Appendix: A Data-Driven Approach to Modeling Choice

In this appendix, we give the proofs of Theorems 1, 2 and 3.

Proof of Theorem 1. 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 2. 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 $1 \leq i \leq K$, 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. ■

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 [9].
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 $1 \leq j \leq J$, when we generate the matrix B by choosing K permutations uniformly at random, let $\mathscr{E}_j$ denote the event that the $K \times K$ matrix B^j is a permutation matrix and let $\mathscr{E}$ denote the event $\cup_j \mathscr{E}_j$. We want to prove that $\mathbb{P}(\mathscr{E}) \to 1$ as $N \to \infty$ as long as $K = o(\log N)$. Let X_j denote the indicator variable of the event $\mathscr{E}_j$, and X denote $\sum_j X_j$. Then, it is easy to see that $\Pr(X=0) = \Pr((\mathscr{E})^c)$. 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:

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

   It thus follows that $\mathbb{P}(X=0) \leq \operatorname{Var}(X)/(\mathbb{E}[X])^2$. We now evaluate $\mathbb{E}[X]$. Since X_js are indicator variables, $\mathbb{E}[X_j] = \mathbb{P}(X_j = 1) = \mathbb{P}(\mathscr{E}_j)$. In order to evaluate $\mathbb{P}(\mathscr{E}_j)$, 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, $\mathbb{P}(\mathscr{E}_j) = K!/2^{K^2}$. Thus, we have:

   $$\mathbb{E}[X] = \sum_{j=1}^{J} \mathbb{E}[X_j] = \sum_{j=1}^{J} \mathbb{P}(\mathscr{E}_j) = \frac{JK!}{2^{K^2}}. \tag{10}$$

   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 (10) that

   $$\mathbb{E}[X] \geq \frac{L^K}{e^K K^K}\frac{K^K}{e^K}\frac{1}{2^{K^2}} = \frac{L^K}{e^{2K}2^{K^2}}. \tag{11}$$

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

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

   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,

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

It now follows that for $K = o(\log L/\log(4e))$, $\Pr(X = 0) \to 0$ 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" [11], the K balls falls into different bins with a high probability provided $K = o(\sqrt{N})$.

This finishes the proof of the theorem. ■