. (2002), "Generalized method of moments estimation when a parameter is on a boundary." *Journal of Business & Economic Statistics*, 20 (4), 530–544. [387, 402]

Andrews, D. K. W. and P. Guggenberger (2009), "Hybrid and size-corrected subsample methods." *Econometrica*, 77 (3), 721–762. [402]

Andrews, D. K. W. and P. Guggenberger (2010a), "Applications of subsampling, hybrid, and size-correction methods." *Journal of Econometrics*, 158, 285–305. [401, 402]

Andrews, D. K. W. and P. Guggenberger (2010b), "Asymptotic size and a problem with subsampling and the *m* out of *n* bootstrap." *Econometric Theory*, 26, 426–468. [402]

Beran, R. and P. W. Millar (1994), "Minimum distance estimation in random coefficient regression models." *The Annals of Statistics*, 22 (4), 1976–1992. [384]

Berry, S. and P. Haile (2010a), "Identification in differentiated products markets using market level data." Working paper, NBER. [398]

Berry, S. and P. Haile (2010b), "Nonparametric identification of multinomial choice demand models with heterogeneous consumers." Working paper, NBER. [383, 398]

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile price in market equilibrium." *Econometrica*, 63 (4), 841–890. [383, 398, 400]

Berry, S., O. Linton, and A. Pakes (2004), "Limit theorems for estimating the parameters of differentiated product demand systems." *Review of Economic Studies*, 71 (3), 613–654. [400]

Biernacki, C., G. Celeux, and G. Govaert (2003), "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models." *Computational Statistics & Data Analysis*, 41, 561–575. [384]

Böhning, D. (1982), "Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process." *The Annals of Statistics*, 10 (3), 1006–1008. [383]

Burda, M., M. Harding, and J. Hausman (2008), "A Bayesian mixed logit–probit for multinomial choice demand models." *Journal of Econometrics*, 147 (2), 232–246. [383]

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. [383]

Dubin, J. A. and D. L. McFadden (1984), "An econometric analysis of residential electric appliance holdings and consumption." *Econometrica*, 52 (2), 345–362. [396]

Duflo, E., R. Hanna, and S. P. Ryan (forthcoming), "Monitoring works: Getting teachers to come to school." *American Economic Review*. [384, 409]

Fox, J. T. and A. Gandhi (2010), "Nonparametric identification and estimation of random coefficients in nonlinear economic models." Working paper, University of Michigan. [383, 385, 398]

Fox, J. T. and A. Gandhi (2011a), "Using selection decisions to identify the joint distribution of outcomes." Working paper, University of Michigan. [396]

Fox, J. T. and A. Gandhi (2011b), "Identifying Demand With Multidimensional Unobservables: A Random Functions Approach." Working paper, University of Michigan. [383, 398]

Geweke, J. (1986), "Exact inference in the inequality constrained normal linear regression model." *Journal of Applied Econometrics*, 1 (2), 127–141. [401]

Golub, G. H., M. Heath, and G. Wahba (1979), "Generalized cross-validation as a method for choosing a good ridge parameter." *Technometrics*, 21 (2), 215–223. [401]

Hansen, B. (2004), "Bandwidth selection for nonparametric distribution estimation." Working paper, University of Wisconsin. [389]

Heckman, J. J. (1981), "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process." In *Structural Analysis of Discrete Data With Econometric Applications* (C. Manski and D. McFadden, eds.), Ch. 4, 179–195, MIT Press, Cambridge. [395]

Heckman, J. and B. Singer (1984), "A method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica*, 52 (2), 271–320. [383]

Judge, G. G. and T. Takayama (1966), "Inequality restrictions in regression analysis." *Journal of the American Statistical Association*, 61 (313), 166–181. [401]

Kamakura, W. A. (1991), "Estimating flexible distributions of ideal-points with external analysis of preferences." *Psychometrika*, 56 (3), 419–431. [384]

Karlis, D. and E. Xekalaki (2003), “Choosing initial values for the EM algorithm for finite mixtures.” *Computational Statistics & Data Analysis*, 41, 577–590. [384]

Kim, K. and A. Petrin (2010), “Control function corrections for unobserved factors in differentiated product models.” Working paper, University of Minnesota. [383, 398]

Laird, N. (1978), “Nonparametric maximum likelihood estimation of a mixing distribution.” *Journal of the American Statistical Association*, 73 (364), 805–811. [383]

Li, J. Q. and A. R. Barron (2000), “Mixture density estimation.” *Advances in Neural Information Processing Systems*, 12, 279–285. [384]

Liew, C. K. (1976), “Inequality constrained least-squares estimation.” *Journal of the American Statistical Association*, 71 (355), 746–751. [401]

Lindsay, B. G. (1983), “The geometry of mixture likelihoods: A general theory.” *The Annals of Statistics*, 11 (1), 86–94. [383]

McLachlan, G. J. and D. Peel (2000), *Finite Mixture Models*. Wiley, New York. [384]

Newey, W. K. and D. McFadden (1994), “Large sample estimation and hypothesis testing.” In *Handbook of Econometrics*, Volume 4, 2111–2245, Elsevier, Amsterdam. [387]

Pilla, R. S. and B. G. Lindsay (2001), “Alternative EM methods for nonparametric finite mixture models.” *Biometrika*, 88 (2), 535–550. [384]

Politis, D. N., J. P. Romano, and M. Wolf (1999), *Subsampling*. Springer, New York. [402]

Quandt, R. E. and J. B. Ramsey (1978), “Estimating mixtures of normal distributions and switching regressions.” *Journal of the American Statistical Association*, 73 (364), 730–738. [384]

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005), *Bayesian Statistics and Marketing*. John Wiley & Sons, West Sussex. [383]

Rust, J. (1987), “Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher.” *Econometrica*, 55 (5), 999–1033. [383, 394]

Rust, J. (1994), “Structural estimation of Markov decision processes.” In *Handbook of Econometrics*, Volume 4 (R. F. Engle and D. L. McFadden, eds.), North-Holland, Amsterdam. [394]

Seidel, W., K. Mosler, and M. Alker (2000), “A cautionary note on likelihood ratio tests in mixture models.” *Annals of the Institute of Statistical Mathematics*, 52 (3), 418–487. [384]

Verbeek, J. J., N. Vlassis, and B. Kröse (2003), “Efficient greedy learning of Gaussian mixture models.” *Neural Computation*, 15, 469–485. [384]

Wolak, F. (1989), “Testing inequality constraints in linear econometric models.” *Journal of Econometrics*, 41 (2), 205–235. [401]

Submitted January, 2010. Final version accepted June, 2011.