


# Online Appendix for “Identification and Estimation of Discrete Choice Models with Unobserved Choice Sets”\*

Victor H. Aguiar  Nail Kashaev<sup>†</sup>



February 29, 2024

**Abstract** This online appendix contains the proofs of all theoretical results, simulations, and omitted details from the empirical application.

JEL classification numbers: C14, C5, D6

Keywords: random utility, discrete choice, random consideration sets, best subset regression

---

\*The “” symbol indicates that the authors’ names are in certified random order, as described by Ray  Robson (2018).

<sup>†</sup>Aguiar: Department of Economics, Simon Fraser University; [vaguiarl@sfu.ca](mailto:vaguiarl@sfu.ca). Kashaev: Department of Economics, University of Western Ontario; [nkashaev@uwo.ca](mailto:nkashaev@uwo.ca).

## A. Proofs

### A.1. Proof of Lemma 1

Assume that  $K = 1$ . Also, to simplify the exposition, we drop  $\mathbf{D}$  and  $\mathbf{x}$  and write  $\mathbb{P}(y_s)$  instead of  $\mathbb{P}(\mathbf{y}_s = y_s)$ . First, note that after applying several times Bayes' theorem and Assumption 2, we get that

$$\begin{aligned} \mathbb{P}(y_1 | y_4, y_3, y_2) &= \frac{\mathbb{P}(y_4, y_3, y_2, y_1)}{\mathbb{P}(y_4, y_3, y_2)} = \frac{\mathbb{P}(y_4 | y_3, y_2, y_1) \mathbb{P}(y_3 | y_2, y_1) \mathbb{P}(y_2, y_1)}{\mathbb{P}(y_4 | y_3, y_2) \mathbb{P}(y_3 | y_2) \mathbb{P}(y_2)} \\ &= \frac{\mathbb{P}(y_4 | y_3) \mathbb{P}(y_3 | y_2) \mathbb{P}(y_2, y_1)}{\mathbb{P}(y_4 | y_3) \mathbb{P}(y_3 | y_2) \mathbb{P}(y_2)} = \frac{\mathbb{P}(y_2, y_1)}{\mathbb{P}(y_2)} = \mathbb{P}(y_1 | y_2). \end{aligned}$$

Similarly,  $\mathbb{P}(y_1 | y_4, y_2) = \mathbb{P}(y_1 | y_2)$ . Hence,

$$\begin{aligned} \mathbb{P}(y_3 | y_4, y_2, y_1) &= \frac{\mathbb{P}(y_3, y_1 | y_4, y_2)}{\mathbb{P}(y_1 | y_4, y_2)} = \frac{\mathbb{P}(y_1 | y_4, y_3, y_2) \mathbb{P}(y_3 | y_4, y_2)}{\mathbb{P}(y_1 | y_2)} \\ &= \frac{\mathbb{P}(y_1 | y_2) \mathbb{P}(y_3 | y_4, y_2)}{\mathbb{P}(y_1 | y_2)} = \mathbb{P}(y_3 | y_4, y_2). \end{aligned}$$

As a result,

$$\begin{aligned} \mathbb{P}(y_5, y_3, y_1 | y_4, y_2) &= \mathbb{P}(y_5 | y_4, y_3, y_2, y_1) \mathbb{P}(y_3 | y_4, y_2, y_1) \mathbb{P}(y_1 | y_4, y_2) \\ &= \mathbb{P}(y_5 | y_4, y_2) \mathbb{P}(y_3 | y_4, y_2) \mathbb{P}(y_1 | y_4, y_2). \end{aligned}$$

Thus,  $\mathbf{y}_5$ ,  $\mathbf{y}_3$ , and  $\mathbf{y}_1$  are conditionally independent conditional on  $\mathbf{y}_4$  and  $\mathbf{y}_2$ . For  $K > 1$  one just need to relabel  $y_5$  as  $y_{2K+3}$ ,  $y_4$  as  $y_{2K+2}$ ,  $y_3$  as  $y(\mathcal{S}_2)$ ,  $y_2$  as  $y_{K+1}$ , and  $y_1$  as  $y(\mathcal{S}_1)$ , and apply the above arguments.

## A.2. Proof of Theorem 1

Fix some  $x, y, y'$ , and  $s = S$ . Take two disjoint subsets of size  $K$ ,  $\mathcal{S}_1$  and  $\mathcal{S}_2$ , that do not contain  $s$ . Let  $g : \mathcal{Y}^K \rightarrow \bar{\mathcal{Y}} = \{1, 2, \dots, Y^K\}$  be any one-to-one mapping. Define two random variable on  $\bar{\mathcal{Y}}$ :  $\mathbf{z}_1 = g(\mathbf{y}(\mathcal{S}_1))$  and  $\mathbf{z}_2 = g(\mathbf{y}(\mathcal{S}_2))$ . To simplify the exposition, we drop  $x, y, y'$ , and  $\mathcal{S}_i$  from the notation. All the probabilities below are defined conditional on  $\mathbf{x} = x, \mathbf{y}_{K+1} = y$ , and  $\mathbf{y}_{2K+2} = y'$ . Define the following matrices

$$\begin{aligned} L_{1,2} &= [\mathbb{P}(\mathbf{z}_1 = i, \mathbf{z}_2 = j)]_{i,j \in \bar{\mathcal{Y}}}, \\ L_{1|D} &= [\mathbb{P}(\mathbf{z}_1 = i \mid \mathbf{D} = D_k)]_{i \in \bar{\mathcal{Y}}, k=1, \dots, d_D}, \\ L_{2|D} &= [\mathbb{P}(\mathbf{z}_2 = i \mid \mathbf{D} = D_k)]_{i \in \bar{\mathcal{Y}}, k=1, \dots, d_D}, \\ A_D &= \text{diag}((\mathbb{P}(\mathbf{D} = D_k))_{k=1, \dots, d_D}) = \text{diag}((m(D_k))_{k=1, \dots, d_D}), \end{aligned}$$

where  $\text{diag}(z)$  is a diagonal matrix with vector  $z$  on the diagonal.

*Step 1.* In this step, we show how to identify the number of choice sets that are considered with positive probability. By the law of total probability, Lemma 1 (recall that we are also conditioning on  $\mathbf{x} = x, \mathbf{y}_{K+1} = y$ , and  $\mathbf{y}_{2K+2} = y'$ ) implies that

$$\begin{aligned} \mathbb{P}(\mathbf{z}_1 = i, \mathbf{z}_2 = j) &= \sum_k \mathbb{P}(\mathbf{z}_1 = i, \mathbf{z}_2 = j \mid \mathbf{D} = D_k) \mathbb{P}(\mathbf{D} = D_k) \\ &= \sum_k \mathbb{P}(\mathbf{z}_1 = i \mid \mathbf{D} = D_k) \mathbb{P}(\mathbf{z}_2 = j \mid \mathbf{D} = D_k) \mathbb{P}(\mathbf{D} = D_k), \end{aligned}$$

or in matrix notation

$$L_{1,2} = L_{1|D} A_D L_{2|D}^\top.$$

Under Assumption 5 the maximal number of the points in the support of  $\mathbf{D}$  is equal to the number of the possible outcomes. That is,  $d_D \leq |\bar{\mathcal{Y}}| = Y^K$  (recall that the number of alternatives is  $Y$ ).

Next, note that Assumption 5 implies that  $L_{1|D}$  and  $L_{2|D}$  have full column rank ( $d_D$ ).

Hence, using the properties of the rank operator we can conclude that

$$\text{rank}(L_{1,2}) = \text{rank}(L_{1|D}A_D L_{2|D}^\top) = \text{rank}(A_D L_{2|D}^\top) = \text{rank}(A_D) = d_D.$$

That is, the rank of  $L_{1,2}$  is equal to  $d_D = |\mathcal{D}_x|$ . Hence, since  $L_{1,2}$  is observed (can be consistently estimated), we can identify (consistently estimate) the number of choice sets that DMs are using.

*Step 2.* Knowing  $d_D$  and the fact that  $L_{1|D}$  and  $L_{2|D}$  have full column rank, we take a collection of alternatives in  $\bar{\mathcal{Y}}$ ,  $\{\tilde{z}_k\}_{k=1}^{d_D}$ , such that the following observable modification of  $L_{1,2}$  is nonsingular (have full rank):

$$\tilde{L}_{1,2} = [\mathbb{P}(\mathbf{z}_1 = \tilde{z}_i, \mathbf{z}_2 = \tilde{z}_j)]_{i,j \in \{1, \dots, d_D\}}.$$

Such collection  $\{\tilde{z}_k\}_{k=1}^{d_D}$  always exists since one can always find  $d_D$  linearly independent rows of  $L_{1|D}$ . Indeed, similar to Step 1

$$\tilde{L}_{1,2} = \tilde{L}_{1|D}A_D\tilde{L}_{2|D}^\top,$$

where

$$\tilde{L}_{1|D} = [\mathbb{P}(\mathbf{z}_1 = \tilde{z}_i | \mathbf{D} = D_k)]_{i,k \in \{1, \dots, d_D\}},$$

$$\tilde{L}_{2|D} = [\mathbb{P}(\mathbf{z}_2 = \tilde{z}_i | \mathbf{D} = D_k)]_{i,k \in \{1, \dots, d_D\}}.$$

Since  $\tilde{L}_{1|D}$  and  $\tilde{L}_{2|D}$  are nonsingular, it implies that  $\tilde{L}_{1,2}$  is nonsingular as well ( $A_D$  has rank  $d_D$ ).

*Step 3.* This step is based on [Hu \(2008\)](#) and [Hu et al. \(2013\)](#). Fix some  $y \in \mathcal{Y}$  and define

$$\tilde{L}_{1,D} = [\mathbb{P}(\mathbf{z}_1 = \tilde{z}_i, \mathbf{D} = D_k)]_{i,k \in \{1, \dots, d_D\}},$$

$$\tilde{L}_{2,1,y} = [\mathbb{P}(\mathbf{z}_2 = \tilde{z}_i, \mathbf{z}_1 = \tilde{z}_j, \mathbf{y}_t = y)]_{i,j \in \{1, \dots, d_D\}},$$

$$A_{y|D} = \text{diag} \left( (\mathbb{P}(\mathbf{y}_s = y | \mathbf{D} = D_k))_{k \in \{1, \dots, d_D\}} \right) = \text{diag} \left( (F^{\text{RUM}}(y | D_k))_{k \in \{1, \dots, d_D\}} \right).$$

By the law of total probability, Lemma 1 implies that

$$\begin{aligned} \mathbb{P}(\mathbf{z}_1 = \tilde{z}_i, \mathbf{z}_2 = \tilde{z}_j) &= \sum_k \mathbb{P}(\mathbf{z}_1 = \tilde{z}_i, \mathbf{z}_2 = \tilde{z}_j | \mathbf{D} = D_k) \mathbb{P}(\mathbf{D} = D_k) \\ &= \sum_k \mathbb{P}(\mathbf{z}_2 = \tilde{z}_j | \mathbf{D} = D_k) \mathbb{P}(\mathbf{z}_1 = \tilde{z}_i, \mathbf{D} = D_k). \end{aligned}$$

Hence, in matrix notation we get

$$\tilde{L}_{1,2}^\top = \tilde{L}_{2|D} \tilde{L}_{1,D}^\top.$$

Since, by construction in Step 2,  $\tilde{L}_{2|D}$  is nonsingular, we have that

$$\tilde{L}_{1,D}^\top = \tilde{L}_{2|D}^{-1} \tilde{L}_{1,2}^\top. \quad (1)$$

Similarly to the previous calculations, Lemma 1 implies that

$$\tilde{L}_{2,1,y} = \tilde{L}_{2|D} A_{y|D} \tilde{L}_{1,D}^\top.$$

Combining the latter with equation (1) we get the following eigenvector-eigenvalue decomposition of  $R_y = \tilde{L}_{2,1,y} (\tilde{L}_{1,2}^\top)^{-1}$

$$R_y = \tilde{L}_{2|D} A_{y|D} \tilde{L}_{2|D}^{-1}. \quad (2)$$

*Step 4.* Note that in the decomposition (2) the change in  $y$  does not affect eigenvectors of  $R_y$ , but affects its eigenvalues. For  $R_y$  let  $\{(\eta_k, \lambda_{y,k})\}_{k=1}^{d_D}$  denote the set of its eigenvectors and eigenvalues. To pin down eigenvectors uniquely note that it suffices to pick those that belong to a simplex (each one of them should sum up to 1). In contrast to the existing results (e.g. Hu et al., 2013), we do not use these eigenvectors to identify  $L_{2|D}$  since  $\tilde{L}_{2|D}$  is only a submatrix of  $L_{2|D}$ .

Take  $y = 1$  and fix the set of eigenvectors of  $R_1$ ,  $\{\eta_k\}_{k=1}^{d_D}$ . Stack them in any order to form

matrix  $\Lambda_1$ . Then we can compute

$$A_{y|D}^* = \Lambda_1^{-1} R_y \Lambda_1$$

for every  $y$ . Since the order of eigenvalues is fixed, the diagonal entries of  $A_{y|D}^*$  correspond to the same sets. Note that  $y \in D$  if and only if  $F^{\text{RUM}}(y|D) > 0$ . Thus, we can identify the identity of choice sets and  $F^{\text{RUM}}(y|D)$  for every  $y$  and  $D$ . Hence, we can identify the conditional distributions of  $\mathbf{z}_1$  and  $\mathbf{z}_2$  conditional on  $\mathbf{D}$ . Thus, we identify  $L_{2|D}$ .

*Step 5.* Finally, let  $m = (m(D_k))_{k \in \{1, \dots, d_D\}}$ , then

$$m = L_{1,D}^\top \cdot \iota,$$

where  $\iota$  is the vector of ones. Hence,

$$L_{1,2}^\top \iota = L_{2|D} L_{1,D}^\top \cdot \iota = L_{2|D} m.$$

Since  $L_{1,2}$  is observed (can be consistently estimated), and  $L_{2|D}$  is constructively identified and has full column rank, we also identify the distribution of choice sets. The fact that the choice of  $x$ ,  $y$ , and  $y'$  was arbitrary completes the proof.

### A.3. Proof of Proposition 1

Fix some  $x \in X$ ,  $y, y' \in \mathcal{Y}$ , and  $\mathcal{S}_i$ . To simplify the exposition, we drop  $x$ ,  $y$ ,  $y'$ , and  $\mathcal{S}_i$  from the notation. Note that if the linear independence condition is satisfied when  $\mathcal{D} = \{A \cup \{y, y'\} : A \subseteq \mathcal{Y}\}$ , then it is automatically satisfied for any smaller  $\mathcal{D}$ . Hence, without loss of generality, we assume that  $\mathcal{D} = \{A \cup \{y, y'\} : A \subseteq \mathcal{Y}\}$ .

First, order all elements in  $\mathcal{D}$  according to their cardinality (with arbitrary order among sets with the same cardinality). For any two sets  $D_k$  and  $D_m$  such that  $k < m$ ,  $D_m$  is never a subset of  $D_k$ .

Given the above order over elements in  $\mathcal{D}$ , consider the following sequence of  $2^{Y-2}$  vectors in  $\mathcal{Y}^K$ . For any  $D_k$  take  $y_k^K$  such that every element in  $D_k$  is some component of  $y_k^K$ . Since  $K \geq Y$  and  $|D_k| \leq Y$  such  $y_k^K$  always exists. Moreover, for any  $D_k \neq D_l$  it has to be true that  $y_k^K \neq y_l^K$ .

Consider the matrix  $\mathcal{G}$  of size  $2^{Y-2} \times 2^{Y-2}$  such that  $(j, k)$ -element of it is

$$\mathcal{G}_{j,k} = G(y_j^K | D_k).$$

Note that this matrix is upper-triangular. Indeed, take any  $j, k$  such that  $j > k$ . Since  $j > k$ , then  $D_j$  is never a subset of  $D_k$ . Hence, there exists a component of  $y_j^K$  that is not an element of  $D_k$ . This means that the probability of observing a sequence  $y_j^K$  given set  $D_k$  is zero. That is,  $\mathcal{G}_{j,k} = 0$  if  $j > k$ . Assumption 4 implies that the diagonal elements,  $\mathcal{G}_{j,j}$ , are nonzero since any element of  $D_j$  can be observed with positive probability. Since  $\mathcal{G}$  is upper-triangular with nonzero diagonal elements, it is of full column rank (the determinant of  $\mathcal{G}$  equals to the product of the diagonal elements).

Adding more rows to  $\mathcal{G}$  does not change its column rank. Hence, the linear independence condition is satisfied. The fact that the choice of  $x, y, y'$ , and  $\mathcal{S}_i$  was arbitrary completes the proof.

#### A.4. Proof of Proposition 2

(i). Fix some  $x \in X$ ,  $y, y' \in \mathcal{Y}$ , and  $\mathcal{S}_i$ . To simplify the exposition, we drop  $x, y, y'$ , and  $\mathcal{S}_i$  from the notation. Note that if the linear independence condition is satisfied when  $\mathcal{D} = \{\{y, y'\}, \{y, y', y_1\}, \{y, y', y_1, y_2\}, \{y, y', y_1, y_2, y_3\}, \dots, \mathcal{Y}\}$ , then it is automatically satisfied for any smaller  $\mathcal{D}$ . So, without loss of generality, we assume that  $|\mathcal{D}| = Y - 1$ .

Nestedness implies that  $D_k \subseteq D_{k+1}$  for all  $k$ . Let  $\{y_j^K\}_{j=1}^Y$  be a sequence in  $\mathcal{Y}^K$  such that  $y_j^K = (y_j, y_j, \dots, y_j)^\top$ . Recall that  $D_k = \{y, y', y_1, y_2, \dots, y_k\}$ . Consider the matrix  $\mathcal{G}$  of size

$Y - 1 \times Y - 1$  such that  $(j, k)$ -element of it is

$$\mathcal{G}_{j,k} = G(y_j^K \mid D_k).$$

Note that this matrix is upper-triangular. Indeed, take any  $j, k$  such that  $j > k$ . Hence, none of the components of  $y_j^K$  are elements of  $D_k$ . This means that the probability of observing a sequence  $y_j^K$  given set  $D_k$  is zero. That is,  $\mathcal{G}_{j,k} = 0$  if  $j > k$ . Assumption 4 implies that the diagonal elements,  $\mathcal{G}_{j,j}$ , are nonzero since any element of  $D_j$  can be observed with positive probability. Since  $\mathcal{G}$  is upper-triangular with nonzero diagonal elements, it is of full column rank (the determinant of  $\mathcal{G}$  equals to the product of the diagonal elements). Adding more rows to  $\mathcal{G}$  does not change its column rank. Hence, the linear independence condition is satisfied. The fact that the choice of  $x, y, y'$ , and  $\mathcal{S}_i$  was arbitrary completes the proof.

(ii). Fix some  $x \in X, y, y' \in \mathcal{Y}$ , and  $\mathcal{S}_i$ . To simplify the exposition, we drop  $x, y, y'$ , and  $\mathcal{S}_i$  from the notation. Let  $\{y_k^K\}_{k=1}^{|\mathcal{D}|}$  be a sequence in  $\mathcal{Y}^K$  such that  $y_k^K = (y_k, y_k, \dots, y_k)^\top$ , where  $y_k \in \mathcal{Y}$  are from condition (ii) of the proposition. Consider the matrix  $\mathcal{G}$  of size  $|\mathcal{D}| \times |\mathcal{D}|$  such that  $(j, k)$ -element of it is

$$\mathcal{G}_{j,k} = G(y_j^K \mid D_k).$$

Note that this matrix is diagonal. Indeed, take any  $j, k$  such that  $j \neq k$ . Hence, none of the components of  $y_j^K$  are elements of  $D_k$ . This means that the probability of observing a sequence  $y_j^K$  given set  $D_k$  is zero. That is,  $\mathcal{G}_{j,k} = 0$  if  $j \neq k$ . Assumption 4 implies that the diagonal elements,  $\mathcal{G}_{j,j}$ , are nonzero since any element of  $D_j$  can be observed with positive probability. Since  $\mathcal{G}$  is diagonal with nonzero diagonal elements, it is of full column rank (the determinant of  $\mathcal{G}$  equals to the product of the diagonal elements). Adding more rows to  $\mathcal{G}$  does not change its column rank. Hence, the linear independence condition is satisfied. The fact that the choice of  $x, y, y'$ , and  $\mathcal{S}_i$  was arbitrary completes the proof.



## B. Finite-sample Performance of the Estimator

This section aims to analyze the finite sample performance of the estimator we propose in Section 4. First, we define the data generating processes (DGPs) used in simulations. In all experiments we assume that there are no covariates, Assumption 2' is satisfied, and that  $S = 3$  (this setup is equivalent to the one with choices depending on the previous choice and  $S = 5$ ). There are  $Y = 5$  alternatives and  $d_{\mathcal{D}} = 5$  choice sets. Every DGP is characterized by two matrices:  $Pyd \in \mathbb{R}^{5 \times 5}$  and  $Pd \in \mathbb{R}^5$ .  $Pyd$  and  $Pd$  are such that  $Pyd_{y,j} = \mathbb{P}(\mathbf{y}_s = y \mid D_j)$  and  $Pd_j = \mathbb{P}(\mathbf{D} = D_j)$ . In other words, every column of  $Pyd$  corresponds to a choice set. For instance, the fourth column of  $Pyd$  in DGP1 indicates that the fourth choice set has two elements,  $\{1, 4\}$ , and that conditional on  $\mathbf{D} = \{1, 4\}$  alternative 1 is picked with probability 0.4. Since the fifth element of  $Pd$  is 0.15, the probability of considering  $\{1, 4\}$  is 0.15.

**DGP1:**

$$Pyd = \begin{pmatrix} 1 & 0.6 & 0.5 & 0.4 & 0.2 \\ 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0.8 \end{pmatrix}, \quad Pd = \begin{pmatrix} 0.2 \\ 0.15 \\ 0.3 \\ 0.15 \\ 0.2 \end{pmatrix}.$$

**DGP2:**

$$Pyd = \begin{pmatrix} 1 & 0.6 & 0.5 & 0.25 & 0.1 \\ 0 & 0.4 & 0.2 & 0.35 & 0.25 \\ 0 & 0 & 0.3 & 0.25 & 0.15 \\ 0 & 0 & 0 & 0.15 & 0.3 \\ 0 & 0 & 0 & 0 & 0.2 \end{pmatrix}, \quad Pd = \begin{pmatrix} 0.2 \\ 0.15 \\ 0.3 \\ 0.15 \\ 0.2 \end{pmatrix}.$$

The cardinality of the choice sets in DGP1 does not vary much. It assigns positive probability to the choice sets of cardinality less than 3 only. For example, alternative 1 enters all choice

sets.<sup>1</sup> All other options enter only one choice set. DGP2 is very heterogeneous in terms of the size of the choice sets. The smallest one contains only alternative 1. The biggest one contains all alternatives. Note that these DGPs satisfy conditions of Proposition 2 (Excluded Choices and Nestedness).

We evaluate the ability of the Step-1 and Step-2 estimators to correctly recover sets. The results are presented in Table 1. As expected, the performance of both estimators improves with the sample size. The Step-2 estimator outperforms the Step-1 estimator in all experiments. The gains from the Step-2 estimator are especially striking for DGP2. For example, for a sample size of 2000 the Step-1 estimator correctly recovers all 5 sets only in about 4 percent of cases. However, the Step-2 estimator recovers all 5 sets in 34 percent of cases.

**Table 1** – Percent of Correctly Estimated Sets

Sample Size		2000	5000	10000	50000
DGP1	Step-1	63.7	73.9	19.8	95.9
	Step-2	68.6	86.6	93.1	99.8
DGP2	Step-1	3.7	12.1	25.5	75.7
	Step-2	34	47	62.8	93.8

Notes: Number of replications=1000,  $\varepsilon = 0.01$ . Results are rounded to 1 digit.

To see how noisy the estimates of choice sets can be we also compute the average number of correctly recovered sets. As Table 2 shows, our estimator on average finds correctly more than 4 out of 5 sets. The Step-1 estimator recovers at least 3 out of 5 sets correctly on average. The performance of both estimators improves with the sample size.

Tables 3-10 present the bias and the mean-squared-error of estimation of  $m$  and  $F^{\text{RUM}}$ . Every experiment was conducted 1000 times. Since some elements of matrix  $Pyd$  are zeros and elements of every column sum up to one, for estimates of  $F^{\text{RUM}}$  we only report the nonzero, linearly independent elements. For instance, for DGP1 we only report estimates of (1,2), (1,3), (1,4), and (1,5) elements of  $Pyd$ . The estimators perform well even in the samples of a moderate size. As expected, both the bias and the root-mean-squared-error decrease with the

<sup>1</sup>We impose this restriction in the estimation step.

**Table 2** – Average Number of Correctly Estimated Sets

Sample Size		2000	5000	10000	50000
DGP1	Step-1	4.66	4.74	4.8	4.96
	Step-2	4.65	4.86	4.93	5
DGP2	Step-1	3.36	3.62	3.91	4.75
	Step-2	4.18	4.41	4.6	4.93

Notes: Number of replications=1000,  $\varepsilon = 0.01$ , number of sets=5. Results are rounded to 2 digits.

sample size.

**Table 3** – Bias in Estimating  $m$ . DGP1 ( $\times 10^{-5}$ )

Sample Size\Set	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
2000	-112.7	35.2	43.8	35.7	-2
5000	-30.8	19.1	8.7	1.9	1.1
10000	-2.6	-7.7	8.2	-1.2	3.2
50000	4.2	-12	-0.0	6.3	1.6

Notes: Number of replications=1000. Results are rounded to 6 digits.

**Table 4** – Root Mean Squared Error in Estimating  $m$ . DGP1 ( $\times 10^{-3}$ )

Sample Size\Set	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
2000	13.1	10.5	12.5	8.5	9.2
5000	8	6.6	7.5	5.4	5.9
10000	5.8	4.8	5.2	3.7	4.2
50000	2.7	2.1	2.4	1.7	1.9

Notes: Number of replications=1000. Results are rounded to 4 digits.

**Table 5** – Bias in Estimating  $m$ . DGP2 ( $\times 10^{-3}$ )

Sample Size\Set	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
2000	-20.2	-8.9	-67.7	55.4	41.4
5000	-13.3	-7.4	-70.5	59.9	31.2
10000	-7.4	-1.8	-50.6	41.4	18.4
50000	0.6	0.4	-6.7	4.5	2.4

Notes: Number of replications=1000. Results are rounded to 4 digits.

**Table 6** – Root Mean Squared Error in Estimating  $m$ . DGP2 ( $\times 10^{-2}$ )

Sample Size\Set	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
2000	5.7	4.8	11	11.2	4.9
5000	4.2	3.3	11	10.3	3.9
10000	2.8	1.8	8.4	7.6	2.6
50000	0.3	0.5	1.5	1.2	0.6

Notes: Number of replications=1000. Results are rounded to 3 digits.

## C. Application Details Omitted From the Main Text

### C.1. Data Construction

We consider  $Y = 5$  brands of RTE cereal: Store brand (CTL), General Mills (GM), Kellogg (K), Quaker (Q), and other brands of RTE cereal (O). We record only purchases of households that buy 1 brand per trip.<sup>2</sup> We focus on households that are frequent buyers. We define frequent buyers as households that buy at least one RTE cereal in  $S = 3$  consecutive trips.<sup>3</sup> The majority of households in our sample makes 1 trip per week. Thus, the predominant time frequency of our dataset is weekly. We focus on trips and households present in the Homescan in 2016-2018.<sup>4</sup> We include only the 3 earliest consecutive trips per household. Each

<sup>2</sup>The households that buy more than 1 brand in a given trip are dropped from the sample to avoid dealing with bundling.

<sup>3</sup>A trip is an instance of a household member going to a store and purchasing at least one item that is recorded in the Homescan.

<sup>4</sup>We eliminate from our sample trips happening in December and January, because of their strong seasonality effects on RTE cereal consumption.

**Table 7** – Bias in Estimating  $F^{\text{RUM}}$ . DGP1  
( $\times 10^{-4}$ )

Sample Size \ (i, j)	1, 2	1, 3	1, 4	1, 5
2000	-1.0	-2.7	0.4	1.8
5000	-3.8	4.4	-8.5	-1.3
10000	-1.1	1.6	-2.9	0.6
50000	2.1	2.1	0.0	0.3

Notes: Number of replications=1000. Results are rounded to 5 digits. Only estimated elements of  $Pyd$  are displayed. Only linearly independent estimated elements of  $\hat{F}^{\text{RUM}}$  are displayed.

**Table 8** – Root Mean Squared Error in Estimating  $F^{\text{RUM}}$ . DGP1 ( $\times 10^{-2}$ )

Sample Size \ (i, j)	1, 2	1, 3	1, 4	1, 5
2000	3.3	2.3	3.0	2.3
5000	2.1	1.4	2.0	1.5
10000	1.5	1.0	1.4	1.0
50000	0.7	0.4	0.6	0.4

Notes: Number of replications=1000. Results are rounded to 3 digits. Only linearly independent estimated elements of  $\hat{F}^{\text{RUM}}$  are displayed.

household appears only once in the cross-section.<sup>5</sup> We consider a balanced panel by dropping any household that does not have 3 consecutive trips in a given year. We end up having  $S = 3$  consecutive choices of  $n = 47,509$  households.

There are only 2 product characteristics available in the Homescan and Nielsen Retail Scanner: price of a unit (USD) and size of it (ounces).<sup>6</sup> The dataset also contains information on zip-codes for every household/purchase in the sample. We use the Nielsen Retail Scanner and the Homescan to construct the dataset on prices and sizes by pooling all the information on prices per UPC code (barcode) of all RTE cereals by week and location (3 digit zip-code).<sup>7</sup>

<sup>5</sup>To ensure this we consider the first 3 trips per year per household. Then we create a unified panel with the information of years 2016 – 2018, and we balance the panel keeping only the first 3 trips. Hence, if any household appears in all three years, we keep only its 2016 observations.

<sup>6</sup>We also know barcodes of every purchase. Unfortunately, it is hard to match these barcodes with actual products to obtain additional product characteristics since these barcodes change over time and some products are not produced anymore.

<sup>7</sup>To obtain prices in the Homescan we use the paid price (including discounts) divided by the number of

**Table 9** – Bias in Estimating  $F^{\text{RUM}}$ . DGP2 ( $\times 10^{-2}$ )

Sample Size \ ( $i, j$ )	1, 2	1, 3	2, 3	1, 4	2, 4	3, 4	1, 5	1, 5	3, 5	4, 5
2000	-1.3	-3.7	-1.0	-0.2	0	-8.5	1.5	1.3	-0.4	-3.6
5000	-2.1	-3.5	-0.9	-0.6	-0.4	-7.4, 1.1	1	-0.2	-2.8	
10000	-1.4	-2.8	-0.6	-0.4	-0	-5.1	0.7	0.7	-0	-1.7
50000	-0.2	-0.4	0	0.2	0.2	-1.0	0.1	0.0	0.1	-0.3

Notes: Number of replications=1000. Results are rounded to 3 digits. Only estimated elements of  $P_{y,d}$  are displayed. Only linearly independent estimated elements of  $\hat{F}^{\text{RUM}}$  are displayed.

**Table 10** – Root Mean Squared Error in Estimating  $F^{\text{RUM}}$ . DGP2 ( $\times 10^{-2}$ )

Sample Size \ ( $i, j$ )	1, 2	1, 3	2, 3	1, 4	2, 4	3, 4	1, 5	1, 5	3, 5	4, 5
2000	16.4	10.4	12.8	15.8	13.7	10.5	4.0	3.5	3.3	4.6
5000	11.0	8.5	10.5	11.4	9.8	9.2	2.7	2.5	2.1	3.6
10000	5.6	6.3	6.3	7.0	6.4	7.0	2.0	1.9	1.5	2.5
50000	1.3	1.2	0.9	2.4	2.0	2.1	0.8	0.7	0.7	0.7

Notes: Number of replications=1000. Results are rounded to 3 digits. Only estimated elements of  $P_{y,d}$  are displayed. Only linearly independent estimated elements of  $\hat{F}^{\text{RUM}}$  are displayed.

Then we compute the mean price of every brand at every location.<sup>8</sup> Brand-location size variable is built similarly. As a result, given the information on the location of every household, we match every purchase with the price and size.

Despite having a relatively large sample, there are too many 3-digit level zip-codes to treat them as markets. Thus, to increase the number of observations per market, similar to [Nevo \(2001\)](#), we use prices and geographic coordinates (i.e. longitude and latitude) of every location to define markets. In particular, we define a market by employing K-means clustering with the Euclidean norm using centroids based on prices and geographic location. In other words, we group together households that live close to each other and face similar prices. We initialize the K-means and fix the number of markets using the 3-digit zip-code. All locations with less than 2000 households are collapsed to a single dummy location.<sup>9</sup> In total we obtain 34 units, and we drop from our sample those that pay a zero price after discount.

<sup>8</sup>We average across weeks to diminish measurement error in prices and because there are some missing prices per brand.

<sup>9</sup>This quantity was chosen on the basis of simulations, to ensure a sufficiently high number of observations per market.

markets.<sup>10</sup> The map depicting the geographical locations of the markets is depicted in Figure 1. Finally, we aggregate prices on the market-brand level.

Since the dataset contains information on the household’s income, the age of the household’s head, and the size of the household, we also compute the average (on the market level) income, the age of the head of the household, and the household size. We use these demographics in our analysis of own-price elasticity. Summary statistics for demographic variables are provided in Table 11

**Table 11** – Summary Statistics of Demographic Variables

Variable	Mean	Median	Std	Min	Max
Average Age (years)	54.33	54.27	1.55	49,87	57.12
Average Income (USD)	23,333	22,503	4,506.25	14,543.4	32,580
Average HH Size	2.7	2.7	0.12	2.49	3.18

Notes: These summary statistics are computed for 34 markets. For instance, the minimum market average age is the smallest among 34 markets market-average age, not the age of youngest head of the household in the sample.

Figure 1 depicts the geographical location of our markets and the fractions of the population in these markets that consider the two most frequent sets. We also add those who only consider GM alone for a better glimpse of the heterogeneity in choice sets.

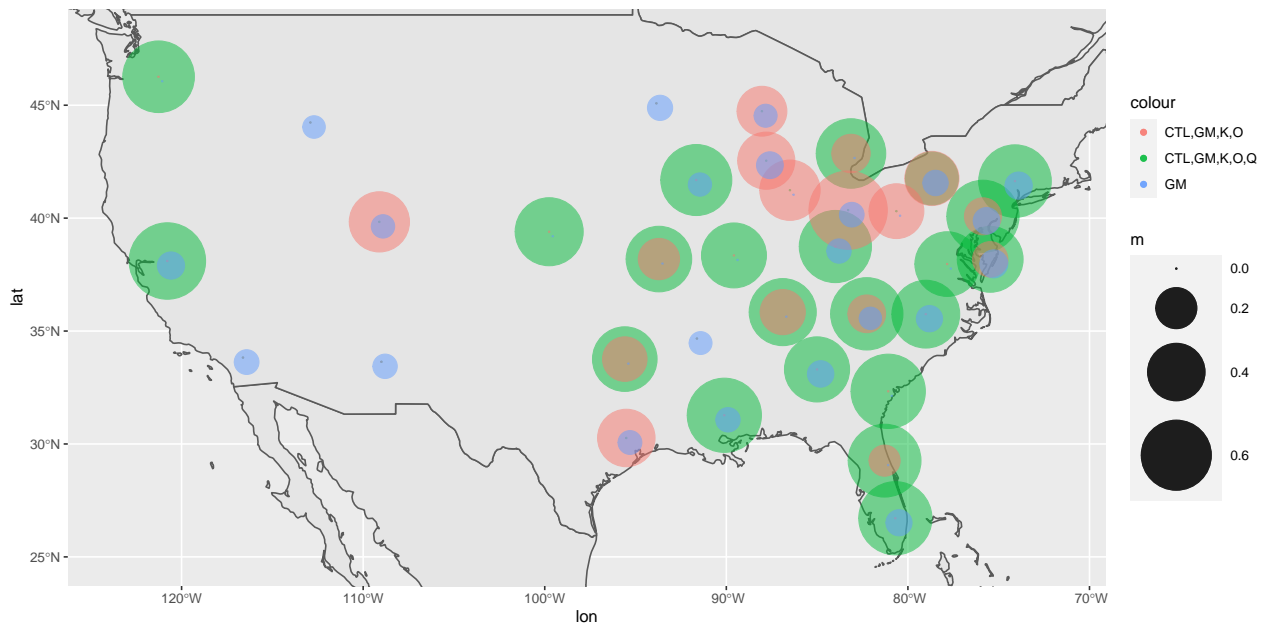
## C.2. Additional Results for the Illustrative Application

We estimate  $m$  and  $F^{\text{RUM}}$  conditioning on every market and covariate value. Since we aggregate the covariates on the market level, variation in covariates is only driven by variation across markets. Let  $\hat{m}(D|j, x)$  denote the estimated probability that set  $D$  is considered in market  $j$  given covariate value  $x$ .

Using the estimated  $\hat{m}$ , first, we find that *all* markets have less than 5 sets that are

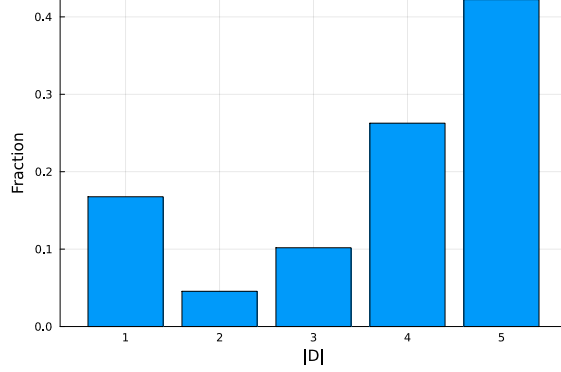
---

<sup>10</sup>We obtain qualitatively the same results when we increased the number of markets to 72.



**Figure 1** – LOCATION OF MARKETS AND CONSIDERATION PROBABILITIES OF SOME CHOICE SETS: CTL=Store brand, GM=General Mills, K=Kellogg, Q=Quaker , O=other brands. The size of every circle corresponds to the estimates of consideration probability  $\hat{m}$ . The color of every circle corresponds to a different choice set: Red={CTL,GM,K,O}, Green={CTL,GM,K,O,Q}, Blue={GM}.





**Figure 2** – PROPORTION OF INDIVIDUALS CONSIDERING SETS OF GIVEN CARDINALITY:  $|D|$  denotes the size of the choice set.

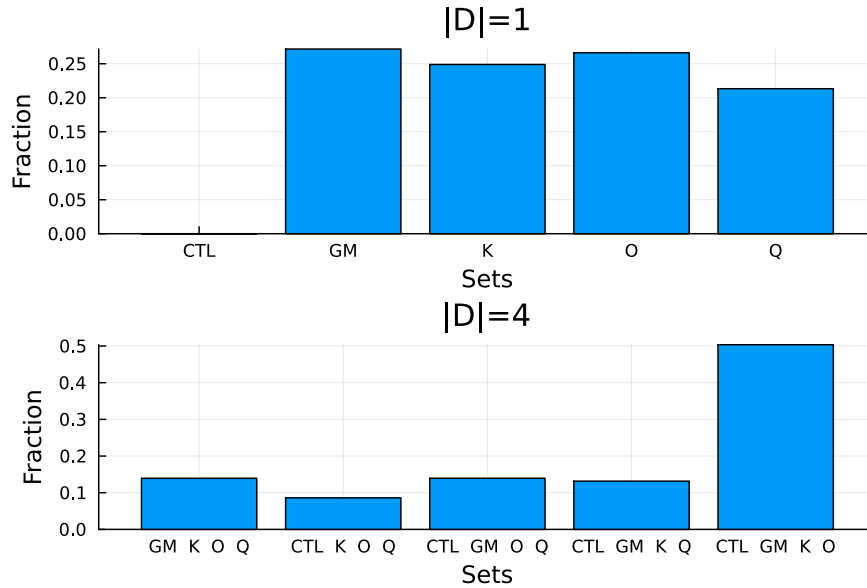
considered by more than 15 percent of households in the market. That is,

$$\sum_j \mathbf{1} \left( \sum_D \mathbf{1} (\hat{m}(D|j) > 0.1) \geq 5 \right) = 0.$$

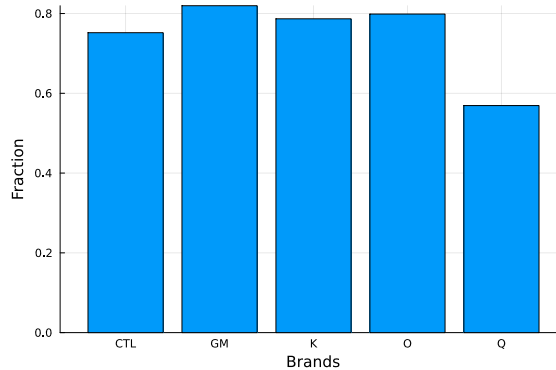
Even if we lower the threshold to 5 percent, more than 15 percent of markets have less than 5 choice sets. That is, among 5 estimated sets at least 1 set is considered by less than 15 percent of population in every market, and a sizable fraction of markets has at least one set that is faced by less than 5 percent of consumers. These findings lend support to our sparsity assumption.

Next, we compute the estimated proportion of individuals in the sample who considered sets of a given cardinality  $l$  as  $\sum_{j,D} \mathbf{1} (|D| = l) \hat{m}(D|j) w_j$ , where  $w_j = N_j/N$  is a fraction of the whole sample (of size  $N$ ) that is coming from market  $j$  (of size  $N_j$ ). As Figure 2 demonstrates, sets of all sizes are considered. The vast majority (about 70 percent) of the sample considered sets of cardinality 4 and 5. Given that most likely all 5 brands are usually present, these individuals can be thought of as full consideration individuals that we usually work with in discrete choice settings. However, about 16 percent of DMs only considered one brand. These are super loyal consumers that always purchase the same brand no matter what.

Next, we consider the composition of sets of cardinality 1 and 4 (there is only one set of size 5). The results are presented in Figure 3. Interestingly, CTL is never considered alone. The



**Figure 3** – DISTRIBUTION OF SETS OF GIVEN CARDINALITY:  $|D|$  denotes the size of the choice set. CTL=Store brand, GM=General Mills, K=Kellogg, Q=Quaker , O=other brands.



**Figure 4** – PROPORTION OF INDIVIDUALS CONSIDERING A BRAND: CTL=Store brand, GM=General Mills, K=Kellogg, Q=Quaker , O=other brands.

rest of the brands are almost equally likely considered by those who only look at one brand (Q has the smallest share of about 21 percent). Among those who considered sets of size 4, almost half considered everything but Q. The rest of sets of cardinality 4 have similar shares.

Next, we compute the fraction of DMs who paid attention to a set that contains a given brand  $b$  as  $\sum_{j,D} \mathbb{1}(b \in D) \hat{m}(D|j)w_j$ . Similar to Figure 3, Figure 4 indicates that Q is considered less often (about 57 percent of DMs) than other brands (about 80 percent of DMs).

Finally, there are just 2 sets that attract more than 5 percent of DMs: the set that contains all 5 brands (about 40 percent of DMs) and the set that contains all brands but Q (about 15

percent).

Overall, we can conclude that although most DMs seem to consider almost all brands, there is a sizeable fraction of those who only consider one brand. Moreover, Q is considered less often than other brands and CTL is always considered with other alternatives.

In market 1, those who consider all 5 brands prefer Kellogg over all other brands. At the same time, those who do not consider Q predominantly buy other brands of cereals (see Table 12).

**Table 12** – Market Shares for Two Choice Sets in Market 1

Brand/Set	{CTL,GM,L,Q,O}	{CTL,GM,L,O}
CTL	0.1	0.081
GM	0.334	0.161
K	0.346	0.11
O	0.132	0.648
Q	0.088	0.0

Notes: Results are rounded to 3 digits.

### Parametric Estimation of Price Elasticity

Given that  $\beta_D$  is estimated using GMM, if we treat the estimated shares as the true shares (i.e. no estimation error), then we can easily construct the 2-step efficient GMM standard errors. Table 13 displays the estimates of  $\beta_D$  together with their standard errors. Following

**Table 13** – Estimates of  $\beta$

	Direct	{CTL, GM, K, O, Q}	{CTL, GM, K, O}
$\hat{\beta}$	-17.28	-10.04	-13.65
std. error	3.69	10.51	14.93

Notes: Standard errors are computed assuming that there is no estimation error in shares. Results are rounded to 2 digits.

Nevo (2001), here we also report the median across markets own-price elasticities in Table 14.

**Table 14** – Estimates of Median Across Markets Own-Price Elasticities

	Direct	{CTL, GM, K, O, Q}	{CTL, GM, K, O}
CTL	-1.96	-1.23	-1.74
GM	-2.58	-1.53	-2.2
K	-2.19	-1.21	-1.97
O	-2.22	-1.34	-1.72
Q	-2.55	-1.48	0

Notes: The first column is computed assuming that consumers face all 5 brands. The second column is computed assuming choice set variation for those consumers who consider all 5 brands. The last column is computed for those consumers who do not consider Q. Results are rounded to 2 digits.

## References

- Hu, Yingyao (2008) “Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution,” *Journal of Econometrics*, 144 (1), 27–61.
- Hu, Yingyao, David McAdams, and Matthew Shum (2013) “Identification of first-price auctions with non-separable unobserved heterogeneity,” *Journal of Econometrics*, 174 (2), 186–193.
- Nevo, Aviv (2001) “Measuring market power in the ready-to-eat cereal industry,” *Econometrica*, 69 (2), 307–342.
- Ray, Debraj (R) Arthur Robson (2018) “Certified random: A new order for coauthorship,” *American Economic Review*, 108 (2), 489–520.