Random Coefficients Model: A Machine Learning Approach

Nick Doudchenko Evgeni Drynkin

August 2, 2018

Abstract

In this paper we propose a new method to estimate a discrite choice demand model when individual level data are available. The method employs a two-step procedure. Step 1 predicts the choice probabilities as functions of the observed individual level characteristic. Step 2 estimates the structural parameters of the model using the estimated choice probabilities at a fixed point. We use simulations to compare the performance of the proposed procedure with the standard methodology. We find that our method delivers an improved precision as well as a substantially faster convergence time. We supplement the analysis by providing the large sample properties of the proposed estimator.

Keywords: Demand Estimation, Discrete Choice, Random Coefficients, Prediction, Machine Learning

1 Introduction

The primary goal of applied economic research is informing policy decisions. In many cases experimental analysis is infeasible or prohibitively expensive. In those cases researchers have to rely on other sources of identification. Moreover, some of the important applications require extrapolating outside of the support of the observed data. For instance, the analysis of a potential merger might require predicting the quantities and prices after two of the three firms in an industry merge. It is quite possible that within the relevant time span the industry has always consisted of three firms. As a result, the training set would not contain the data necessary to obtain a model capable of accurate counterfactual predictions. Traditionally researchers would rely on theory to specify a functional form and extrapolate outside of the support of the observed data. Issues like that have limited the adoption of purely predictive methods from statistical and machine learning in economic research.

However, in some cases parts of the overall empirical strategy can be posed as predictions problems. In that case the full force of the machine learning methods can be invoked providing the flexibility of the functional form and computational efficiency.[1] In this paper we consider the discrete choice demand estimation (McFadden, 1973; Berry, Levinsohn, and Pakes, 1995; Nevo, 2000) with coefficients that depend on observable individual level characteristics. The standard approach utilized in the literature is to assume a spicific parametric form of the dependence on the individual level variable. We propose an alternative two-step procedure. First, we solve a prediction problem that links the individual level variable to the choice probabilities. Second, we estimate the standard discrete choice model to find the coefficient at a pre-defined value of the covariate. This allows us to obtain the values of the coefficient at any other point by solving a system of linear equations.

The standard approach relies heavily on either correctly specifying the functional form or using a functional form that doesn't lead to a high bias or variance. It doesn't take into account the potential structure of the space, such as, for example, sparcity, either. Another issue with this approach is that it may become computationally burdensome when the dimensionality of the individual level characteristic increases. Using a prediction approach can address all of these issues.

We compare the proposed method to the more common approach by simulating the distribution of estimated elasticities. We consider two variations of the standard procedure: (i) the case when the parametric form of the coefficient as a function of the individual level variable is misspecified, and (ii) the case when it is specified correctly (the oracle case). When compared

[1] For a review of commonly used prediction methods see, for example, Friedman, Hastie, and Tibshirani (2001) and Murphy (2012).

in terms of the root-mean-square error our method performs about 65% better than the oracle[2] while the misspecified estimation performs almost 100% worse than the oracle. Perhaps the biggest advantage of the proposed procedure is its computational efficiency. We find that in the simplest case—when the individual level characteristic is a scalar—our method converges almost twice as fast as the considered alternatives. The efficiency gains become more significant when the dimensionality increases.

We also provide theoretical results that justify the use of the proposed estimator—it is consistent and has the same asymptotic distribution as the oracle estimator.

2 Related Literature

Berry and Haile (2014) and Dunker, Hoderlein, and Kaido (2017) address the problem of non-parametric identification in discrete choice demand models and Compiani (2018) proposes a specific way to estimate a model given the data commonly available in applications.

A study somewhat related to ours is Gillen, Shum, and Moon (2014) that uses ideas similar to those of Belloni, Chernozhukov, and Hansen (2014b) and Farrell (2015) in the context of demand estimation when the space of product-level characteristics is high-dimensional and sparse. Gillen, Montero, Moon, and Shum (2015) studies the issue of selection from a set of demographic variables to include in demand estimation. Athey, Blei, Donnelly, Ruiz, and Schmidt (2018) use individual level data and machine learning methods to estimate demand for restaurants and travel time.

There are a number of studies that apply ideas from machine learning to causal inference. See, for example, Hartford, Lewis, Leyton-Brown, and Taddy (2016); Wager and Athey (2017); Athey and Imbens (2016); Fan (2012); Belloni, Chernozhukov, and Hansen (2011b); Gautier and Tsybakov (2011); Hansen and Kozbur (2014); Chernozhukov, Hansen, and Spindler (2015); Bloniarz, Liu, Zhang, Sekhon, and Yu (2016); Athey, Imbens, and Wager (2016); Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2017); Belloni, Chernozhukov, and Hansen (2011a, 2014a); Belloni, Chen, Chernozhukov, and Hansen (2012).

[2] Why out method is more precise than the one based on the correct specification needs to be investigated further.

3 Model

Consider the following discrete choice setting. There are M separate markets populated by N_m individuals for $m = 1, \ldots, M$. Each individual $i = 1, \ldots, N_m$ is characterized by a set of observable covariates $Z_{im} \in \mathbb{R}^p$ and her product choice $d_{im} \in \{0, 1, \ldots, J\}$, where J is the number of products available in each market[3] and $d_{im} = 0$ corresponds to the outside good. Product $j = 1, \ldots, J$ is characterized by an observable k-dimensional variable X_{jm} and an unobservable $\xi_{jm} \in \mathbb{R}$. We also assume the existence of a set of instimental variables $W_{jm} \in \mathbb{R}^l$ that are uncorrelated with ξ_{jm}. Utility u_{ijm} derived by individual i from buying good $j = 1, \ldots, J$ in market m is given by

$$u_{ijm} \quad = \quad \beta(Z_{im})^T X_{jm} - \alpha P_{jm} + \xi_{jm} + \varepsilon_{ijm},$$

where P_{jm} is the price of product j in market m and ε_{ijm} is an idiosyncratic error distibuted according to a Generalized Extreme Value Type-I (Gumbel) distribution. The utility of the outside good is $u_{i0m} = \varepsilon_{i0m}$.

We assume that each person chooses the good that provides the highest level of utility in which case the probability that individual i chooses good $j = 1, \ldots, J$ in market m is[4]

$$s_{jm}(Z_{im}) \quad = \quad \frac{\exp\left(\beta(Z_{im})^T X_{jm} - \alpha P_{jm} + \xi_{jm}\right)}{1 + \sum_{j'} \exp\left(\beta(Z_{im})^T X_{j'm} - \alpha P_{j'm} + \xi_{j'm}\right)}.$$

Averaging across individuals we obtain the market shares,

$$s_{jm} \quad = \quad \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{\exp\left(\beta(Z_{im})^T X_j - \alpha P_{jm} + \xi_{jm}\right)}{1 + \sum_{j'} \exp\left(\beta(Z_{im})^T X_{j'} - \alpha P_{j'm} + \xi_{j'm}\right)}.$$

4 Estimation

4.1 Standard Approach

The standard procedure utilizes the following algorithm:

[3] It is straightforward to generalize to the case when each market has a separate set of J_m products.

[4] See, for example, Nevo (2000).

1. Specify the parametric functional form of $\beta(Z) = g(Z, \gamma)$, where g is known and γ is a parameter.

2. Initialize the parameters to be estimated: $(\alpha, \gamma) = (\alpha_0, \gamma_0)$.

3. Iterate (α, γ) until convergence. At step n:

   (a) Solve for ξ_{jm}, $j = 1, \dots, J$, $m = 1, \dots, M$ by inverting the market shares—find $\hat{\xi}_{jm}$ such that the implied market shares $s_{jm}(\alpha_n, \gamma_n, \hat{\xi}_{jm})$ coincide with the observed market shares s_{jm} (BLP inversion).

   (b) Compute the cost function $L(\alpha_n, \gamma_n, \hat{\xi}_{jm})$ (usually based on generalized method of moments/minimum distance estimation).

   (c) Update the parameters $(\alpha, \gamma) = (\alpha_{n+1}, \gamma_{n+1})$.

4.2 Proposed Method

The method we propose in this paper is based on the following idea. If in each market the market shares for a given $Z = Z_0$ were observed, the problem would reduce to a simple discrete choice estimation. Indeed,

$$s_{jm}(Z_0) = \frac{\exp\left(\beta(Z_0)^T X_{jm} - \alpha P_{jm} + \xi_{jm}\right)}{1 + \sum_{j'} \exp\left(\beta(Z_0)^T X_{j'm} - \alpha P_{j'm} + \xi_{j'm}\right)}.$$

We can let $\beta = \beta(Z_0)$ and obtain the estimates $(\hat{\alpha}, \hat{\beta})$ using steps 2–3 from the algorithm described above. As $s_{jm}(Z_0)$ are unobserved, we attempt to estimate them using the individual level data, (Z_{im}, d_{im}).[5]

We propose the following algorithm:

1. Use a prediction method of choice to fit $s(z, j, m) = P\left(d_{im} = j | Z_{im} = z\right)$.

2. Pick Z_0 within the support of Z.[6]

[5] Another issue is estimating $\beta(Z)$ at $Z \neq Z_0$. We address this in the next section.

[6] The optimal choice of Z_0 is an important question that we leave for future research. In the simulations we use the median of the observed values for every dimension of Z.

3. Predict $\hat{s}(Z_0, j, m)$ for every $j = 1, \ldots, J$ and $m = 1, \ldots, M$.

4. Initialize the parameters to be estimated: $(\alpha, \beta) = (\alpha_0, \beta_0)$.

5. Iterate (α, β) until convergence. At step n:

    (a) Solve for ξ_{jm}, $j = 1, \ldots, J$, $m = 1, \ldots, M$ by inverting the predicted market shares—find $\hat{\xi}_{jm}$ such that the implied market shares $s_{jm}(\alpha_n, \beta_n, \hat{\xi}_{jm})$ coincide with the predicted market shares $\hat{s}_{jm}$.

    (b) Compute the cost function $L(\alpha_n, \beta_n, \hat{\xi}_{jm})$.

    (c) Update the parameters $(\alpha, \beta) = (\alpha_{n+1}, \beta_{n+1})$.

4.2.1 Estimating $\hat{\beta}(Z)$ for an Arbitrary Z

So far we have estimated $\hat{\beta}(Z)$ for a single evalue of $Z = Z_0$. However, $\hat{\beta}(Z)$ can be easily derived from $\hat{s}(Z, j, m)$ by solving a system of linear equations. As before, let $\beta = \beta(Z_0)$. Additionally, define $c_{jm}(Z) = (\beta(Z) - \beta)^T X_{jm}$. Then, $\beta(Z)^T X_{jm} = \beta^T X_{jm} + c_{jm}(Z)$ and

$$s_{jm}(Z_{im}) \quad = \quad \frac{w_{jm}(Z_{im}) \exp\left(\beta^T X_{jm} - \alpha P_{jm} + \xi_{jm}\right)}{1 + \sum_{j'} w_{j'm}(Z_{im}) \exp\left(\beta^T X_{j'm} - \alpha P_{j'm} + \xi_{j'm}\right)},$$

where $w_{jm}(Z) = \exp\left(c_{jm}(Z)\right)$.

Note that the set of equations

$$\hat{s}_{jm}(Z_0) \quad = \quad \frac{\exp\left(\hat{\beta}^T X_{jm} - \hat{\alpha} P_{jm} + \hat{\xi}_{jm}\right)}{1 + \sum_{j'} \exp\left(\hat{\beta}^T X_{j'm} - \hat{\alpha} P_{j'm} + \hat{\xi}_{j'm}\right)}$$

for $j = 1, \ldots, J$ and $m = 1, \ldots, M$ uniquely defines the values of $\exp\left(\hat{\beta}^T X_j - \hat{\alpha} P_{jm} + \hat{\xi}_{jm}\right)$ for all j and m. Let their estimates produced from $\hat{s}_{jm}(Z_0)$ be $\hat{E}_{jm}$. Then, we have the following system of linear equations:

$$\hat{s}_{jm}(Z) + \hat{s}_{jm}(Z) \sum_{j'=1}^{J} \hat{w}_{j'm}(Z) \hat{E}_{j'm} \quad = \quad \hat{w}_{jm}(Z) \hat{E}_{jm}$$

We can solve for $\hat{w}_{jm}(Z)$, which is equivalent to $\hat{c}_{jm}(Z)$. Finally, once we have $\hat{c}_{jm}(Z)$ for all j and m, we can use these values to recover $\hat{\beta}(Z)$ by regressing $\hat{c}_{jm}(Z)$ on X_{jm}.

5 Theoretical Results

There are two main results in this paper. The first one shows that if we have any consistent estimator of $s_{jm}(Z_0)$, the resulting procedure leads to a consistent estimator for α. Consequently, many of the known flexible machine learning algorithms for non-parametric probability estimation result in consistent estimators of α.

Second important result is the oracle property. That is, once the number of individuals per market is large relative to the number of markets, the distribution of the proposed estimator is the same as of the estimator obtained when the functional form of $\beta(Z)$ if known up to an additive constant. In other words, $\beta(Z) = \beta(Z_0) + f(Z)$, where $f(Z)$ is a known function and $\beta(Z_0)$ is a parameter to be estimated.

Theorem 1. *If $\hat{s}_{jm}(Z_0) \xrightarrow{p} s_{jm}(Z_0)$ uniformly over $m = 1, \ldots, M$ as $M \longrightarrow \infty$, then $\hat{\alpha}_M \xrightarrow{p} \alpha$ and $\hat{\beta} \xrightarrow{p} \beta$.*

Proof. Note that BLP inversion of the form $\xi(s; a, b)$ for fixed parameters (a, b) is a continuous function of s. Hence, if $\hat{s}_{jm}(Z_0) \xrightarrow{p} s_{jm}(Z_0)$, we have $\xi(\hat{s}_{jm}(Z_0); a, b) \xrightarrow{p} \xi(s_{jm}(Z_0); a, b)$ uniformly. Denote by $\Xi_{jm}(a, b) \xrightarrow{p} 0$ the difference between $\xi(\hat{s}_{jm}(Z_0); a, b)$ and $\xi(s_{jm}(Z_0); a, b)$. The moment driven objective function has the property:

$$\frac{1}{M}\sum_{m=1}^{M} g(\xi_{jm}, X_{jm}, P_{jm}, W_{jm}) \xrightarrow{p} 0, \tag{5.1}$$

and has two important properties: (i) it is continuous in ξ, (ii) it satisfies the identification assumption,[7] meaning that:

$$\frac{1}{M}\sum_{m=1}^{M} g(\xi(s_{jm}(Z_0); a, b), X_{jm}, P_{jm}, W_{jm}) \xrightarrow{p} 0, \quad \text{if } (a, b) = (\alpha, \beta), \tag{5.2}$$

[7] For the proof that steps 4–5 of the proposed procedure lead to consistent estimators as $M \longrightarrow \infty$ see, for example, Freyberger (2015).

and converges to something positive otherwise. We thus have:

$$\frac{1}{M}\sum_{m=1}^{M} g(\xi(\hat{s}_{jm}(Z_0); a, b), X_{jm}, P_{jm}, W_{jm}) =$$

$$\frac{1}{M}\sum_{m=1}^{M} g(\xi(s_{jm}(Z_0); a, b) + \Xi_{jm}(a, b), X_{jm}, P_{jm}, W_{jm}) \xrightarrow{p}$$

$$\lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} g(\xi(s_{jm}(Z_0); a, b), X_{jm}, P_{jm}, W_{jm}) \quad (5.3)$$

as $\Xi_{jm}(a, b)$ converges in probability to zero uniformly over m. To complete the proof we need to fix any neighborhood of (α, β) and to observe that the last equation implies that for large enough M $(\hat{\alpha}, \hat{\beta})$ must lie inside this neighborhood. □

There are a number of estimation procedures that guarantee the condition required by Theorem 1. The following result provides several sufficient conditions.

Lemma 1. *The following procedures provide a (uniformly over m) consistent estimator of $s_{jm}(Z_0)$:*

(i) Kernel ridge regression with the tuning parameter approaching zero.

(ii) Sieve estimator with the limiting space containing $s(Z)$.

(iii) AdaBoost with the appropriate stopping rule.

(iv) Sufficiently flexible neural network.

Proof. The result for (i) follows from Evgeniou, Pontil, and Poggio (2000) and Theorem 29.8 of Devroye, Györfi, and Lugosi (1996). Chen (2007) provides the technical conditions under which (ii) produces a consistent estimator. The results from Bartlett and Traskin (2007) specify the stopping rule that depends on the sample size and guarantees the convergence. Cybenko (1989) provides the results for (iv) for the case of a single hidden layer neural network with a sufficiently large number of neurons. □

Theorem 2. *Assume that $M, N \longrightarrow \infty$. Then the asymptotic distribution of $(\hat{\alpha}^{oracle}, \hat{\beta})$ is the same as the asymptotic distribution of the limiting proposed estimator, that is $\lim_{N\to\infty}(\hat{\alpha}^{ML}_{M,N}, \hat{\beta}^{ML}_{M,N})$.*

Proof. In fact, this statement is nothing more than saying that the values of unobservables, ξ_{jm}, are the same in population regardless of whether we use just a single point, $s_{jm}(Z_0)$, or the whole curve, $s_{jm}(\cdot)$. This is true as the whole curve is uniquely defined by the value at Z_0 and ξ_{jm}. In fact, that is exactly what we are going to show. Let us fix parameters (a, b). With just one point, we invert ξ^1_{jm} as a solution to the following equation:

$$\log(s_{jm}(Z_0)) - \log(s_{0m}(Z_0)) = b^T X_{jm} - aP_{jm} + \xi^1_{jm} \tag{5.4}$$

If we invert from the full curve, then ξ^o_{jm} is set to match the share on average. If $h_m(Z)$ is the density of consumers in market m, this results in equation:

$$\int s_{jm}(Z)h_m(Z)\,dZ = \int \frac{\exp\left((b + f(Z))^T X_{jm} - aP_{jm} + \xi^o_{jm}\right)}{1 + \sum_{j'} \exp\left((b + f(Z))^T X_{j'm} - aP_{j'm} + \xi^o_{j'm})\right)} h_m(Z)\,dZ \tag{5.5}$$

However, from the structural form, for any Z we have:

$$\log(s_{jm}(Z)) - \log(s_{0m}(Z)) = \log(s_{jm}(Z_0)) - \log(s_{0m}(Z_0)) + f(Z)^T X_{jm} \tag{5.6}$$

As a result, once we've matched $s_{jm}(Z_0)$, we've matched the whole curve perfectly and ξ^1_{jm} solves integral equality above pointwise. Thus, $\xi^o_{jm} = \xi^1_{jm}$. □

6 Simulations

To illustrate the performance differences of different methods we consider a setting with $J = 2$ goods (plus the outside good) and $M = 50$ separate markets. Each market is populated by $N = 1000$ consumers. Both X_{jm} and Z_{im} are one-dimensional ($p = k = 1$) and there price is exogenous.[8]

We let $\beta(Z) = \gamma_0 + \gamma_1 Z + \gamma_2 Z^2$ and consider three different cases: (i) the standard approach when $\beta(Z)$ is misspecified and $\tilde{\beta}(Z) = \tilde{\gamma}_0 + \tilde{\gamma}_2 Z^2$ is used instead, (ii) the standard approach when $\beta(Z)$ is correctly specified, and (iii) the proposed algorithm. The true value of the price

[8] This is done for simplicity to ignore the instruments, W_{jm}, but does not affect the results.

coefficient, α, is 1.

We use the minimum distance estimator with the following moments:

1. The covariance between X_{jm} and $\hat{\xi}_{jm}$ has to be close to zero.

2. The covariance between P_{jm} and $\hat{\xi}_{jm}$ has to be close to zero.

3. The covariance between Z_{im} and $X_{d_{im}m}$ (where $X_{d_{im}m}$ denotes the characteristics of the good chosen by individual i in market m) has to be close to the covariance between Z_{im} and $\sum_j \hat{s}_{jm}(Z_{im})X_{jm}$ (only for the oracle case).

4. The covariance between Z_{im}^2 and $X_{d_{im}m}$ has to be close to the covariance between Z_{im}^2 and $\sum_j \hat{s}_{jm}(Z_{im})X_{jm}$ (only for the misspecified and the oracle cases).

For the first step of the algorithm we use the kernel ridge regression as presented in Murphy (2012).

We estimate the model $B = 50$ times and obtain an estimate $\hat{\alpha}_b$ at each iteration $b = 1, \ldots, B$. The distributions of these values are shown in Figure 1. In the plot, "Misspecified" refers to case (i), "Oracle" to case (ii), and "Proposed" to case (iii).

6.1 Precision

As reported in Table 1, the proposed procedure performs well compared to the oracle case in terms of both—the bias and the root-mean-square error—while the misspecified procedure produces the estimated that are very imprecise.

Bias and Root-Mean-Square Error of the Algorithms			
	Misspecified	Proposed	Oracle
Bias	0.42	0.01	0.07
RMSE	0.42	0.07	0.21

Table 1: Bias and Root-Mean-Square Error

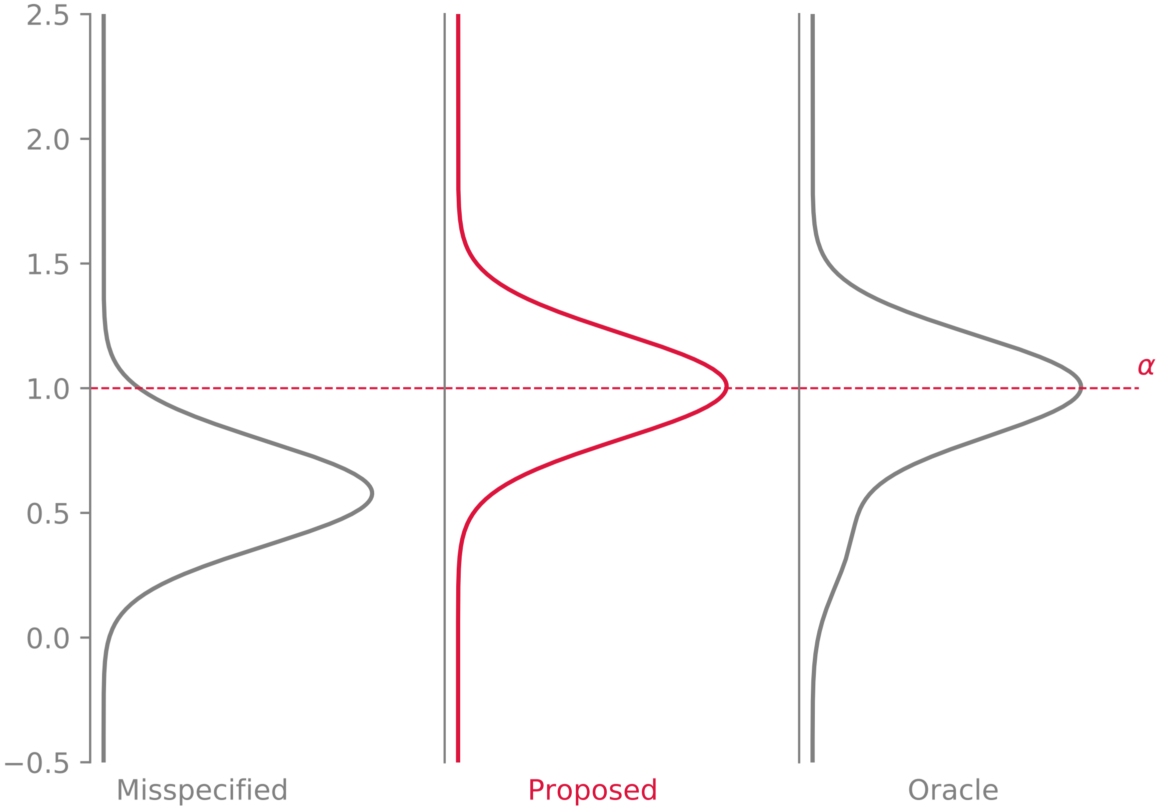

Figure 1: Simulated Distributions of the Price Coefficient

6.2 Computational Costs

One of the main advantages of our procedure is its computational simplicity. Even for a single dimensional individual level characteristic, Z, it converges on average almost two times faster than the misspecified procedure (which requires just a single additional parameter) and almost four times faster than the correct specification (two additional parameters). When the dimensionality of Z or the complexity of $\beta(Z)$ as a function of Z increase, the difference becomes even more substantial.

Computational Times of the Algorithms			
	Misspecified	Proposed	Oracle
Average time	191 s	107 s	406 s
Standard deviation	164 s	105 s	317 s

Table 2: Computational Costs

7 Conclusion

The method of estimation that we propose is likely to outperform the approach commonly used in the literature unless the correct functional form is known to the researcher. We also show that the proposed estimator has nice large sample properties and converges substantially faster than the alternatives.

There are several potential directions for future research. First, the optimal choice of Z_0 should be investigated. As different markets may be heterogeneous in terms of the distributions of Z, the precision of the estimates may suffer from the choice of Z_0 that is the same for every market. Second, most of the machine learning procedures require the choice of a tuning parameter (or parameters) that was largely ignored in this paper. Third, the statistical properties of the relevant hypotheses tests may be investigated. Finally, it would be useful to see how the proposed method performs in applications compared to the alternative procedures.

References

Athey, S., D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt (2018). Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. arXiv preprint arXiv:1801.07826.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113(27), 7353–7360.

Athey, S., G. W. Imbens, and S. Wager (2016). Efficient inference of average treatment effects in high dimensions via approximate residual balancing. Technical report.

Bartlett, P. L. and M. Traskin (2007). Adaboost is consistent. Journal of Machine Learning Research 8(Oct), 2347–2368.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2011a). Inference for high-dimensional sparse econometric models. arXiv preprint arXiv:1201.0220.

Belloni, A., V. Chernozhukov, and C. Hansen (2011b). Lasso methods for gaussian instrumental variables models.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives 28(2), 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies 81(2), 608–650.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. Econometrica: Journal of the Econometric Society, 841–890.

Berry, S. T. and P. A. Haile (2014). Identification in differentiated products markets using market level data. Econometrica 82(5), 1749–1797.

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. Proceedings of the National Academy of Sciences 113(27), 7383–7390.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. Handbook of econometrics 6, 5549–5632.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017). Double/debiased machine learning for treatment and causal parameters. Technical report.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach.

Compiani, G. (2018). Nonparametric demand estimation in differentiated products markets.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2(4), 303–314.

Devroye, L., L. Györfi, and G. Lugosi (1996). A probabilistic theory of pattern recognition, Volume 31. Springer Science & Business Media.

Dunker, F., S. Hoderlein, and H. Kaido (2017). Nonparametric identification of random coefficients in endogenous and heterogeneous aggregate demand models. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Evgeniou, T., M. Pontil, and T. Poggio (2000). Regularization networks and support vector machines. Advances in computational mathematics 13(1), 1.

Fan, Q. (2012). The adaptive lasso method for instrumental variable selection. North Carolina State University.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics 189(1), 1–23.

Freyberger, J. (2015). Asymptotic theory for differentiated products demand models with many markets. Journal of Econometrics 185(1), 162–181.

Friedman, J., T. Hastie, and R. Tibshirani (2001). The elements of statistical learning, Volume 1. Springer series in statistics New York.

Gautier, E. and A. Tsybakov (2011). High-dimensional instrumental variables regression and confidence sets. arXiv preprint arXiv:1105.2454.

Gillen, B., S. Montero, H. Moon, and M. Shum (2015). Blp-lasso for aggregate discrete-choice models applied to elections with rich demographic covariates. Technical report, Working Paper. California Institute of Technology.

Gillen, B. J., M. Shum, and H. R. Moon (2014). Demand estimation with high-dimensional product characteristics. In Bayesian Model Comparison, pp. 301–323. Emerald Group Publishing Limited.

Hansen, C. and D. Kozbur (2014). Instrumental variables estimation with many weak instruments using regularized jive. Journal of Econometrics 182(2), 290–308.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2016). Counterfactual prediction with deep instrumental variables networks. arXiv preprint arXiv:1612.09596.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Murphy, K. P. (2012). Machine learning - a probabilistic perspective. In Adaptive computation and machine learning series.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. Journal of economics & management strategy 9(4), 513–548.

Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association (just-accepted).