Appendix

A. Proofs for Section 6

A.1. Proof of Theorem 1

The result of Theorem 1 follows immediately from the following lemma, which we prove below.

LEMMA 1. *Let d denote the column rank of matrix A and $\mathcal{Y}$ denote the convex hull of the columns of A. Then, it must be that y belongs to a $d-1$ dimensional subspace, $\|\lambda^{\min}(y)\|_0 \le d+1$, and*

$$\mathrm{vol}_{d-1}(\mathcal{Y}^{\mathrm{sparse}}) = \mathrm{vol}_{d-1}(\mathcal{Y}),$$

where $\mathcal{Y}^{\mathrm{sparse}} \subset \mathcal{Y}$ denotes the set of all data vectors such that

$$\|\lambda^{\min}(y)\|_0 \le d+1 \text{ and } \|\lambda^{\mathrm{sparse}}(y)\|_0 \ge d$$

and $\mathrm{vol}_{d-1}(S)$ denotes the $d-1$ dimensional volume of a set S of points.

Proof of Lemma 1 We prove this lemma in two parts: (1) $\mathcal{Y}$ belongs to a $d-1$ dimensional subspace and $\|\lambda^{\min}(y)\|_0 \le d+1$ for all $y \in \mathcal{Y}$, and (2) $\mathrm{vol}_{d-1}(\mathcal{Y} \setminus \mathcal{Y}^{\mathrm{sparse}}) = 0$.

To prove the first part, note that any data vector $y \in \mathcal{Y}$ belongs to $d-1$ dimensional subspace because A has a d dimensional range space and y belongs to the intersection of the range space of A and the hyperplane $\sum_\sigma \lambda_\sigma = 1$. Let $\tilde{A}$ denote the augmented matrix, which is obtained by augmenting the last row of matrix A with a row of all 1s. Similarly, let $\tilde{y}$ denote the vector obtained by augmenting vector y with 1. The equality constraints of (2) can now be written as $\tilde{y} = \tilde{A}\lambda$, $\lambda \ge 0$. Since A has rank d, the rank of $\tilde{A}$ will be at most $d+1$. Therefore, for any data vector $y \in \mathcal{Y}$, an optimal BFS solution to (2) must be such that

$$\|\lambda^{\min}(y)\|_0 \le d+1, \quad \forall\, y \in \mathcal{Y}. \tag{12}$$

Coming to the second part of the proof, for any $r \le d-1$, let $\mathcal{Y}_r$ denote the set of all data vectors that can be written as a convex combination of at most r columns of matrix A. Let L denote the number of columns of A of size at most r, and let $S_1, S_2, \ldots, S_L$ denote the corresponding subsets of columns of A of size at most r. Then, it is easy to see that $\mathcal{Y}_r$ can be written as the union of disjoint subsets $\mathcal{Y}_r = \cup_{i=1}^{L} \mathcal{Y}_{ri}$, where for each i, $\mathcal{Y}_{ri}$ denotes the set of data vectors that can be written as the convex combination of the columns in subset S_i. For each i, since $\mathcal{Y}_{ri}$ is a polytope residing in $r-1 \le d-2$ dimensional space, it must follow that $\mathrm{vol}_{d-1}(\mathcal{Y}_{ri}) = 0$. Since L is finite, it follows that $\mathrm{vol}_{d-1}(\mathcal{Y}_r) = 0$. Therefore, we can conclude that

$$\mathrm{vol}_{d-1}\left(y \in \mathcal{Y} \colon \|\lambda^{\mathrm{sparse}}(y)\|_0 \le d-1\right) = 0 \tag{13}$$

The result of the lemma now follows from (12) and (13).

A.2. Proof of Theorem 2

Before we prove Theorem 2, we propose a simple combinatorial algorithm that recovers the model λ whenever λ satisfies the signature and linear independence conditions; we make use of this algorithm in the proof of the theorem.

The algorithm recovers λ when the signature and linear independence conditions are satisfied. If the conditions are not satisfied, the algorithm provides a certificate to that effect. The algorithm takes y as an explicit input with the prior knowledge of the structure of A as an auxiliary input. It's aim is to produce λ. In particular, the algorithm outputs the sparsity of λ, $K = \|\lambda\|_0$, permutations $\sigma_1, \ldots, \sigma_K$ so that $\lambda(\sigma_i) \ne 0$, $1 \le i \le K$ and the values $\lambda(\sigma_i)$, $1 \le i \le K$. Without loss of generality, assume that the values $y_1, \ldots, y_m$ are sorted with $y_1 \le \cdots \le y_m$ and further that $\lambda(\sigma_1) \le \lambda(\sigma_2) \le \ldots \lambda(\sigma_K)$.

Before we describe the algorithm, we observe the implication of the two conditions. The *Linear Independence* condition says that for any two non-empty distinct subsets $S, S' \subset \{1, \ldots, K\}$, $S \ne S'$,

$$\sum_{i \in S} \lambda(\sigma_i) \ne \sum_{j \in S'} \lambda(\sigma_j).$$

This means that if we know all $\lambda(\sigma_i)$, $1 \le i \le K$ and since we know $y_d, 1 \le d \le m$, then we can recover $A(\sigma_i)_d, i = 1, 2, \ldots, K$ as the unique solution to $y_d = \sum_{i=1}^{K} A(\sigma_i)_d \lambda(\sigma_i)$ in $\{0,1\}^K$.

Therefore, the non-triviality lies in finding K and $\lambda(\sigma_i)$, $1 \le i \le K$. This issue is resolved by use of the *Signature* condition in conjunction with the above described properties in an appropriate recursive manner.

Specifically, recall that the *Signature* condition implies that for each σ_i for which $\lambda(\sigma_i) \neq 0$, there exists d such that $y_d = \lambda(\sigma_i)$. By *Linear Independence*, it follows that all $\lambda(\sigma_i)$s are distinct and hence by our assumption

$$\lambda(\sigma_1) < \lambda(\sigma_2) < \cdots < \lambda(\sigma_K).$$

Therefore, it must be that the smallest value, y_1 equals $\lambda(\sigma_1)$. Moreover, $A(\sigma_1)_1 = 1$ and $A(\sigma_i)_1 = 0$ for all $i \neq 1$. Next, if $y_2 = y_1$ then it must be that $A(\sigma_1)_2 = 1$ and $A(\sigma_i)_2 = 0$ for all $i \neq 1$. We continue in this fashion until we reach a d' such that $y_{d'-1} = y_1$ but $y_{d'} > y_1$. Using similar reasoning it can be argued that $y_{d'} = \lambda(\sigma_2)$, $A(\sigma_2)_{d'} = 1$ and $A(\sigma_i)_{d'} = 0$ for all $i \neq 2$. Continuing in this fashion and repeating essentially the above argument with appropriate modifications leads to recovery of the sparsity K, the corresponding $\lambda(\sigma_i)$ and $A(\sigma_i)$ for $1 \leq i \leq K$. The complete procedural description of the algorithm is given below.

Sparsest Fit Algorithm:

Initialization: $k(1) = 1$, $d = 1$, $\lambda(\sigma_1) = y_1$ and $A(\sigma_1)_1 = 1$, $A(\sigma_1)_\ell = 0$, $2 \leq \ell \leq m$.
for $d = 2$ to m
  if $y_d = \sum_{i \in T} \lambda(\sigma_i)$ for some $T \subseteq \{1, \ldots, k(d-1)\}$
    $k(d) = k(d-1)$
    $A(\sigma_i)_d = 1 \quad \forall \quad i \in T$
  else
    $k(d) = k(d-1) + 1$
    $\lambda(\sigma_{k(d)}) = y_d$
    $A(\sigma_{k(d)})_d = 1$ and $A(\sigma_{k(d)})_\ell = 0$, for $1 \leq \ell \leq m, \ell \neq d$
  end if
end for
Output $K = k(m)$ and $(\lambda(\sigma_i), A(\sigma_i)), 1 \leq i \leq K$.

Now, we have the following theorem justifying the correctness of the above algorithm:

THEOREM 5. *Suppose we are given $y = A\lambda$ and λ satisfies the "Signature" and the "Linear Independence" conditions. Then, the Sparsest Fit algorithm recovers λ.*

We present the proof of Theorem 2 followed by the proof of Theorem 5.

Proof of Theorem 2 Suppose, to arrive at a contradiction, assume that there exists a distribution μ over the permutations such that $y = A\mu$ and $\|\mu\|_0 \leq \|\lambda\|_0$. Let $v_1, v_2, \ldots, v_K$ and $u_1, u_2, \ldots, u_L$ denote the values that λ and μ take on their respective supports. It follows from our assumption that $L \leq K$. In addition, since λ satisfies the "signature" condition, there exist $1 \leq d(i) \leq m$ such that $y_{d(i)} = v_i$, for all $1 \leq i \leq K$. Thus, since $y = A\mu$, for each $1 \leq i \leq K$, we can write $v_i = \sum_{j \in T(i)} u_j$, for some $T(i) \subseteq \{1, 2, \ldots, L\}$. Equivalently, we can write $v = Bu$, where B is a $0-1$ matrix of dimensions $K \times L$. Consequently, we can also write $\sum_{i=1}^{k} v_i = \sum_{j=1}^{L} \zeta_j u_j$, where ζ_j are integers. This now implies that $\sum_{j=1}^{L} u_j = \sum_{j=1}^{L} \zeta_j u_j$ since $\sum_{i=1}^{K} v_i = \sum_{j=1}^{L} u_j = 1$.

Now, there are two possibilities: either all the ζ_js are > 0 or some of them are equal to zero. In the first case, we prove that μ and λ are identical, and in the second case we arrive at a contradiction. In the case when $\zeta_j > 0$ for all $1 \leq j \leq L$, since $\sum_j u_j = \sum_j \zeta_j u_j$, it should follow that $\zeta_j = 1$ for all $1 \leq j \leq L$. Thus, since $L \leq K$, it should be that $L = K$ and $(u_1, u_2, \ldots, u_L)$ is some permutation of $(v_1, v_2, \ldots, v_K)$. By relabeling the u_js, if required, without loss of generality, we can say that $v_i = u_i$, for $1 \leq i \leq K$. We have now proved that the values of λ and μ are identical. In order to prove that they have identical supports, note that since $v_i = u_i$ and $y = A\lambda = A\mu$, μ must satisfy the "signature" and the "linear independence" conditions. Thus, the algorithm we proposed accurately recovers μ and λ from y. Since the input to the algorithm is only y, it follows that $\lambda = \mu$.

Now, suppose that $\zeta_j = 0$ for some j. Then, it follows that some of the columns in the B matrix are zeros. Removing those columns of B, we can write $v = \tilde{B}\tilde{u}$ where $\tilde{B}$ is B with the zero columns removed and $\tilde{u}$ is u with u_js such that $\zeta_j = 0$ removed. Let $\tilde{L}$ be the size of $\tilde{u}$. Since at least one column was removed $\tilde{L} < L \leq K$. The condition $\tilde{L} < K$ implies that the elements of vector v are not linearly independent i.e., we can find integers c_i such that $\sum_{i=1}^{K} c_i v_i = 0$. This is a contradiction, since this condition violates our "linear independence" assumption. The result of the theorem now follows.

Proof of Theorem 5 Let $\sigma_1, \sigma_2, \ldots, \sigma_K$ be the permutations in the support and $\lambda_1, \lambda_2, \ldots, \lambda_K$ be their corresponding probabilities. Since we assumed that λ satisfies the "signature" condition, for each $1 \le i \le K$, there exists a $d(i)$ such that $y_{d(i)} = \lambda_i$. In addition, the "linear independence" condition guarantees that the condition in the "if" statement of the algorithm is not satisfied whenever $d = d(i)$. To see why, suppose the condition in the "if" statement is true; then, we will have $\lambda_{d(i)} - \sum_{i \in T} \lambda_i = 0$. Since $d(i) \notin T$, this clearly violates the "linear independence" condition. Therefore, the algorithm correctly assigns values to each of the λ_is. We now prove that the $A(\sigma)$s that are returned by the algorithm do indeed correspond to the σ_is. For that, note that the condition in the "if" statement being true implies that y_d is a linear combination of a subset T of the set $\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$. Again, the "linear independence" condition guarantees that such a subset T, if exists, is unique. Thus, when the condition in the "if" statement is true, the only permutations with $A(\sigma)_d = 1$ are the ones in the set T. Similarly, when the condition in the "if" statement is false, then it follows from the "signature" and "linear independence" conditions that only for σ_i, $A(\sigma)_{d(i)} = 1$. From this, we conclude that the algorithm correctly finds the true underlying distribution.

A.3. Proof of Theorem 3

First, we note that, irrespective of the form of observed data, the choice model generated from the "generation model" satisfies the "linear independence" condition with probability 1. The reason is as follows: the values $\lambda(\sigma_i)$ obtained from the generation model are i.i.d uniformly distributed over the interval $[a, b]$. Therefore, the vector $(\lambda(\sigma_1), \lambda(\sigma_2), \ldots, \lambda(\sigma_K))$ corresponds to a point drawn uniformly at random from the hypercube $[a, b]^K$. In addition, the set of points that satisfy $\sum_{i=1}^{K} c_i \lambda(\sigma_i) = 0$ lie in a lower-dimensional space. Since c_is are bounded, there are only finitely many such sets of points. Thus, it follows that with probability 1, the choice model generated satisfies the "linear independence" condition.

The conditions under which the choice model satisfies the "signature" condition depends on the form of observed data. We consider each form separately.

1. Ranking Data: The bound of $K = O(n)$ directly follows from Lemma 2 of Jagabathula and Shah (2008).
2. Comparison Data: For each permutation σ, we truncate its corresponding column vector $A(\sigma)$ to a vector of length $N/2$ by restricting it to only the disjoint unordered pairs: $\{0,1\}, \{2,3\}, \ldots, \{N-2, N-1\}$. Denote the truncated binary vector by $A'(\sigma)$. Let $\tilde{A}$ denote the matrix A with each column $A(\sigma)$ truncated to $A'(\sigma)$. Clearly, since $\tilde{A}$ is just a truncated form of A, it is sufficient to prove that $\tilde{A}$ satisfies the "signature" condition.

   For brevity, let L denote $N/2$, and, given K permutations, let B denote the $L \times K$ matrix formed by restricting the matrix $\tilde{A}$ to the K permutations in the support. Then, it is easy to see that a set of K permutations satisfies the "signature" condition iff there exist K rows in B such that the $K \times K$ matrix formed by the K rows is a permutation matrix.

   Let $R_1, R_2, \ldots, R_J$ denote all the subsets of $\{1, 2, \ldots, m\}$ with cardinality K; clearly, $J = \binom{L}{K}$. In addition, let B^j denote the $K \times K$ matrix formed by the rows of B that are indexed by the elements of R_j. Now, for each $1 \le j \le J$, when we generate the matrix B by choosing K permutations uniformly at random, let $\mathcal{E}_j$ denote the event that the $K \times K$ matrix B^j is a permutation matrix and let $\mathcal{E}$ denote the event $\cup_j \mathcal{E}_j$. We want to prove that $\mathbb{P}(\mathcal{E}) \to 1$ as $N \to \infty$ as long as $K = o(\log N)$. Let X_j denote the indicator variable of the event $\mathcal{E}_j$, and X denote $\sum_j X_j$. Then, it is easy to see that $\mathbb{P}(X = 0) = \mathbb{P}((\mathcal{E})^c)$. Thus, we need to prove that $\mathbb{P}(X = 0) \to 0$ as $N \to \infty$ whenever $K = o(\log n)$. Now, note the following:

$$\mathrm{Var}(X) \ge (0 - \mathbb{E}[X])^2 \, \mathbb{P}(X = 0)$$

   It thus follows that $\mathbb{P}(X = 0) \le \mathrm{Var}(X)/(\mathbb{E}[X])^2$. We now evaluate $\mathbb{E}[X]$. Since X_js are indicator variables, $\mathbb{E}[X_j] = \mathbb{P}(X_j = 1) = \mathbb{P}(\mathcal{E}_j)$. In order to evaluate $\mathbb{P}(\mathcal{E}_j)$, we restrict our attention to the $K \times K$ matrix B^j. When we generate the entries of matrix B by choosing K permutations uniformly at random, all the elements of B will be i.i.d Be(1/2) i.e., uniform Bernoulli random variables. Therefore, there are 2^{K^2} possible configurations of B^j and each of them occurs with a probability $1/2^{K^2}$. Moreover, there are $K!$ possible $K \times K$ permutation matrices. Thus, $\mathbb{P}(\mathcal{E}_j) = K!/2^{K^2}$. Thus, we have:

$$\mathbb{E}[X] = \sum_{j=1}^{J} \mathbb{E}[X_j] = \sum_{j=1}^{J} \mathbb{P}(\mathcal{E}_j) = \frac{JK!}{2^{K^2}}. \tag{14}$$

Since $J = \binom{L}{K}$, it follows from Stirling's approximation that $J \geq L^K/(eK)^K$. Similarly, we can write $K! \geq K^K/e^K$. It now follows from (14) that

$$\mathbb{E}[X] \geq \frac{L^K}{e^K K^K} \frac{K^K}{e^K} \frac{1}{2^{K^2}} = \frac{L^K}{e^{2K} 2^{K^2}}. \tag{15}$$

We now evaluate $\mathrm{Var}(X)$. Let ρ denote $K!/2^{K^2}$. Then, $\mathbb{E}[X_j] = \rho$ for all $1 \leq j \leq J$. We can write,

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \sum_{i=1}^{J} \sum_{j=1}^{J} \mathbb{P}(X_i = 1, X_j = 1) - J^2\rho^2.$$

Suppose $|R_i \cap R_j| = r$. Then, the number of possible configurations of B^i and B^j is $2^{(2K-r)K}$ because, since there is an overlap of r rows, there are $2K - r$ distinct rows and, of course, K columns. Since all configurations occur with the same probability, it follows that each configuration occurs with a probability $1/2^{(2K-r)K}$, which can also be written as $2^{rK}\rho^2/(K!)^2$. Moreover, the number of configurations in which both B^i and B^j are permutation matrices is equal to $K!(K-r)!$, since, fixing the configuration of B^i will leave only $K - r$ rows of B^j to be fixed.

For a fixed R_i, we now count the number of subsets R_j such that $|R_i \cap R_j| = r$. We construct an R_j by first choosing r rows from R_i and then choosing the rest from $\{1, 2, \ldots, l\} \setminus R_i$. We can choose r rows from the subset R_i of K rows in $\binom{K}{r}$ ways, and the remaining $K - r$ rows in $\binom{L-K}{K-r}$ ways. Therefore, we can now write:

$$\begin{aligned}
\sum_{j=1}^{J} \mathbb{P}(X_i = 1, X_j = 1) &= \sum_{r=0}^{K} \binom{K}{r}\binom{L-K}{K-r} K!(K-r)! \frac{2^{rK}\rho^2}{(K!)^2} \\
&\leq \rho^2 \sum_{r=0}^{K} \binom{L}{K-r} \frac{2^{rK}}{r!}, \quad \text{Using } \binom{L-K}{K-r} \leq \binom{L}{K-r} \\
&= \binom{L}{K}\rho^2 + \rho^2 \sum_{r=1}^{K} \binom{L}{K-r} \frac{2^{rK}}{r!} \\
&\leq J\rho^2 + \rho^2 L^K \sum_{r=1}^{K} \left(\frac{e2^K}{L}\right)^r \frac{1}{r^r (K-r)^{K-r}}
\end{aligned}$$

The last inequality follows from Stirling's approximation: $\binom{L}{K-r} \leq (L/(K-r))^{K-r}$ and $r! \geq (r/e)^r$; in addition, we have used $J = \binom{L}{K}$. Now consider

$$\begin{aligned}
r^r (K-r)^{K-r} &= \exp\{r\log r + (K-r)\log(K-r)\} \\
&= \exp\{K\log K - KH(r/K)\} \\
&\geq \frac{K^K}{2^K}
\end{aligned}$$

where $H(x)$ is the Shannon entropy of the random variable distributed as $\mathrm{Be}(x)$, defined as $H(x) = -x\log x - (1-x)\log(1-x)$ for $0 < x < 1$. The last inequality follows from the fact that $H(x) \leq \log 2$ for all $0 < x < 1$. Putting everything together, we get

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{J} \left[\sum_{j=1}^{J} \mathbb{P}(X_i = 1, X_j = 1)\right] - \mathbb{E}[X]^2 \\
&\leq J\left[J\rho^2 + \rho^2 L^K \frac{2^K}{K^K} \sum_{r=1}^{K} \left(\frac{e2^K}{L}\right)^r\right] - J^2\rho^2 \\
&= \frac{J\rho^2 2^K L^K}{K^K} \sum_{r=1}^{K} \left(\frac{e2^K}{L}\right)^r
\end{aligned}$$

We can now write,

$$\begin{aligned}
\mathbb{P}(X=0) &\leq \frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2} \\
&\leq \frac{1}{J^2\rho^2}\frac{J\rho^2 2^K L^K}{K^K}\sum_{r=1}^{K}\left(\frac{e2^K}{L}\right)^r \\
&= \frac{1}{J}\,\frac{2^K L^K}{K^K}\,\frac{e2^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r \\
&\leq \frac{e^K K^K}{L^K}\,\frac{2^K L^K}{K^K}\,\frac{e2^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r, \quad \text{Using } J=\binom{L}{K}\leq\left(\frac{L}{eK}\right)^K \\
&= e\frac{(4e)^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r
\end{aligned}$$

It now follows that for $K = o(\log L/\log(4e))$, $\mathbb{P}(X=0)\to 0$ as $N\to\infty$. Since, by definition, $L = N/2$, this completes the proof of the theorem.

3. Top Set Data: For this type of data, note that it is sufficient to prove that $A^{(1)}$ satisfies the "signature" property with a high probability; therefore, we ignore the comparison data and focus only on the data corresponding to the fraction of customers that have product i as their top choice, for every product i. For brevity, we abuse the notation and denote $A^{(1)}$ by A and $y^{(1)}$ by y. Clearly, y is of length N and so is each column vector $A(\sigma)$. Every permutation σ ranks only one product in the first position. Hence, for every permutation σ, exactly one element of the column vector $A(\sigma)$ is 1 and the rest are zeros.

   In order to obtain a bound on the support size, we reduce this problem to a balls-and-bins setup. For that, imagine K balls being thrown uniformly at random into N bins. In our setup, the K balls correspond to the K permutations in the support and the N bins correspond to the N products. A ball is thrown into bin i provided the permutation corresponding to the ball ranks product i to position 1. Our "generation model" chooses permutations independently; hence, the balls are thrown independently. In addition, a permutation chosen uniformly at random ranks a given product i to position 1 with probability $1/N$. Therefore, each ball is thrown uniformly at random.

   In the balls-and-bins setup, the "signature" condition translates into all K balls falling into different bins. By "Birthday Paradox" McKinney (1966), the K balls falls into different bins with a high probability provided $K = o(\sqrt{N})$.

This finishes the proof of the theorem.

A.4. Proof of Theorem 4

To show existence of a choice model $\hat{\lambda}$ with sparse support, that approximates expected revenue of all offer sets of size at most C with respect to the true model, we shall utilize the probabilistic method. Specifically, consider M samples chosen as per the true choice model λ : let these be $\sigma_1,\ldots,\sigma_M$. Let $\hat{\lambda}$ be the empirical choice model (or distribution on permutations) induced by these M samples. We shall show that for M large enough (as claimed in the statement of Theorem 4), this empirical distribution $\hat{\lambda}$ satisfies the desired properties with positive probability. That is, there exists a distribution with sparse support that satisfies the desired property and hence implies Theorem 4.

To this end, consider an offer set $\mathcal{M}$ of size at most C. As noted earlier, the expected revenue $R(\mathcal{M})$ is given by

$$R(\mathcal{M}) = \sum_{j\in\mathcal{M}} p_j\lambda_j(\mathcal{M}),$$

where p_j is the price of product $j\in\mathcal{M}$ and $\lambda_j(\mathcal{M})$ is the probability of customer choosing j to purchase, i.e. $\lambda(\mathcal{S}_j(\mathcal{M}))$. We wish to show that $\hat{\lambda}_j(\mathcal{M}) = \hat{\lambda}(\mathcal{S}_j(\mathcal{M}))$ is good approximation of $\lambda_j(\mathcal{M})$, for all $j\in\mathcal{M}$ and for all $\mathcal{M}$ of size at most C. To show this, we shall use a combination of Chernoff/Hoeffding bound and union bound.

To this end, consider the given $\mathcal{M}$ and a fixed $j\in\mathcal{M}$. For $1\leq\ell\leq M$, define

$$X_\ell^j = \begin{cases} 1 & \text{if } \sigma_\ell\in\mathcal{S}_j(\mathcal{M}), \\ 0 & \text{otherwise.}\end{cases}$$

Then, $X_\ell^j,\ 1 \le \ell \le M$, are independent and identically distributed Bernoulli random variables with $\mathbb{P}(X_\ell^j = 1) = \lambda^j(\mathcal{M})$. By definition,

$$\hat{\lambda}^j(\mathcal{M}) = \frac{1}{M}\sum_{\ell=1}^{M} X_\ell^j. \tag{16}$$

Using (16) and Chernoff/Hoeffding bound for $\sum_{\ell=1}^{M} X_\ell^j$, it follows that for any $t > 0$,

$$\mathbb{P}\left(\left|\hat{\lambda}^j(\mathcal{M}) - \lambda^j(\mathcal{M})\right| > t\right) \le 2\exp\left(-\frac{t^2 M}{2}\right). \tag{17}$$

Let $p_{\max} = \max_{i=1}^{N} p_i$. By selecting, $t = \frac{\varepsilon}{C p_{\max}}$ in (17) we have

$$\mathbb{P}\left(\left|\hat{\lambda}^j(\mathcal{M}) - \lambda^j(\mathcal{M})\right| > \frac{\varepsilon}{C p_{\max}}\right) \le 2\exp\left(-\frac{\varepsilon^2 M}{2C^2 p_{\max}^2}\right). \tag{18}$$

Therefore, for the given $\mathcal{M}$ of size at most C, by union bound we have

$$\mathbb{P}\left(\left|\sum_{j\in\mathcal{M}} p_j \hat{\lambda}^j(\mathcal{M}) - \sum_{j\in\mathcal{M}} p_j \lambda^j(\mathcal{M})\right| > \varepsilon\right) \le 2C\exp\left(-\frac{\varepsilon^2 M}{2C^2 p_{\max}^2}\right). \tag{19}$$

There are at most N^C sets of size upto C. Therefore, by union bound and (19) it follows that

$$\mathbb{P}\left(\max_{\mathcal{M}:|\mathcal{M}|\le C}\left|R(\mathcal{M}) - \sum_{j\in\mathcal{M}} p_j \hat{\lambda}_j(\mathcal{M})\right| > \varepsilon\right) \le 2CN^C\exp\left(-\frac{\varepsilon^2 M}{2C^2 p_{\max}^2}\right). \tag{20}$$

For choice of M such that

$$M > \frac{2C^2 p_{\max}^2}{\varepsilon^2}\left(\log 2C + C\log N\right),$$

the right hand size of (20) becomes < 1. This establishes the desired result.

B. The Exact Approach to Solving the Robust Problem

Here we provide further details on the second approach described for the solution to the (dual of) the robust problem (4). In particular, we first consider the case of ranking data, where an efficient representation of the constraints in the dual may be produced. We then illustrate a method that produces a sequence of 'outer-approximations' to (6) for general types of data, and thereby allows us to produce a sequence of improving lower bounding approximations to our robust revenue estimation problem, (2). This provides a general procedure to address the task of solving (4), or equivalently, (2).

B.1. A Canonical Representation for Ranking Data

Recall the definition of *ranking data* from Section 2: This data yields the fraction of customers that rank a given product i as their rth choice. Thus, the partial information vector y is indexed by i, r with $0 \le i, r \le N$. For each i, r, y_{ri} denotes the probability that product i is ranked at position r. The matrix A is thus in $\{0,1\}^{N^2 \times N!}$ and for a column of A corresponding to the permutation σ, $A(\sigma)$, we will thus have $A(\sigma)_{ri} = 1$ iff $\sigma(i) = r$. We will now construct an efficient representation of the type (6) for this type of data.

Consider partitioning $\mathcal{S}_j(\mathcal{M})$ into $D_j = N$ sets wherein the dth set is given by

$$\mathcal{S}_{jd}(\mathcal{M}) = \{\sigma \in \mathcal{S}_j(\mathcal{M}) \colon \sigma(j) = d\}.$$

and define, as usual, $\mathcal{A}_{jd}(\mathcal{M}) = \{A(\sigma) : \sigma \in \mathcal{S}_{jd}(\mathcal{M})\}$. Thus, $\mathcal{A}_{jd}(\mathcal{M})$ is the set of columns of A whose corresponding permutations rank the jth product as the dth most preferred choice.

It is easily seen that the set $\mathcal{A}_{jd}(\mathcal{M})$ is equal to the set of all vectors x^{jd} in $\{0,1\}^{N^2}$ satisfying:

$$\begin{aligned}
&\sum_{i=0}^{N-1} x_{ri}^{jd} = 1 \quad \text{for } 0 \le r \le N-1\\
&\sum_{r=0}^{N-1} x_{ri}^{jd} = 1 \quad \text{for } 0 \le i \le N-1\\
&x_{ri}^{jd} \in \{0,1\} \quad \text{for } 0 \le i, r \le N-1.\\
&x_{dj}^{jd} = 1\\
&x_{d'i}^{jd} = 0 \qquad \text{for all } i \in \mathcal{M}, i \ne j \text{and } 0 \le d' < d.
\end{aligned} \tag{21}$$

The first three constraints in (21) enforce the fact that x^{jd} represents a valid permutation. The penultimate constraint requires that the permutation encoded by x^{jd}, say σ^{jd}, satisfies $\sigma^{jd}(j) = d$. The last constraint simply ensures that $\sigma^{jd} \in \mathcal{S}_j(\mathcal{M})$.

Our goal is, of course, to find a description for $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ of the type (6). Now consider replacing the third (integrality) constraint in (21)

$$x^{jd}_{ri} \in \{0,1\} \quad \text{for } 0 \le i, r \le N-1$$

with simply the non-negativity constraint

$$x^{jd}_{ri} \ge 0 \quad \text{for } 0 \le i, r \le N-1$$

We claim that the resulting polytope is precisely the convex hull of $\mathcal{A}_{jd}(\mathcal{M}), \bar{\mathcal{A}}_{jd}(\mathcal{M})$. To see this, we note that all feasible points for the resulting polytope satisfy the first, second, fourth and fifth constraint of (21). Further, the polytope is integral, being the projection of a matching polytope with some variables forced to be integers (Birkhoff (1946), von Neumann (1953)), so that any feasible solution must also satisfy the third constraint of (21). We consequently have an *efficient* canonical representation of the type (6), which via (8) yields, in turn, an efficient solution to our robust revenue estimation problem (2) for ranking data, which we now describe for completeness.

Let us define for convenience the set $\mathcal{V}(\mathcal{M}) = \{(j,d) : j \in \mathcal{M}, 0 \le d \le N-1\}$, and for each pair (j,d), the sets $\mathcal{B}(j,d,\mathcal{M}) = \{(i,d') : i \in \mathcal{M}, i \ne j, 0 \le d' < d\}$. Then, specializing (8) to the canonical representation just proposed, we have that the following simple program in the variables α, ν and $\gamma^{jd} \in \mathbb{R}^{2N}$ is, in fact, equivalent to (2) for ranking data:

$$\begin{array}{ll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu \\ \text{subject to} & \gamma^{jd}_i + \gamma^{jd}_{N+r} \ge \alpha_{ri} \qquad \text{for all } (j,d) \in \mathcal{V}(\mathcal{M}), (i,r) \notin \mathcal{B}(j,d,\mathcal{M}) \\ & \sum_{i \ne j} \gamma^{jd}_i + \sum_{r \ne d} \gamma^{jd}_{N+r} + \nu \le p_j - \alpha_{dj} \quad \text{for all } (j,d) \in \mathcal{V}(\mathcal{M}) \end{array} \tag{22}$$

B.2. Computing a Canonical Representation: The General Case

While it is typically quite easy to 'write down' a description of the sets $\mathcal{A}_{jd}(\mathcal{M})$ as all integer solutions to some set of linear inequalities (as we did for the case of ranking data), relaxing this integrality requirement will typically *not* yield the convex hull of $\mathcal{A}_{jd}(\mathcal{M})$. In this section we describe a procedure that starting with the former (easy to obtain) description, solves a sequence of linear programs that yield improving solutions. More formally, we assume a description of the sets $\mathcal{A}_{jd}(\mathcal{M})$ of the type

$$\mathcal{I}_{jd}(\mathcal{M}) = \{x^{jd} : A^{jd}_1 x^{jd} \ge b^{jd}_1, \quad A^{jd}_2 x^{jd} = b^{jd}_2, \quad A^{jd}_3 x^{jd} \le b^{jd}_3, \quad x^{jd} \in \{0,1\}^m\} \tag{23}$$

This is similar to (6), with the important exception that we now allow integrality constraints. Given a set $\mathcal{I}_{jd}(\mathcal{M})$ we let $\bar{\mathcal{I}}^0_{jd}(\mathcal{M})$ denote the polytope obtained by relaxing the requirement $x^{jd} \in \{0,1\}^m$ to simply $x^{jd} \ge 0$. In the case of ranking data, $\bar{\mathcal{I}}^o_{jd}(\mathcal{M}) = \text{conv}(\mathcal{I}_{jd}(\mathcal{M})) = \bar{\mathcal{A}}_{jd}(\mathcal{M})$ and we were done; we begin with an example where this is not the case.

Example 1. Recall the definition of *comparison data* from Section 2. In particular, this data yields the fraction of customers that prefer a given product i to a product j. The partial information vector y is thus indexed by i,j with $0 \le i,j \le N; i \ne j$ and for each i,j, $y_{i,j}$ denotes the probability that product i is preferred to product j. The matrix A is thus in $\{0,1\}^{N(N-1)\times N!}$. A column of A, $A(\sigma)$, will thus have $A(\sigma)_{ij} = 1$ if and only if $\sigma(i) < \sigma(j)$.

Consider $\mathcal{S}_j(\mathcal{M})$, the set of all permutations that would result in a purchase of j assuming $\mathcal{M}$ is the set of offered products. It is not difficult to see that the corresponding set of columns $\mathcal{A}_j(\mathcal{M})$ is equal to the set of vectors in $\{0,1\}^{(N-1)N}$ satisfying the following constraints:

$$\begin{array}{ll} x^j_{il} \ge x^j_{ik} + x^j_{kl} - 1 & \text{for all } i,k,l \in \mathcal{N}, i \ne k \ne l \\ x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i \ne k \\ x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \ne j \\ x^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \ne k \end{array} \tag{24}$$

Briefly, the second constraint follows since for any $i,k, i \ne k$, either $\sigma(i) > \sigma(k)$ or else $\sigma(i) < \sigma(k)$. The first constraint enforces transitivity: $\sigma(i) < \sigma(k)$ and $\sigma(k) < \sigma(l)$ together imply $\sigma(i) < \sigma(l)$. The third constraint enforces that all $\sigma \in S_j(\mathcal{M})$ must satisfy $\sigma(j) < \sigma(i)$ for all $i \in \mathcal{M}$. Thus, (24) is a description of the type (23) with $D_j = 1$ for all j. Now consider the polytope $\bar{\mathcal{I}}^o_j(\mathcal{M})$ obtained by relaxing the fourth (integrality)

constraint to simply $x^j_{ik} \geq 0$. Of course, we must have $\bar{\mathcal{I}}^o_j(\mathcal{M}) \supseteq \text{conv}(\mathcal{I}_j(\mathcal{M})) = \text{conv}(\mathcal{A}_j(\mathcal{M}))$. Unlike the case of ranking data, however, $\bar{\mathcal{I}}^o_j(\mathcal{M})$ can in fact be shown to be *non*-integral [9], so that $\bar{\mathcal{I}}^o_j(\mathcal{M}) \neq \text{conv}(\mathcal{A}_j(\mathcal{M}))$ in general.

We next present a procedure that starting with a description of the form in (23), solves a sequence of linear programs each of which yield improving solutions to (2) along with bounds on the quality of the approximation:

1. Solve (8) using $\bar{\mathcal{I}}^o_{jd}(\mathcal{M})$ in place of $\text{conv}(\mathcal{I}_{jd}(\mathcal{M})) = \bar{\mathcal{A}}_{jd}(\mathcal{M})$. This yields a lower bound on (2) since $\bar{\mathcal{I}}^o_{jd}(\mathcal{M}) \supset \bar{\mathcal{A}}_{jd}(\mathcal{M})$. Call the corresponding solution $\alpha_{(1)}, \nu_{(1)}$.
2. Solve the optimization problem $\max \alpha^\top_{(1)} x^{jd}$ subject to $x^{jd} \in \bar{\mathcal{I}}^o_{jd}(\mathcal{M})$ for each pair (j,d). If the optimal solution $\hat{x}^{jd}$ is integral for each (j,d), then stop; the solution computed in the first step is in fact optimal.
3. Otherwise, let $\hat{x}^{jd}$ possess a non-integral component for some (j,d); say $\hat{x}^{jd}_c \in (0,1)$. Partition $\mathcal{A}_{jd}(\mathcal{M})$ on this variable - i.e. define

$$\mathcal{A}_{jd_0}(\mathcal{M}) = \{A(\sigma) : A(\sigma) \in \mathcal{A}_{jd}(\mathcal{M}), A(\sigma)_c = 0\}$$

and

$$\mathcal{A}_{jd_1}(\mathcal{M}) = \{A(\sigma) : A(\sigma) \in \mathcal{A}_{jd}(\mathcal{M}), A(\sigma)_c = 1\},$$

and let $\mathcal{I}_{jd_0}(\mathcal{M})$ and $\mathcal{I}_{jd_1}(\mathcal{M})$ represent the corresponding sets of linear inequalities with integer constraints (i.e. the projections of $\mathcal{I}_{jd}(\mathcal{M})$ obtained by restricting x^{jd}_c to be 0 and 1 respectively). Of, course, these sets remain of the form in (23). Replace $\mathcal{I}_{jd}(\mathcal{M})$ with $\mathcal{I}_{jd_0}(\mathcal{M})$ and $\mathcal{I}_{jd_1}(\mathcal{M})$ and go to step 1.

The above procedure is akin to a cutting plane method and is clearly finite, but the size of the LP we solve increases (by up to a factor of 2) at each iteration. Nonetheless, each iteration produces a lower bound to (2) whose quality is easily measured (for instance, by solving the maximization version of (2) using the same procedure, or by sampling constraints in the program (4) and solving the resulting program in order to produce an upper bound on (2)). Moreover, the quality of our solution improves with each iteration. In our computational experiments with a related type of data, it sufficed to stop after a single iteration of the above procedure.

B.3. Explicit LP solved for censored comparison data in Section 4

The LP we want to solve is

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \tag{25}$$

For the 'censored' comparison data, the partial information vector is indexed by i,j with $0 \leq i,j \leq N-1$, $i \neq j$. For each i,j such that $i \neq 0$, y_{ij} denotes the fraction of customers that prefer product i to both products j and 0; in other words, y_{ij} denotes the fraction of customers that purchase product i when then offer set is $\{i,j,0\}$. Further, for each $j \neq 0$, y_{0j} denotes the fraction of customers who prefer the 'no-purchase' option to product j; in fact, y_{0j} is the fraction of customers who don't purchase anything when the set $\{j,0\}$ is on offer. The matrix A is then in $\{0,1\}^{N(N-1)}$, with the column of A corresponding to permutation σ, $A(\sigma)$, having $A(\sigma)_{ij} = 1$ if $\sigma(i) < \sigma(j)$ and $\sigma(i) < \sigma(0)$ for each $i \neq 0, j$, and $A(\sigma)_{0j} = 1$ if $\sigma(0) < \sigma(j)$ for $j \neq 0$, and $A(\sigma)_{ij} = 0$ otherwise.

For reasons that will become apparent soon, we modify the LP in (25) by replacing the constraint $A\lambda = y$ with $A\lambda \geq y$. It is now easy to see the following:

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda \geq y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \quad \leq \quad \begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \tag{26}$$

[9] for $N \geq 5$; the polytope can be shown to be integral for $N \leq 4$

We now take the dual of the modified LP. In order to do that, recall from section 3 that $\mathcal{S}_j(\mathcal{M}) = \{\sigma \in S_N : \sigma(j) < \sigma(i), \forall i \in \mathcal{M}, i \neq j\}$ denotes the set of all permutations that result in the purchase of the product $j \in \mathcal{M}$ when the offered assortment is $\mathcal{M}$. In addition, $\mathcal{A}_j(\mathcal{M})$ denotes the set $\{A(\sigma) : \sigma \in \mathcal{S}_j(\mathcal{M})\}$. Now, the dual of the modified LP is

$$\begin{array}{ll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu \\ \text{subject to} & \max\limits_{z^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\ & \alpha \geq 0. \end{array} \tag{27}$$

where α and ν are dual variables corresponding respectively to the data consistency constraints $A\lambda = y$ and the requirement that λ is a probability distribution (i.e. $\mathbf{1}^\top \lambda = 1$) respectively.

Now, consider the following representation of the set $\mathcal{A}_j(\mathcal{M})$, for a fixed j.

$$\begin{array}{ll} z^j_{ik} = \min\left\{x^j_{ik}, x^j_{i0}\right\} & \text{for all } i,k \in \mathcal{N}, i \neq k, i \neq 0 \\ z^j_{0k} = x^j_{0k} & \text{for all } k \in \mathcal{N}, k \neq 0 \\ z^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \neq k \\ x^j_{il} \geq x^j_{ik} + x^j_{kl} - 1 & \text{for all } i,k,l \in \mathcal{N}, i \neq k \neq l \\ x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i \neq k \\ x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \neq j \\ x^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \neq k \end{array} \tag{28}$$

The last four constraints are the same as the set of inequalities in (24), which correspond to the representation of the set $\mathcal{A}_j(\mathcal{M})$ for comparison data; thus, every point satisfying the set of last four constraints in (28) corresponds to a permutation $\sigma \in \mathcal{S}_j(\mathcal{M})$ such that $x^j_{ik} = 1$ if and only if $\sigma(i) < \sigma(k)$. We now claim that the set of points z^j that satisfy the constraints in (28) is equal to the set of vectors in $\mathcal{A}_j(\mathcal{M})$. To see that, note that $z^j_{ik} = 1$ if and only if the corresponding $x^j_{ik} = 1$ *and* $x^j_{i0} = 1$, for $i \neq 0$. This implies that $z^j_{ik} = 1$ if and only if i is preferred to k *and* i is preferred to 0. Similarly, $z^j_{0k} = 1$ if and only if $x^j_{0k} = 1$ i.e., 0 is preferred to k.

Let $\bar{\mathcal{I}}_j(\mathcal{M})$ denote the convex hull of the vectors in $\mathcal{A}_j(\mathcal{M})$, equivalently, of the vectors z^j satisfying the set of constraints in (28). Let $\bar{\mathcal{I}}^o_j(\mathcal{M})$ be the convex hull of the vectors z^j satisfying the constraints in (28) with the constraint $z^j_{ik} = \min\left\{x^j_{ik}, x^j_{i0}\right\}$ replaced by the constraints $z^j_{ik} \leq x^j_{ik}$ and $z^j_{ik} \leq x^j_{i0}$, and the constraint $z^j_{0k} = x^j_{0k}$ replaced by the constraint $z^j_{0k} \leq x^j_{0k}$. Finally, let $\bar{\mathcal{I}}^1_j(\mathcal{M})$ represent the polytope $\bar{\mathcal{I}}^0_j(\mathcal{M})$ with the integrality constraints relaxed to $z^j_{ik} \geq 0$ and $x^j_{ik} \geq 0$. We now have the following relationships:

$$\begin{aligned} \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} &= \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}^o_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} \\ &\supseteq \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}^1_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} \end{aligned} \tag{29}$$

The first equality follows because $\alpha \geq 0$ and, hence, at the optimal solution, $z^j_{ik} = 1$ if $x^j_{ik} = x^j_{i0} = 1$, and $z^j_{0k} = 1$ if $x^j_{0k} = 1$. It should be now clear that in order to establish this equality we considered the modified LP. The second relationship follows because of the relaxation of constraints. It now follows from (26), (27) and (29) that

$$\begin{array}{lll} \underset{\lambda}{\text{minimize}} & \sum\limits_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) & \underset{\alpha,\nu}{\text{maximize}}\ \alpha^\top y + \nu \\ \text{subject to} & A\lambda = y, & \geq \text{subject to} \max\limits_{z^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\ & \mathbf{1}^\top \lambda = 1, & \qquad\qquad \alpha \geq 0. \\ & \lambda \geq 0. & \underset{\alpha,\nu}{\text{maximize}}\ \alpha^\top y + \nu \\ & & \geq \text{subject to} \max\limits_{z^j \in \bar{\mathcal{I}}^1_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\ & & \qquad\qquad \alpha \geq 0. \end{array} \tag{30}$$

Using the procedure described in Section 3.3, we solve the last LP in (30) by taking the dual of the constraint in the LP. For convenience, we write out the program $\max_{z^j \in \mathcal{I}^1_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right)$ and the corresponding

dual variables we use for each of the constraints.

$$\begin{array}{lll}
\underset{z^j}{\text{maximize}} & \alpha^\top z^j + \nu & \\
\text{subject to} & & \text{Dual Variables} \\
z^j_{ik} - x^j_{ik} \le 0 & \text{for all } i,k \in \mathcal{N}, i \neq k & \Omega 1^j_{ik} \\
z^j_{ik} - x^j_{i0} \le 0 & \text{for all } i,k \in \mathcal{N}, i \neq k, i \neq 0 & \Omega 2^j_{ik} \\
x^j_{ik} + x^j_{kl} - x^j_{il} \le 1 & \text{for all } \ i,k,l \in \mathcal{N}, i \neq k \neq l & \Gamma^j_{ikl} \\
x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i < k & \Delta^j_{ik} \\
x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \neq j & \Theta^j_i \\
x^j_{ik}, z^j_{ik} \ge 0 & \text{for all } i,k \in \mathcal{N}, i \neq k &
\end{array} \tag{31}$$

Let P denote the set $\{(i,k)\colon i \neq k, 0 \le i,k \le N-1\}$, and T denote the set $\{(i,k,l)\colon i \neq k \neq l, 0 \le i,k,l \le N-1\}$. Moreover, let $g(a,b,k,j)$ denote $\sum_{k\in\mathcal{N},k\neq a,b} \Gamma^j_{abk} + \sum_{k\in\mathcal{N},k\neq a,b} \Gamma^j_{kab} - \sum_{k\in\mathcal{N},k\neq a,b} \Gamma^j_{akb}$. Then, the LP we solve is

$$\underset{\nu,\alpha}{\text{maximize}} \quad \nu + \textstyle\sum_{(i,k)\in P} \alpha_{ik} y_{ik}$$

subject to

$$\begin{array}{ll}
\displaystyle\sum_{(i,k,l)\in T} \Gamma^j_{ikl} + \sum_{(i,k)\in P, i<k} \Delta^j_{ik} + \sum_{i\in\mathcal{M}, i\neq j} \Theta^j_i \le p_j - \nu \ \forall\, j \in \mathcal{M} & \\
g(a,b,k,j) + \Delta^j_{ab} - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a,b \in \mathcal{N}, a < b;\ \text{if } a = j, b \notin \mathcal{M} \\
g(a,b,k,j) + \Delta^j_{ba} - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a,b \in \mathcal{N}, a > b, b \neq 0;\ \text{if } a = j, b \notin \mathcal{M} \\
g(a,b,k,j) + \Delta^j_{ab} + \Theta^j_b - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a = j, b \in \mathcal{M}, a < b \\
g(a,b,k,j) + \Delta^j_{ba} + \Theta^j_b - \Omega 1^j_{ab} \ge 0 & \forall\ j \in \mathcal{M}, a = j, b \in \mathcal{M}, a > b, b \neq 0 \\
g(a,b,k,j) + \Delta^j_{ba} - \sum_{k\in\mathcal{N},k\neq a} \Omega 2^j_{ak} \ge 0 & \forall\ j \in \mathcal{M}, a \in \mathcal{N}, a \neq j, b = 0 \\
g(a,b,k,j) + \Delta^j_{ba} + \Theta^j_b - \sum_{k\in\mathcal{N},k\neq a} \Omega 2^j_{ak} \ge 0 & \forall\ j \in \mathcal{M}, a = j, b = 0 \\
\Omega 1^j_{ab} + \Omega 2^j_{ab} \ge \alpha_{ab} & \forall\ j \in \mathcal{M}, a,b \in P, a \neq 0, b \neq 0 \\
\Omega 2^j_{ab} \ge \alpha_{ab} & \forall\ j \in \mathcal{M}, a \in \mathcal{N} \setminus \{0\}, b = 0 \\
\Omega 1^j_{ab} \ge \alpha_{ab} & \forall\ j \in \mathcal{M}, a = 0, b \in \mathcal{N} \setminus \{0\} \\
\alpha, \Gamma, \Omega 1, \Omega 2 \ge 0. &
\end{array} \tag{32}$$

C. An Overview of Common Structural Models

Here we give a brief overview of each the parametric choice models we compare our approach with. The descriptions we provide are brief and we refer an interested reader to Ben-Akiva and Lerman (1985) for more details.

C.1. Multinomial logit (MNL) family

The MNL model is a popular and most commonly used parametric model in economics, marketing and operations management (see Ben-Akiva and Lerman (1985), Anderson et al. (1992)) and is the canonical example of a *Random Utility Model*. In the MNL model, the utility of the customer from product j takes the form $U_j = V_j + \xi_j$, where V_j is the deterministic component and the error terms $\xi_0, \xi_1, \ldots, \xi_{N-1}$ are i.i.d. random variables having a Gumbel distribution with location parameter 0 and scale parameter 1. Since only differences of utilities matter, without loss of generality, it is assumed that the mean utility of the "no-purchase" option, $V_0 = 0$. Let w_j denote e^{μ_j}; then, according to the MNL model, the probability that product j is purchased from an assortment $\mathcal{M}$ is given by

$$\mathbb{P}(j|\mathcal{M}) = w_j / \sum_{i\in\mathcal{M}} w_i.$$

A major advantage of the MNL model is that it is analytically tractable. In particular, it has a closed form expression for the choice probabilities. However, it has several shortcomings. One of the major limitations of the MNL model is that it exhibits Independent of Irrelevant Alternatives (IIA) property i.e., the relative likelihood of the purchase of any two given product variants is independent of the other products on offer. This property may be undesirable in contexts where some product are 'more like' other products so that the

randomness in a given customers utility is potentially correlated across products. There are other – more complicated – variants that have been proposed to alleviate the IIA issue – the most popular being the NL model, which we describe next.

C.2. Nested logit family (NL)

The nested logit (NL) family of models, first derived by Ben-Akiva (1973), was designed to explicitly capture the presence of shared unobserved attributes among alternatives. In particular, the universe of products is partitioned into L mutually exclusive subsets called *nests* denoted by $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_L$ such that

$$\mathcal{N} = \bigcup_{\ell=1}^{L} \mathcal{N}_\ell \qquad \text{and} \qquad \mathcal{N}_\ell \cap \mathcal{N}_m = \emptyset, \text{ for } m \neq \ell.$$

The partitioning is such that products sharing unobserved attributes lie in the same nest. Each customer has utility U_j for product j given by $U_j = V_j + \xi_\ell + \xi_{j,\ell}$; here, ξ_ℓ is the error term shared by all the products in nest $\mathcal{N}_\ell$, and $\xi_{j,\ell}$ is the error term that is product specific and assumed to be i.i.d across different products. In this logit case, $\xi_{j,\ell}$ are assumed to be i.i.d standard Gumbel distributed with location parameter 0 and scale parameter 1. The nest specific error terms $\xi_1, \xi_2, \ldots, \xi_L$ are assumed i.i.d., distributed such that for each ℓ, j, $\xi_\ell + \xi_{j,\ell}$ is Gumbel distributed with location parameter 0 and scale parameter $\rho < 1$. The no-purchase option is assumed to be in a nest of its own. Let w_j denote e^{μ_j} and let

$$w(\ell, \mathcal{M}) \stackrel{\text{def}}{=} \sum_{i \in \mathcal{N}_\ell \cap \mathcal{M}} w_i.$$

Then, with the above assumptions on the error terms, it can be shown that (see Ben-Akiva and Lerman (1985)) the probability that product j is purchased when offered assortment $\mathcal{M}$ is

$$\mathbb{P}(j|\mathcal{M}) = \mathbb{P}(\mathcal{N}_\ell|\mathcal{M})\,\mathbb{P}(j|\mathcal{N}_\ell, \mathcal{M}) = \frac{(w(\ell, \mathcal{M}))^\rho}{\sum_{m=1}^{L} (w(m, \mathcal{M}))^\rho} \; \frac{w_j}{w(\ell, \mathcal{M})}. \tag{33}$$

Nested MNL models alleviate the issue of IIA exhibited by the MNL models. Further, they have a closed form expression for the choice probabilities, which makes them computationally tractable. However, these models still exhibit IIA property within a nest. Moreover, it is often a challenging task to partition the products into different nests. Further, the model is limited because it requires each product to belong to exactly one nest.

The fact that NL model requires each product to be placed in exactly one nest is a limitation in several applications where a particular product is correlated with products across nests; for example, the no-purchase option in our setup is clearly correlated with all the products. In order to overcome this problem the NL model was extended to a *Cross Nested Logit (CNL)* model where each product can belong to multiple nests. The name cross-nested seems to be due to Vovsha (1997) and Vovsha's model is similar to the Ordered GEV model proposed by Small (1987). For our experiments, we assume that the no-purchase option has membership in all the nests and all other products have membership in only one nest. For this formulation, the probability of purchase of product j is given by (33) (see Ben-Akiva and Lerman (1985)), where $w(\mathcal{M}, \ell)$ is now defined as

$$w(\ell, \mathcal{M}) \stackrel{\text{def}}{=} \alpha_\ell w_0 + \sum_{i \in (\mathcal{N}_\ell \cap \mathcal{M}) \setminus \{0\}} w_i.$$

Here α_ℓ is the parameter capturing the level of membership of the no-purchase option in nest ℓ. The following conditions are imposed on the parameters α_ℓ, $\ell = 1, 2, \ldots, L$

$$\sum_{\ell=1}^{L} \alpha_\ell^\rho = 1, \quad \alpha_\ell \geq 0, \text{ for } \ell = 1, 2, \ldots, L.$$

The first condition is a normalization condition that is imposed because it is not possible to identify all the parameters. In our setup, it is only natural to assume that the no-purchase option has equal membership in all nests. This assumption translates into assuming that $\alpha_\ell = (1/L)^{1/\rho}$, for all ℓ. Further, we note that we assume without loss of generality that $V_0 = 1$ since only differences between utilities matter.

While the CNL model overcomes the limitations of the NL model, it is less tractable and Marzano and Papola (2008) showed that it cannot capture all possible types of correlations among products. Further, both

Table 2 Relevant attributes of DVDs from Amazon.com data and mean utilities of MNL model fit by Rusmevichientong et al. (2008)

Product ID	Mean utility	Price (dollars)	Average price per disc (dollars)	Total number of helpful votes
1	-4.513	115.49	5.7747	462
2	-4.600	92.03	7.6694	20
3	-4.790	91.67	13.0955	496
4	-4.514	79.35	13.2256	8424
5	-4.311	77.94	6.4949	6924
6	-4.839	70.12	14.0242	98
7	-4.887	64.97	16.2423	1116
8	-4.757	49.95	12.4880	763
9	-4.552	48.97	6.9962	652
10	-4.594	46.12	7.6863	227
11	-4.552	45.53	6.5037	122
12	-3.589	45.45	11.3637	32541
13	-4.738	45.41	11.3523	69
14	-4.697	44.92	11.2292	1113
15	-4.706	42.94	10.7349	320

the MNL and NL models don't account for heterogeneity in customer tastes. The MMNL family of models, described next, explicitly account for heterogeneity in customer tastes.

C.3. Mixed multinomial logit (MMNL)[10] family

The Mixed multinomial logit (MMNL) family of models is the most general of the three parametric families we compare our approach with. In fact, it is considered to be the most widely used and the most promising of the discrete choice models currently available Hensher and Greene (2003). It was introduced by Boyd and Mellman (1980) and Cardell and Dunbar (1980). McFadden and Train (2000) show that under mild regularity conditions, an MMNL model can approximate arbitrarily closely the choice probabilities of *any* discrete choice model that belongs to the class of RUM models. This is a strong result showing that MMNL family of models is very rich and models can be constructed to capture various aspects of consumer choice. In particular, it also overcomes the IIA limitation of the MNL and nested MNL (within a nest) families of models. However, MMNL models are far less computationally tractable than both the MNL and nested MNL models. In particular, there is in general no closed form expression for choice probabilities, and thereby the estimation of these models requires the more complex simulation methods. In addition – and more importantly – while the MMNL family can in principle capture highly complex consumer choice behavior, appropriate choice of features and distributions of parameters is highly application dependent and is often very challenging Hensher and Greene (2003).

In this model, the utility of customer c from product j is given by $U_{c,j} = \beta_c^T x_j + \varepsilon_{c,j}$, where x_j is the vector of *observed* attributes of product j; β_c is the vector of regression coefficients that are *stochastic* and not fixed[11] to account for the *unobserved* effects that depend on the *observed explanatory variables*; and $\varepsilon_{c,j}$ is the stochastic term to account for the rest of the unobserved effects. In this logit context, it is assumed that the variables $\varepsilon_{c,j}$ are i.i.d across customers and products Gumbel distribution of location parameter 0 and scale parameter 1. The distribution chosen for β_c depends on the application at hand and the variance of the components of β_c accounts for the heterogeneity in customer tastes. Assuming that β has a distribution $G(\beta;\theta)$ parameterized by θ, probability that a particular product j is purchased from assortment $\mathcal{M}$ is

$$\mathbb{P}(j|\mathcal{M}) = \int \frac{\exp\{\beta^T x_j\}}{\sum_{i\in\mathcal{M}} \exp\{\beta^T x_i\}} G(d\beta;\theta).$$

References

S. P. Anderson, A. De Palma, and J. F. Thisse. *Discrete choice theory of product differentiation*. MIT press, Cambridge, MA, 1992.

[10] This family of models is also referred to in the literature as Random parameter logit (RPL), Kernel/Hybrid logit.

[11] This is unlike in the MNL model where the coefficients are assumed to be fixed, but unknown.

M. E. Ben-Akiva. *Structure of passenger travel demand models*. PhD thesis, Department of Civil Engineering, MIT, 1973.

M. E. Ben-Akiva and S. R. Lerman. *Discrete choice analysis: theory and application to travel demand*. CMIT press, Cambridge, MA, 1985.

G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman Rev. Ser. A*, 5:147–151, 1946.

J. H. Boyd and R. E. Mellman. The effect of fuel economy standards on the u.s. automotive market: An hedonic demand analysis. *Transportation Research Part A: General*, 14(5-6):367 – 378, 1980.

N. S. Cardell and F. C. Dunbar. Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General*, 14(5-6):423 – 434, 1980.

D. A. Hensher and W. H. Greene. The mixed logit model: the state of practice. *Transportation*, 30(2): 133–176, 2003.

S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. In *Advances in Neural Information Processing Systems*, volume 21, pages 753–760, 2008.

V. Marzano and A. Papola. On the covariance structure of the cross-nested logit model. *Transportation Research Part B: Methodological*, 42(2):83 – 98, 2008.

D. McFadden and K. Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15 (5):447–470, September 2000.

E. H. McKinney. Generalized birthday problem. *American Mathematical Monthly*, pages 385–387, 1966.

P. Rusmevichientong, Z. J. Shen, and D. B. Shmoys. Dynamic Assortment Optimization with a Multinomial Logit Choice Model and Capacity Constraint. Technical report, Working Paper, 2008.

K. A. Small. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society*, 55(2):409–424, 1987.

J. von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *In Contributions to the theory of games*, 2, 1953.

P. Vovsha. Cross-nested logit model: an application to mode choice in the Tel-Aviv metropolitan area. *Transportation Research Record*, 1607:6–15, 1997.