RAND Journal of Economics
Vol. 38, No. 4, Winter 2007
pp. 1002–1019

Semiparametric estimation of multinomial discrete-choice models using a subset of choices

Jeremy T. Fox*

Nonlogit maximum-likelihood estimators are inconsistent when using data on a subset of the choices available to agents. I show that the semiparametric, multinomial maximum-score estimator is consistent when using data on a subset of choices. No information is required for choices outside of the subset. The required conditions about the error terms are the same conditions as for using all the choices. Estimation can proceed under additional restrictions if agents have unobserved, random consideration sets. A solution exists for instrumenting endogenous continuous variables. Monte Carlo experiments show the estimator performs well using small subsets of choices.

1. Introduction

■ Demand estimates are typically inconsistent when estimation uses data on a subset of the choices available to agents in the data-generating process. This article develops an unordered, discrete-choice estimator that is consistent with data on a subset of the available choices.

One motivation for estimation using data on a subset of choices is data availability. Chevalier and Goolsbee (2003) and Bajari, Fox, and Ryan (2006) use data on purchases from the online retailer Amazon. Not all products in a given retail category are offered on Amazon. One assumption is that a consumer sees all products, both on and offline, and only buys on Amazon if the chosen product is offered by Amazon. Using data from one retailer imposes choice-based sampling.

Another motivation for using a subset of all choices is that the computational burden of estimating discrete-choice models increases with the size of the choice set. Large choice sets occur often in applications. Train, McFadden, and Ben-Akiva (1987) estimate the demand for telephone calling plans, where each plan is a combination of several options such as a monthly fee, usage charge, and so forth. Bayer, McMillan, and Reuben (2004) estimate a housing-demand model where the choices are hundreds of thousands of houses in a large metropolitan area. Bajari

* University of Chicago; fox@uchicago.edu.
Thanks to Patrick Bajari, Jean-Pierre Dube, Matthew Gentzkow, Christian Hansen, James Heckman, Han Hong, Thomas MaCurdy, Aprajit Mahajan, Rosa Matzkin, Aviv Nevo, Susanne Schennach, Frank Wolak, anonymous referees, and the Editor Philip Haile, as well as workshop participants at Chicago Econ, Chicago Marketing, Northwestern, and Stanford for helpful comments.

and Fox (2007) study an FCC spectrum auction, where the combinatorics of gathering individual spectrum licenses for sale into packages of multiple licenses results in a choice set with more elements than the number of atoms in the universe.

The only consistent estimators using data on a subset of choices to compute choice probabilities as though the true choice set is the subset with data used in estimation have been developed by McFadden (1978) and Bierlaire, Bolduc, and McFadden (2006). These estimators rely on known closed-form choice probabilities for a class of discrete-choice models characterized by the error terms having a "block additive generalized extreme value" (GEV) distribution. The block additive GEV class includes only the pure multinomial logit among well-known estimators; the class does not include the nested logit and methods with auxiliary distributions for agent heterogeneity, such as the random coefficients and mixed logit.

The GEV class is defined to include only choice models where the marginal distributions for the error terms are type I extreme value with a common scale parameter. The type I extreme value distribution is restrictive. For example, it rules out the economically interesting possibility that the marginal distribution has multiple modes: either a consumer hates or loves a product.

Although I am motivated by the same computational and missing data on choice characteristics concerns as McFadden (1978), this article uses a different set of mathematical tools to address estimation. I work with semiparametric discrete-choice models, where the term "semiparametric" refers to the fact that I specify a set of parameters to estimate, but I do not specify a particular functional form for the error term. Working with semiparametric models forces me to consider identification using properties of models that hold across a wide number of possible distributions for the error terms, rather than imposing a known function to directly compute choice probabilities. A semiparametric proof of consistency is also a constructive proof of semiparametric identification, so this article clarifies the identification of multinomial choice models using data on a subset of choices.

The pioneering work of Manski (1975) introduces semiparametric maximum-score estimation for discrete-choice models. Maximum score estimators are consistent when the choice probabilities for a given agent are rank ordered by the agent's deterministic choice payoffs. The property of rank ordering choice probabilities by non-stochastic payoffs can hold across a wide number of distributions for the error terms, and is the key to semiparametric identification and estimation using maximum-score methods. This article's major contribution is to prove that multinomial maximum score is consistent using data on a subset of choices. The reason is that when one conditions on observations that selected a choice from a subset, the choice probabilities in the subset are still rank ordered by the deterministic payoffs. All maximum score needs is rank ordering, and so maximum score is consistent when the researcher has choice and covariate data on a subset of the choices available to decision makers. More choices could improve the finite-sample accuracy of the estimates, but only two choices are needed for consistency. Although I modernize and therefore weaken the sufficient conditions on the covariates and error terms needed for consistency of multinomial maximum-score estimators in Manski (1975), the assumptions for consistency using a subset of choices are the same as or weaker than the conditions for using data on all choices.

I introduce a multinomial maximum-score estimator that focuses on comparisons between the deterministic payoffs of pairs of choices. I call the estimator pairwise maximum score. Pairwise maximum score makes full use of the rank-order property driving identification and consistency by considering the relative ranking of choice probabilities for, if included, all pairs of choices. By using the restrictions of the rank-order property for comparisons of all pairs of choices, pairwise maximum score may improve the finite-sample accuracy of the estimates.

Evaluating the pairwise maximum-score objective function requires only addition, multiplication, and pairwise comparison. Further, the estimator is consistent using covariate and choice data on a subset of choices in the true model. Compare this to estimating a semiparametric or parametric maximum-likelihood method using covariate data on all choices in the true model. In maximum likelihood with a large number of choices, a fitted probability that is a function of

perhaps millions of covariates needs to be computed for each parameter value. Computing this fitted probability can be a daunting numerical challenge that involves evaluating densities in the far right tails. The role of maximum score as a computationally simple alternative to maximum likelihood differs from the usual pitch for semiparametric methods, which concerns relaxing distributional assumptions.

I present implementation advice about how to use pairwise maximum score in applications. The availability of new global optimization routines has decreased the computational difficulty of numerically maximizing a step function. I also introduce a two-stage instrumental variables estimator.

I present Monte Carlo studies about the finite-sample properties of maximum score when using data on a subset of choices. The Monte Carlo experiments compare maximum score to the parametric logit estimator for a true model where the error terms do not have the extreme value distribution, so that the logit is inconsistent.

This article revisits the multinomial maximum-score estimator of Manski (1975). Most other work on maximum-score estimators, including Manski's later work, focuses on the binary (two) choice case. I do not believe multinomial maximum-score methods have been used in any applications other than Briesch, Chintagunta, and Matzkin (2002) and my own (Bajari and Fox, 2007; Bajari, Fox, and Ryan, 2006). By improving and describing new properties of multinomial maximum score, I hope to make maximum score an attractive and practical method that can be used by applied researchers.

2. Rank ordering of choice probabilities

■ The identification strategy relies only on the rank ordering of choice probabilities. This section defines rank ordering and provides sufficient conditions on the distribution of errors in a random utility model for rank ordering.

□ **Model.** The model is completely standard. Consider a single-agent, unordered discrete-choice, random-utility model. An agent makes a choice among $i = 1, \ldots, J$ products. In a duplication of notation, J refers to both the set of choices and the number of choices. The computational cost in estimation that this article in part addresses arises when J is large. The number and set of choices can vary from agent to agent. Ignoring ties, the agent picks choice i if

$$u_i > u_j \forall j \in J, j \neq i. \tag{1}$$

The agent chooses i when the payoff from i exceeds the payoffs from all $J - 1$ alternatives. If i satisfies (1), I refer to i as the selection.

The payoff from choosing $i \in J$ is

$$u_i = x_i'\beta + \varepsilon_i,$$

where x_i is a vector of d covariates, β is a vector of d parameters multiplying x_i, and ε_i represents the sum of factors the agent feels are important but are unobserved to the econometrician. Let the $J \times d$ matrix x be the observable covariates for all choices. Also let ε be the vector of all J ε_i's.

In applications, x_i is typically formed from the characteristics of choice i and the interactions of the characteristics of choice i with the characteristics of the agent. Point identification requires varying the x's across observations on agents. Variation across agents can come from differences in the characteristics of agents facing the same choice set, or from agents facing different choice sets.[1]

[1] The article assumes the standard $x'\beta + \varepsilon$ parametric functional form, although the maximum-score estimators do not exploit any special properties of this functional form. A previous draft used results from Matzkin (1993) to replace $x'\beta$ with a nonparametric function of x. As in Abrevaya (2000), I also replaced additive separability between $x'\beta$ and ε with weak separability in an unknown monotone function. These weaker restrictions make the maximum-score estimator nonparametric.

□ **The rank ordering property.** The estimators in this article are semiparametric, as I will not specify a parametric functional form for $F(\varepsilon \mid x, J)$, the distribution function of the J ε_i's. This distinguishes semiparametric estimation from parametric estimators such as the logit and probit, which assume a known functional form for $F(\varepsilon \mid x, J)$. Instead, semiparametric estimators require that some particular property of the underlying choice model holds across a range of functional forms for the distribution of the error terms.

For the case of discrete choice, the pioneering work of Manski (1975) introduces the property that a given agent makes choices that have higher deterministic payoffs with greater frequency. The deterministic payoffs of choices rank order the choice probabilities. Manski's formal property is

Assumption 1. For a given agent, and for $i, j \in J$,

$$x_i'\beta > x_j'\beta \tag{2}$$

if and only if

$$P(i \mid x, J, \beta) > P(j \mid x, J, \beta).$$

Here $P(i \mid x, J, \beta)$ is the probability of i being selected. The probability is an integral over the J unknown ε_i's over the domain where the decision rule in (1) is satisfied. The probability $P(i \mid x, J, \beta)$ is a function of the $J \times d$ matrix x and the number of choices J. Therefore, Assumption 1 is a property of integrating out the unknown ε_i's over the correct domain, while fixing the matrix x and therefore the payoffs $x_i'\beta$ for all choices. As a consequence, the specific linear functional form $x_i'\beta$ is irrelevant to whether Assumption 1 holds.

Assumption 1 states that the contributions $x_i'\beta$ of observable characteristics *rank order* the choice probabilities for any given agent. Choices with higher deterministic payoffs are more likely to be chosen. Assumption 1 holds for two choices at a time, regardless of the true number of choices J. The pairwise comparison property inherent in rank ordering is why the estimators in this article are consistent using data on a subset of choices. This will be elaborated in more detail soon.

Assumption 1 has nothing to do with comparisons of choices across agents. For a given agent, choice probabilities are rank ordered by deterministic payoffs, but two agents with only slightly different deterministic payoffs can make choices with different probabilities. For example, male agents can have errors from Laplace distributions, and female agents can have errors from normal distributions, and this difference in error distributions will generate different choice probabilities. Assumption 1 allows the functional form for the distribution of the errors to vary relatively flexibly with x under the strong sufficient conditions mentioned below. The fact that the maximum-score estimators in this article do not rely on the same shape restrictions holding across agents also distinguishes maximum score from parametric methods such as the logit and probit.

□ **Rank ordering holds for a subset of choices.** In this article, estimation uses data on choices and covariates for choices in $K \subseteq J$. The rank ordering of choice probabilities still holds when conditioning on a set of endogenous choices K. The preservation of the rank ordering of Assumption 1 through conditioning is the key to understanding why maximum-score estimation is consistent when using data on a subset of choices.

The underlying population random variables generating the data are still $x_1, \ldots, x_J$ and $\varepsilon_1, \ldots, \varepsilon_J$, as well as possibly J itself. Define $P_K(i \mid x, J, \beta)$ to be the probability of picking choice i conditional on making a choice in the subset K. It is easy to show that choice probabilities are still rank ordered in the subset of choices K. If

$$P(i \mid x, J, \beta) > P(j \mid x, J, \beta),$$

then by division of a positive number,

$$\frac{P(i \mid x, J, \beta)}{\sum_{h \in K} P(h \mid x, J, \beta)} > \frac{P(j \mid x, J, \beta)}{\sum_{h \in K} P(h \mid x, J, \beta)},$$

which is just

$$P_K(i \mid x, J, \beta) > P_K(j \mid x, J, \beta).$$

Therefore, conditioning on a choice set preserves the rank ordering of the choice probabilities by the deterministic payoffs. The following result follows directly from Assumption 1.

Lemma 1. Under Assumption 1, for an agent who is known to have made a choice in a subset of choices K, and for choices $i, j \in K$,

$$x_i'\beta > x_j'\beta \tag{3}$$

if and only if

$$P_K(i \mid x, J, \beta) > P_K(j \mid x, J, \beta).$$

The covariates for all J choices, the $J \times d$ matrix x, enter even the conditional choice probabilities. The choice-based data are an issue only for the econometrician. The data are generated by agents who have information on all J choices. Maximum-score estimators do not involve computing $P_K(i \mid x, J, \beta)$.

□ **Sufficient conditions for Assumption 1.** Assumption 1 is satisfied by a wide class of functional forms for $F(\varepsilon \mid x, J)$, but also is restrictive in that many other popular parametric specifications do not satisfy Assumption 1.

Manski (1975) shows that a sufficient condition for Assumption 1 is that each ε_i has *support equal to the real line* and an absolutely continuous, *independent*, and *identical* distribution for all choices. The exact distribution $F(\varepsilon_i \mid x, J)$ can vary across agents with the entire $J \times d$ matrix x, but under i.i.d. errors the marginal distribution $F(\varepsilon_i \mid x, J)$ must be the same across all choices for a given agent, and further the random variables ε_i must be independent across all choices for a given agent.

In a discussion of the empirical content of a game-theoretic concept known as quantal response equilibrium, Goeree, Holt, and Palfrey (2004) prove that a weaker sufficient condition for Assumption 1 is as follows.

Assumption 2. The errors ε have an absolutely continuous joint distribution with full support on $\mathbb{R}^J$. The associated joint density $f(\varepsilon_1, \ldots, \varepsilon_J \mid x, J)$ exists and is exchangeable.

Let Ψ_J be the set of all permutations of J objects. The joint density of the errors is exchangeable if, for all permutations $\psi \in \Psi_J$,

$$f(\varepsilon_1, \ldots, \varepsilon_J \mid x, J) = f(\varepsilon_{\psi(1)}, \ldots, \varepsilon_{\psi(J)} \mid x, J).$$

Exchangeability is also known as interchangeability and label independence. Any collection of i.i.d. random variables has an exchangeable joint density. The most common example of an exchangeable density for nonindependent random variables is a multivariate normal density with equicorrelation, or

$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_J \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho & \rho \\ \rho & \ddots & \rho \\ \rho & \rho & \sigma^2 \end{bmatrix} \right).$$

The correlation ρ between all choices J is the same, and the univariate variance σ^2 of each choice is also identical. For Assumption 1 to hold, the multivariate normal parameters μ, ρ, and σ^2 can vary randomly across agents, and can be arbitrary functions of the $J \times d$ matrix x. However, each agent must have only the three parameters μ, σ^2, and ρ.

Consider a heterogeneous taste v_a for quality $x_i'\beta$. Thus let $u_{ai} = (x_i'\beta)v_a + \varepsilon_{ai}$, where v_a is an agent-specific random variable. Here v_a scales the utility parameters of agent a. As the rank-order property does not involve comparisons of choices across agents, the relative scales of utility across agents can vary flexibly. What is important is that the conditional on x variance of u_{ai} can be written as a weakly increasing function of $x_i'\beta$, in place of a function of each of the components of x_i separately.

There are economically interesting models that do not satisfy Assumption 1. With product characteristics representing the main variation in x_i across choices, parametric demand specifications often rely on random coefficients to add unobserved heterogeneity in tastes for observable characteristics (Berry, Levinsohn, and Pakes, 1995). A random-coefficient specification for u_{ai} for agent a is

$$u_{ai} = x_i'\beta_a + \eta_{ai} = x_i'\beta + x_i'(\beta_a - \beta) + \eta_{ai} = x_i'\beta + \varepsilon_{ai},$$

where the random coefficient β_a enters the composite error term $\varepsilon_{ai} = x_i'(\beta_a - \beta) + \eta_{ai}$, and the object of estimation β is interpreted as the mean random coefficient in the population. Random coefficients generate heteroskedasticity across choices that cannot be written as a weakly increasing function of only $x_i'\beta$. Random coefficients also generate correlation between choices.

The deterministic payoffs $x_i'\beta$ may not rank order the choice probabilities under heteroskedasticity across choices, for a given agent. A choice with a greater variance and the same mean payoff $x_i'\beta$ as another choice will have more frequent draws from the right tails of the support of the error distribution. The decision rule in (1) shows that maximal order statistics drive discrete choices. A choice with a higher variance will have more draws from the right tail and will be made with a higher probability. This violates Assumption 1, which says that choices with the same mean payoffs must be made with equal probabilities.

The rank ordering of choice probabilities by deterministic payoffs in Manski's Assumption 1 is strong for many applications such as demand estimation. Bajari, Fox, and Ryan (2006) show that estimation can allow for agent-specific fixed effects over nests of alternatives. Fixed effects allow the error terms to be correlated within classes of similar products, one of the motivations behind random coefficients.

For binary choice ($J = 2$), Manski (1975) and a large follow-up literature, such as Manski (1985) and Horowitz (1992), detail the properties of maximum-score estimators. In binary choice, an agent who does not pick choice 1 must pick choice 2, so these authors choose to work with a single composite error term $\varepsilon_1 - \varepsilon_2$ and a single covariate vector $x_1 - x_2$. In binary choice,

$$P(1 \mid x, 2, \beta) + P(2 \mid x, 2, \beta) = 1,$$

so choice 1 has a greater probability when its probability is greater than $1/2$. Substituting the CDF of $\varepsilon_1 - \varepsilon_2$ and applying the decision rule in (1) gives

$$P(1 \mid x, 2, \beta) = 1 - F_{\varepsilon_1 - \varepsilon_2}(-(x_1 - x_2) \mid x, \beta) = \frac{1}{2} \text{ if } x_1 - x_2 = 0$$

and the composite error term $\varepsilon_1 - \varepsilon_2$ has a median of 0 regardless of the value of $x_1 - x_2$. The rank ordering of choice probabilities by deterministic payoffs in Assumption 1 follows from this median independence assumption.

For binary choice, random coefficients

$$u_{a1} - u_{a2} = (x_1 - x_2)'\beta + (x_1 - x_2)'(\beta_a - \beta) + \eta_{a1} - \eta_{a2} = (x_{a1} - x_{a2})'\beta + \varepsilon_{a1} - \varepsilon_{a2}$$

can be tolerated, as some distributions for the random coefficients preserve median independence. To see this for an example, let $\beta_a \sim N(0, \Sigma)$ and $\eta_{a1} - \eta_{a2} \sim N(0, \sigma_\eta^2)$. Then

$$\varepsilon_{a1} - \varepsilon_{a2} \sim N\big(0, \sigma_\eta^2 + (x_{a1} - x_{a2})'\Sigma(x_{a1} - x_{a2})\big)$$

and has median 0 conditional on x_1 and x_2. With two choices, random coefficients induce only heteroskedasticity across agents in $\varepsilon_{a1} - \varepsilon_{a2}$. Heteroskedasticity across agents does not affect the

rank-order property, which involves orderings of choice probabilities across choices for a given agent.

□ **Rank ordering with unobserved, random consideration sets.** It may be unrealistic to assume that a decision maker is aware of his or her payoffs for all J choices. An alternative is that each agent is aware of a subset H of the J available choices. H is not observable to the econometrician. The set H is often called a consideration set (Mehta, Rajiv, and Srinivasan, 2003).

Let $P^S(H \mid x, J)$ be the probability that an agent is aware of the choices in H. The probability of an agent choosing product i conditional on J (but not H) is

$$P(i \mid x, J, \beta) = \sum_{H \subseteq J \text{ s.t. } i \in H} P^S(H \mid x, J) P(i \mid x^H, H, \beta),$$

where x^H is the matrix of characteristics for the products in H. I derive conditions on $P^S(H \mid x, J)$ so that the rank-order property, Assumption 1, holds for the entire choice set J when agents have random and unobserved consideration sets. If Assumption 1 holds, then Lemma 1 holds, and subset estimation using the typically nonrandom set $K \subseteq J$ can proceed. For example, β can be estimated using data on two choices, even if some consumers were not aware of one of the two choices.

I maintain that the rank-order property holds for each decision problem H. The key additional assumption is that decision makers are weakly more likely to be aware of a set of choices than an otherwise identical set where a choice is swapped for a choice with a lower deterministic payoff.

Lemma 2. Assume:

(i) For all $H \subseteq J$, $\{\varepsilon_j\}_{j \in H}$ has an exchangeable, absolutely continuous joint distribution with support $\mathbb{R}^{|H|}$.
(ii) For all $H \subseteq J$, for any $i, j \in H$ and $H \subset J \backslash \{i, j\}$, $P^S(H \cup \{i\} \mid x, J) \geq P^S(H \cup \{j\} \mid x, J)$ if and only if $x_i'\beta \geq x_j'\beta$.
(iii) For two sets $H_1 \subseteq J$ and $H_2 \subseteq J$, where $|H_1| = |H_2|$, $F_{H_1}(\varepsilon_1, \ldots, \varepsilon_{H_1}) = F_{H_2}(\varepsilon_1, \ldots, \varepsilon_{H_2})$.

Then $x_i'\beta \geq x_j\,\beta$ if and only if $P(i \mid x, J, \beta) \geq P(j \mid x, J, \beta)$.

Proof. $P(i \mid x, J, \beta) - P(j \mid x, J, \beta)$ can be decomposed into the sum of four parts or

$$\begin{aligned}
&\sum_{H \subseteq J} P^S(H \mid x, J)(P(i \mid x^H, H, \beta) - P(j \mid x^H, H, \beta)) \\
&\quad = \sum_{H \subseteq (J \backslash \{i,j\}) \cup \emptyset} \Big[P^S(H \mid x, J) \cdot 0 + P^S(H \cup \{i\} \mid x, J) \cdot P\big(i \mid x^{H \cup \{i\}}, H \cup \{i\}, \beta\big) \\
&\qquad - P^S(H \cup \{j\} \mid x, J) \cdot P(j \mid x^{H \cup \{j\}}, H \cup \{j\}, \beta) + P^S(H \cup \{i, j\} \mid x, J) \\
&\qquad \cdot (P(i \mid x^{H \cup \{i,j\}}, H \cup \{i, j\}, \beta) - P(j \mid x^{H \cup \{i,j\}}, H \cup \{i, j\}, \beta)) \Big].
\end{aligned}$$

For the "only if" part of the lemma, let $x_i'\beta \geq x_j'\beta$. Fix $H \subseteq J \backslash \{i, j\}$. I will show that the term in the summation is nonnegative. As this is true for all $H \subseteq J \backslash \{i, j\}$, the entire sum is nonnegative, so $P(i \mid x, J, \beta) \geq P(j \mid x, J, \beta)$. The result of Goeree, Holt, and Palfrey (2004) shows that an exchangeable joint density leads to the rank-order property: $x_i'\beta \geq x_j'\beta$ iff $P(i \mid x^{H \cup \{i,j\}}, H \cup \{i, j\}, \beta) > P(j \mid x^{H \cup \{i,j\}}, H \cup \{i, j\}, \beta)$. In the summation, the first term is the case where neither i nor j is in H, the second term is the case where i is in H but j is not, the third term is when j is in H but i is not, and the fourth term is when both i and j are in H.

The term inside the summation can only be negative if $P^S(H \cup \{i\} \mid x, J) \cdot P(i \mid x^{H \cup \{i\}}, H \cup \{i\}, \beta) < P^S(H \cup \{j\} \mid x, J) \cdot P(j \mid x^{H \cup \{j\}}, H \cup \{j\}, \beta)$. $P^S(H \cup \{i\} \mid x, J) \geq P^S(H \cup \{j\} \mid x, J)$ by the conditions of the lemma. The "only if" part of the lemma will hold if I

prove that $x_i'\beta \geq x_j'\beta$ implies $P(i \mid x^{H\cup\{i\}}, H \cup \{i\}, \beta) \geq P(j \mid x^{H\cup\{j\}}, H \cup \{j\}, \beta)$. Writing out the formula for choice probabilities gives

$$P(i \mid x^{H\cup\{i\}}, H \cup \{i\}, \beta) = \int_{-\infty}^{\infty} \int_{-\infty}^{x_i'\beta+\varepsilon_i-x_H'\beta} \cdots \int_{\infty}^{x_i'\beta+\varepsilon_i-x_1'\beta} dF_{H\cup\{i\}}(\varepsilon_1, \ldots, \varepsilon_H, \varepsilon_i),$$

where WLOG the elements of H are ordered $1, \ldots, H$, with the convention that the i in the upper limits is the i in $H \cup \{i\}$ rather than the ith element of $1, \ldots, H$. As $x_i'\beta$ enters only the upper limits of integrals, and the integrand is everywhere positive, then $P(i \mid x^{H\cup\{i\}}, H \cup \{i\}, \beta)$ is increasing in $x_i'\beta$. By the condition in the lemma that $F_{H\cup\{i\}} = F_{H\cup\{j\}}$, the formula for $P(j \mid x^{H\cup\{j\}}, H \cup \{j\}, \beta)$ is identical to $P(i \mid x^{H\cup\{i\}}, H \cup \{i\}, \beta)$ with $x_j'\beta$ replacing $x_i'\beta$. Therefore, $P(i \mid x^{H\cup\{i\}}, H \cup \{i\}, \beta) \geq P(j \mid x^{H\cup\{j\}}, H \cup \{j\}, \beta)$.

The "if" direction of the lemma assumes that $P(i \mid x, J, \beta) \geq P(j \mid x, J, \beta)$ and concludes that $x_i'\beta \geq x_j'\beta$. Assume not, so that $x_j'\beta > x_i'\beta$. Then by the above argument, $P(j \mid x, J, \beta) > P(i \mid x, J, \beta)$, a contradiction.

For space reasons, I do not explore the case where a researcher has observables v_i that enter the consideration set probabilities but that do not enter the payoff of a choice conditional on it being in a consideration set. Such exclusion restrictions might allow one to identify $P^S(H \mid x, v, J)$ from observations on $P(i \mid x, v, J, \beta)$. Goeree (2005) uses data on media exposure to help identify $P^S(H \mid x, v, J)$.

3. Data on a subset of choices

■ This section presents sufficient conditions on the sampling rule of the data for consistency of the maximum-score estimators that I will later introduce. The econometrician has access to a sample of data on agents who made selections in the subset K. K is also the number of observed choices, so $K \leq J$. If the econometrician knows that a particular agent was not selecting from all elements of K, the estimators can be modified to include this information.

The econometrician has an i.i.d. sample of n observations on agents who selected a choice from the set K. Importantly, the econometrician observes the covariate data x_i for all choices $i \in K$ regardless of the selection of the agent, as long as the selection is in K to begin with. Without a loss of generality, reindex the choices $i = 1, \ldots, K$. Let x^K be the $K \times d$ matrix of one agent's covariate data for choices in K. Estimation does not use data about the frequencies of agents who make choices outside of K, or about the characteristics of choices outside of K, or even about the number J of choices in the true model. This article does not consider the case where individual regressors in the matrix x^K are missing.

Point identification requires a special regressor with a continuous distribution. For each choice i, factor the vector of covariates x_i into $(x_{1i}, \tilde{x}_i)$, where $x_{1,i}$ is the first element of the vector of covariates, and $\tilde{x}_i$ is the $d-1$ other covariates. For the other covariates, the assumptions are similar to those made in parametric maximum-likelihood models, such as the logit and probit.

Assumption 3. For each pair of choices $i, j \in K$, let $w_{ij} = x_{1,i} - x_{1,j}$. The following properties are true.

(i) $G^K_{1,ij}(w_{ij} \mid \tilde{x}_i, \tilde{x}_j)$ is absolutely continuous with density $g^K_{1,ij}(w_{ij} \mid \tilde{x}_i, \tilde{x}_j)$.
(ii) $g^K_{1,ij}(w_{ij} \mid \tilde{x}_i, \tilde{x}_j)$ is nonzero everywhere on $\mathbb{R}$.
(iii) The parameter β_1 on the first covariate $x_{1,i}$ is nonzero.
(iv) The support of the covariates $x_i - x_j$ does not lie in a proper linear subspace of $\mathbb{R}^d$, where again $d = \dim\{x_i - x_j\}$.

The freely varying covariate $w_{ij} = x_{1,i} - x_{1,j}$ in Assumption 3 is a standard sufficient condition for point (versus set) identification in semiparametric discrete-choice models (Manski, 1985). Covariates $x_{1,i}$ and $x_{1,j}$ that are individually always positive can fulfill the role of $w_{ij} = x_{1,i} - x_{1,j}$ as long as sometimes $x_{1,i} > x_{1,j}$ by sufficient margins, and other times

$x_{1,i} < x_{1,j}$ by sufficient margins. In demand estimation, price is a good covariate to play the role of the continuously varying $x_{1,i}$, if price varies across agents in the sample. Alternatively, the continuously varying characteristic might be a consumer characteristic such as income that is interacted with some product characteristic.

The large support assumption on $w_{ij} = x_{1,i} - x_{1,j}$ can be weakened. Manski (1988) and Horowitz (1998) show that, under additional conditions, semiparametric discrete-choice models can be identified if the first element of the covariate vector has bounded support.

4. Pairwise maximum score

■ The econometrician wants to use the covariate and choice data to estimate β, the unknown parameters in the linear index of observable payoffs in utility. The econometrician is willing to assume the data are generated by the random-utility model in (1), but he or she does not want to choose a functional form for $F(\varepsilon \mid x, J)$, the distribution of the errors.

This section introduces a new objective function for multinomial maximum score. The objective function is inspired by the functional form for binary (two) choice maximum score, in that the pairwise maximum-score objective function compares pairs of choices at a time. It turns out that the pairwise maximum-score objective function produces the same objective-function value as a member of a general class of scoring rules introduced by Manski (1975). Later I compare pairwise maximum score to the functional forms considered in Manski's pioneering work.

□ **Only two choices used in estimation.** For illustration, consider an econometrician who wants to use data on only choices i and j. In this example, $K = 2$. Let there be $a = 1, \ldots, n$ observations, where every agent selects either i or j. A multinomial maximum-score estimator is then any parameter vector (the objective function is a step function) that maximizes

$$Q_n^{ij}(\beta) = \frac{1}{n}\sum_{a=1}^{n}(1[ai] \cdot 1[x'_{ai}\beta > x'_{aj}\beta] + 1[aj] \cdot 1[x'_{aj}\beta > x'_{ai}\beta]). \tag{4}$$

There are several things to note about this pairwise maximum-score estimator. The $1[\cdot]$ functions are indicator functions equal to 1 when the condition in brackets is true, and 0 otherwise. The dependent variable data are of the form $1[ai]$, which is equal to 1 if, in the data, agent a's selection is i. If a does select i, $1[x'_{ai}\beta > x'_{aj}\beta]$ enters the objective function.

For any agent i, at most one of the two indicator functions $1[x'_{ai}\beta > x'_{aj}\beta]$ and $1[x'_{aj}\beta > x'_{ai}\beta]$ can be 1. Therefore, the objective function increases by 1 when, for example, a selects i and a's deterministic payoff for i, $x'_{ai}\beta$, is greater than a's deterministic payoff for j. The term maximum score is appropriate because the econometrician maximizes the score of correct predictions according to Lemma 1. Now agent a could select j because of a high ε_{aj} draw even when $x'_{ai}\beta > x'_{aj}\beta$. This is expected for a handful of observations. However, asymptotically, if $x'_{ai}\beta > x'_{aj}\beta$, then under Lemma 1 there will be more instances where i is selected than j.

For a given β, $Q_n^{ij}(\beta)$ is the model's fraction of correct predictions according to Lemma 1. Therefore, $Q_n^{ij}(\beta)$ is a measure of model fit and can be listed in a table of output just as one would report R^2 in a linear regression.

Note that calculating $Q_n^{ij}(\beta)$ involves only addition, multiplication, and pairwise comparison. There are no numerical integrations and complex nonlinearities to work with.

□ **More than two choices.** It is easy to generalize the objective function in (4) to include data on more choices. Now the econometrician uses n observations, all of whom selected a choice in K. The pairwise maximum-score objective function using choice-based data on K choices is

$$Q_n^{K}(\beta) = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{1}{n}\sum_{a=1}^{n}(1[ai] \cdot 1[x'_{ai}\beta > x'_{aj}\beta] + 1[aj] \cdot 1[x'_{aj}\beta > x'_{ai}\beta]). \tag{5}$$

The estimator is any parameter vector $\hat{\beta}_n^K$ that maximizes $Q_n^K(\beta)$. It can be seen that $Q_n^K(\beta)$ repeats $Q_n^{ij}(\beta)$ for $\frac{1}{2}K(K-1)$ pairs of choices in K. Only a single pair of choices is needed for consistency, but adding more choices makes more use of the data.

A researcher has the flexibility to pick the set K if there are data on more than two choices. The researcher may want to balance computational speed and the finite-sample precision of the estimates of β.

Itemizing over all $\frac{1}{2}K(K-1)$ possible combinations of pairs of choices in K in (5) is not necessary for consistency. A more primitive notion is a pair of choices j and k such that $j \neq k$. A researcher can arbitrarily include any number of distinct pairs of choices.

The objective function is written as though K is fixed across observations. Say data on K choices are available, but the objective function uses only a subset I_a for observation a. The included subset I is random with probability $P(I \mid x^K)$. Point identification requires the support of the exogenous data to remain the same, so $P(I \mid x^K) > 0$ for all x^K and $I \subseteq K$. Randomly drawing I_a requires throwing away observation a with selection y_a if $y_a \notin I_a$.

□ **Consistency.** As a discrete choice is qualitative and cannot pin down a cardinalization of utility, one needs to impose location and scale normalizations on the parameter space. Therefore, make the following standard assumption.

Assumption 4. The parameter β^0 generating the data is known to lie in a compact space Θ that imposes location and scale normalizations.

I choose the normalization that makes the consistency proof the easiest. In particular, recall the freely varying covariate $x_{1,i}$, the first element of the covariate vector, from the covariates assumption. Assumption 3 states that $\beta_1 \neq 0$. Without any further loss of generality, choose the scale normalization $\beta_1 = \pm 1$. The entire parameter vector can be estimated by first estimating the model for $\beta_1 = 1$, and then reestimating the model for $\beta_1 = -1$, and picking the estimates of β to be the case with the largest objective-function value.

Pairwise maximum score is consistent under the same assumptions as the original multinomial maximum-score estimator of Manski (1975). The formal statement of consistency as $n \to \infty$ follows.

Theorem 1. Under Assumptions 1, 3, and 4, the pairwise maximum-score estimator $\hat{\beta}_n^K \in \Theta$ is a consistent estimator for β^0, the true parameter in the data-generating process.

The proof of Theorem 1 is in the Appendix. The proof is written to be rather detailed because applied users of demand estimation techniques may not be familiar with proof techniques from the semiparametric discrete-choice literature.

If the covariates do not satisfy Assumption 3, β^0 can be bounded by finding the set of β^0 that satisfy Assumption 1.

□ **Comparison of pairwise maximum score to Manski's multinomial maximum score.** Manski (1975) introduces a general class of scoring rules. It is easy to see that pairwise maximum score produces the same objective-function value as the objective function

$$Q_n^{\text{Rank}}(\beta) = \sum_{i=1}^{K} \frac{1}{n} \sum_{a=1}^{n} \left(1[ai] \cdot \text{Rank}\left[x_{ai}'\beta \mid \{x_{aj}'\beta\}_{j=1}^{K}\right]\right),$$

where $\text{Rank}[x_{ai}'\beta \mid \{x_{aj}'\beta\}_{j=1}^{K}]$ is the number of other choices in K that have a deterministic payoff $x_{aj}'\beta$ less than $x_{ai}'\beta$. $\text{Rank}[x_{ai}'\beta \mid \{x_{aj}'\beta\}_{j=1}^{K}]$ is a function of the characteristics of all K choices. If $K = J$, $Q_n^{\text{Rank}}(\beta)$ is a member of the general class of estimators in Manski (1975).

The form of the multinomial maximum score used by Manski (1975) in his Monte Carlo study, and reported in the textbook of Amemiya (1985), is

$$Q_n^{\text{Max}}(\beta) = \sum_{i=1}^{J} \frac{1}{n} \sum_{a=1}^{n} (1[ai] \cdot 1[x'_{ai}\beta > x'_{aj}\beta \,\forall\, j \in J, j \neq i]).$$

The objective function $Q_n^{\text{Max}}(\beta)$ increases the score of correct predictions if an observed selection i has a greater linear index $x'_{ai}\beta$ than the linear indices of all J choices other than i. Although previous authors have not discovered this property, if J were replaced with $K \leq J$ in $Q_n^{\text{Max}}(\beta)$, the objective function produces a consistent estimator under the same conditions as Theorem 1.

There are three advantages of the pairwise maximum-score estimator formed by maximizing $Q_n^K(\beta)$ over one or both of $Q_n^{\text{Rank}}(\beta)$ and $Q_n^{\text{Max}}(\beta)$. First, $Q_n^{\text{Max}}(\beta)$ does not incorporate all the restrictions of Assumption 1. If J or K is 100, the presence of 100 error terms means that the choice with the highest $x'_{ai}\beta$ will often not be picked. Say $x'_{ai}\beta$ for the observed choice is the 50th highest deterministic payoff. $Q_n^{\text{Max}}(\beta)$ would increase by 0 for observation a. Assumption 1 says that the probability of choosing i must be higher than 49 other probabilities. The pairwise maximum-score objective function would increase the score of correct predictions by 49 when evaluated at the true parameter value β^0, and makes better use of the data from observation a.

A second, programming advantage of pairwise maximum score over $Q_n^{\text{Rank}}(\beta)$ is that a researcher does not need to code a sorting algorithm to compute $\text{Rank}[x'_{ai}\beta \mid \{x'_{aj}\beta\}_{j=1}^{K}]$. The evaluation of the pairwise maximum-score objective function requires only addition, multiplication, and pairwise comparison.

The third advantage of pairwise maximum score over both $Q_n^{\text{Rank}}(\beta)$ and $Q_n^{\text{Max}}(\beta)$ involves the rate of convergence and asymptotic distribution of maximum-score estimators. For the binary-choice maximum-score estimator, Horowitz (1992) proves that, under additional smoothness assumptions about the data-generating process, his smoothed estimator converges at a rate close to $\sqrt{n}$ and is asymptotically normal with a variance-covariance matrix that can be estimated and used for inference. Effective smoothing requires a one-dimensional function of the unknown parameters, β. Pairwise maximum score includes only one-dimensional functions that are easy to smooth, whereas $Q_n^{\text{Rank}}(\beta)$ and $Q_n^{\text{Max}}(\beta)$ are multidimensional functions of β that will have slower rates of convergence when smoothed.

Smoothing the maximum-score step function does not solve the main issue in the computational cost of numerically maximizing the objective function: the presence of local hills providing tempting regions for a greedy optimization routine to converge to.

5. Implementation suggestions

■ **Programming the objective function for a given data set.** If an otherwise intractable number of choices are observed in the data, a feasible estimator splits all choices into nests. For example, if K is 1 million and there are $n = 1000$ observations on agents, then the researcher could create 100 nests, each with 10 selections made by agents in the data and 10 unobserved choices. Say the researcher has chosen D distinct nests of choices $K_1, \ldots, K_D$. In notation appropriate for understanding the asymptotic properties of the estimator, the new pairwise maximum-score objective function is

$$Q_n^{K,D}(\beta) = \sum_{d=1}^{D} \sum_{i \in K_d} \sum_{j \in K_d, j > i} \frac{1}{n} \sum_{a=1}^{n} \{1[ai] \cdot 1[x'_{ai}\beta > x'_{aj}\beta] + 1[aj] \cdot 1[x'_{aj}\beta > x'_{ai}\beta]\}.$$

Bajari, Fox, and Ryan (2006) show that this objective function preserves consistency under agent and nest fixed effects. If the econometrician only compares two choices in the same nest, an unobserved nest and agent-specific fixed effect drops out. If nests are included for computational reasons, allowing for agent and nest fixed effects comes for free.

For any given data set, many of the dependent variable indicator functions of the form $1[ai]$ will be zero. Define $K(a)$ to be a nest that includes agent a's observed selection, y_a. For a, the

estimator will only relate the deterministic linear index of y_a to the linear indices of other options in $K(a)$. For a given data set, the objective function with nests reduces to

$$Q_n^{K,D}(\beta) = \frac{1}{n}\sum_{a=1}^{n}\sum_{j\in K(a), j\neq y_a} 1[x'_{ay_a}\beta > x'_{aj}\beta].$$

This is an easy objective function to program. Evaluating the objective function requires only multiplication, addition, and pairwise comparison.

□ **Numerical maximization of the objective function.** The maximum-score objective function is a step function, so there will always be multiple global maxima. The idea underlying consistency is that with a large number of observations, there is likely to be only one connected range of maxima. In the limit, this connected range converges to a point, if the assumptions for point identification and consistency specified in Theorem 1 are satisfied.

A numerical search procedure must be employed in order to compute one global maximum. As every point on a step function is a local maximum, local search methods are not useful. A researcher should use a global search routine. In terms of speed conditional on finding a global maximum at all, I have found good results using the method of differential evolution developed by Storn and Price (1997). Storn and Price discuss how differential evolution beats many other optimization routines, include simulated annealing, on a suite of test problems commonly used as benchmarks in the numerical optimization literature. The objective functions in the Monte Carlo experiments in Section 6 are almost always properly maximized by the differential evolution routine in Mathematica, using its default settings. The differential evolution procedure is an option of the NMinimize command.

□ **Inference.** Neither estimating the components of an analytically derived asymptotic distribution nor using the bootstrap are currently useful methods for inference for maximum-score estimators. Kim and Pollard (1990) show that the binary-choice maximum-score estimator converges at the rate of $\sqrt[3]{n}$, and Abrevaya and Huang (2005) show that the standard bootstrap is not consistent in the case of the class of $\sqrt[3]{n}$-consistent estimators studied by Kim and Pollard.

Delgado, Rodríguez-Poo, and Wolf (2001) show that an alternative resampling procedure, subsampling, consistently estimates the asymptotic distribution of test statistics for the class of $\sqrt[3]{n}$-consistent estimators studied by Kim and Pollard. Subsampling is described in the book by Politis, Romano, and Wolf (1999). I have used subsampling in my own empirical applications (Bajari and Fox, 2007).

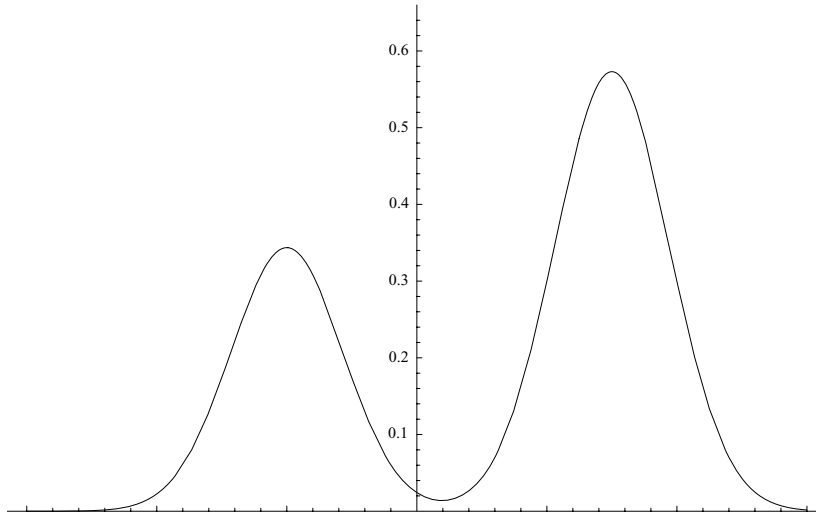
□ **Price endogeneity and instruments.** Say a researcher includes an endogenous covariate, say price, and has a vector of instruments z for price. Normalize the coefficient on price to be -1. The first stage of a two-stage instrumental variables estimator runs OLS to produce the estimate $\hat{\gamma}_n$ for the parameters of the pricing prediction equation. Then $z'_i\hat{\gamma}_n$ is substituted for price in the discrete-choice objective function, and the maximum-score estimator is computed normally. The two-stage IV pairwise maximum-score estimator is any parameter vector that maximizes the objective function

$$\begin{aligned} Q_n^{K,\mathrm{IV}}(\beta) = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{1}{n}\sum_{a=1}^{n}\{&1[ai]\cdot 1[x'_{ai}\beta - z'_{ai}\hat{\gamma}_n > x'_{aj}\beta - z'_{aj}\hat{\gamma}_n] \\ &+ 1[aj]\cdot 1[x'_{aj}\beta - z'_{aj}\hat{\gamma}_n > x'_{ai}\beta - z'_{ai}\hat{\gamma}_n]\}. \end{aligned}$$

This estimator will be consistent if the true residual from the price regression is i.i.d. across observations. In this case, the sum of the (itself assumed i.i.d.) random utility error and the pricing residual is i.i.d., giving the rank-order property needed for identification.

FIGURE 1

MIXED NORMAL DENSITY FUNCTION USED IN MONTE CARLO EXPERIMENTS



The density is a mixed normal: $0.369 \cdot N(-1.0, 0.184) + (1 - 0.369) \cdot N(1.5, 0.193)$. These parameters are chosen to create an asymmetric, bimodal distribution with the same mean and variance as the extreme value distribution assumed by the logit.

6. Monte Carlo experiments

■ This section presents Monte Carlo experiments to study the finite-sample properties of the subset maximum-score estimator. For an observation a, data are generated from the random-utility model

$$u_{ai} = x_{1,ai} + \beta_2 x_{2,ai} + \varepsilon_{ai}, \quad \text{for } i = 1, \ldots, J,$$

where ε_{ai} is i.i.d. across observations and choices. The covariates $x_{1,ai}$ and $x_{2,ai}$ are i.i.d. $N(0, 2)$. The true parameter value is always $\beta_2 = 1$. As choice behavior alone cannot identify the cardinalization of utility functions, discrete-choice estimators require location and scale normalizations. I impose the normalizations that the mean of ε_{ai} is equal to Euler's constant ($\gamma \approx 0.577$), the variance of ε_{ai} is equal to the logit variance ($\pi^2/6 \approx 1.65$), and the coefficient on $x_{1,ai}$ is 1. The first two normalizations correspond to those made in the logit maximum-likelihood estimator. The coefficient normalization is more common in semiparametric estimators. A semiparametric estimator can estimate the sign on $x_{1,ai}$ (whether the coefficient is 1 or -1), although the estimate of the sign would converge at such a fast rate that further analysis of its finite-sample properties is unhelpful.

The distribution of the error term ε_{ai} is mixed normal, which in this case is a bimodal, asymmetric distribution. Figure 1 displays a plot of the probability density function, and gives details of the exact parameters chosen. Again, the parameters are chosen to match the mean and variance of the extreme value distribution in the logit.[2] The mixed normal density is chosen specifically so the logit estimator will be misspecified, to highlight the greater robustness of semiparametric estimation.

[2] An alternative normalization is to make the median (not the mean) of ε_{ai} equal to its value for the logit ($-\log\log 2 \approx 0.367$) while preserving the variance of $\pi^2/6$. Under one such mixed normal satisfying these alternative normalizations, the finite-sample bias of the logit maximum-likelihood estimator is 0.01 farther from 0.0 than the logit ML bias reported in the first row ($N = 100$, $J = 10$) of Table 1.

TABLE 1 Monte Carlo Calculations Full-Choice Sets and Estimation Subsets

	# of True Choices	Logit Maximum Likelihood			Pairwise Maximum Score ($K = J$)			# of Est. Choices	Logit Sampling			Subset Pairwise Maximum Score		
N	(J)	Bias	MSE	Time	Bias	MSE	Time	(K)	Bias	MSE	Time	Bias	MSE	Time
100	10	.072	.014	.035	.006	.019	1.09	5	.089	.021	.018	.008	.039	.480
	100	.178	.036	.314	-	-	-	10	.178	.046	.039	.025	.046	1.16
	1000	.257	.067	3.25	-	-	-	10	.198	.072	.043	.084	.142	1.27
500	10	.071	.007	.207	.004	.006	6.85	5	.082	.009	.121	.009	.010	2.93
	100	.178	.033	1.72	-	-	-	10	.169	.031	.219	.008	.013	6.76
	1000	.253	.065	19.3	-	-	-	10	.151	.029	.233	.015	.030	7.91

The true model is i.i.d. multinomial mixed normal. One thousand replications are performed for all experiments. The reported time is the mean number of seconds to perform one replication for that estimator. For each replication, the four estimators use the same fake data. The logit estimates are computed using a gradient-based maximization routine. The maximum-score estimates are computed using a genetic algorithm because the objective function is a step function. Computations are done in Mathematica 5 for (32-bit) Windows on a machine with a 2.2 GHz Athlon 64 CPU.

Table 1 reports the finite-sample mean biases, mean-squared errors (MSE), and mean execution times from the Monte Carlo experiments. In all experiments, only the parameter β_2 is estimated. The bias of an estimator $\hat{\beta}_2$ is $E[\hat{\beta}_2 - \beta_2]$, and the mean-squared error is $E[(\hat{\beta}_2 - \beta_2)^2]$. The first estimator in Table 1 is the logit maximum-likelihood estimator. The second estimator is the maximum-score estimator when all choices are included.

The table also includes the two estimators that are computationally feasible when the true number of choices is large. The logit sampling estimator was introduced by McFadden (1978). It exploits the independence from irrelevant alternatives property of the logit to allow estimation using agent-specific random choice sets. The logit sampling estimator is compared to the pairwise maximum-score estimator, where observations are grouped into smaller nests.

For the maximum-score estimator, I use all observations by creating distinct estimation nests, as in Section 5. The randomly sampled choice set (logit) or nest (maximum score) sizes (K_d) are equal to 5 or 10, as marked in Table 1. In the sampling logit, the choice set for an agent is its observed choice plus 4 or 9 other choices, randomly sampled without replacement from all possible choices.[3] The other choices are included with uniform probabilities. In the subset maximum score with nests, nests are created from the total set of observed (in the data) choices. Two ($K_d = 5$) or five ($K_d = 10$) of the choices in each nest are sampled without replacement from choices seen in the data. The total number of nests is the number of subsets formed by this partition. Some observed selections are unassigned if the number of unique choices in the data is not evenly partitionable into nests. I assign the remainder of the observed selections to all nests. The remaining choices (to make 5 or 10) in a nest are randomly sampled without replacement from all choices not already assigned to that nest. Estimation proceeds by assigning each agent the nest that includes the agent's observed selection. To break ties, an agent with a selection in more than one nest is assigned the largest arbitrary nest index.

When examining the execution times, it is clear that both the logit sampling estimator and the subset maximum-score estimator require less computer time for estimation than methods that require itemization of the entire choice set.[4] In fact, performing 1000 replications of pairwise

[3] The distinction between random and deterministic choice sets should not affect the distribution of included covariates in this Monte Carlo study, and thus should not affect the comparison of the estimators.

[4] The Monte Carlo experiments are not designed to compare the execution time of the logit and maximum score. In practice, the logit requires the estimation of one more parameter than the maximum-score estimator, as semiparametric estimators make location and scale normalizations on the parameter space, rather than the distribution of the unobservables. As optimization routines suffer from a curse of dimensionality in the number of parameters, it is possible that the logit might take longer than the maximum score. The maximum-score estimator should be computed twice, once for the first parameter $\beta_1 = 1$, and again for $\beta_1 = -1$. The set of estimates corresponding to the minimum of the two sets of

maximum score with the full choice set is burdensome enough that I do not investigate its small-sample properties when the number of choices is greater than 10. For all estimators, and for a given number of observations, execution time is roughly equal to a constant times the number of choices considered in estimation for each observation. The logit without choice sampling estimator takes 19 seconds with 500 observations and 1000 choices. By extrapolation, a model where there are one million choices and only one parameter might take five hours. Optimization routines suffer from a curse of dimensionality in the number of parameters, so a multivariate logit model with one million choices and five covariates might take days to estimate.

Table 1 shows that the maximum-score estimators have lower levels of finite-sample bias than the logit estimators. For example, in the first row ($N = 100$, $J = 100$), the logit sampling estimator has a bias of 0.09, while the bias of the semiparametric estimator is an order of magnitude lower, at 0.01. The most likely explanation is that the semiparametric estimator is consistent under the mixed normal distribution, whereas the logit is inconsistent. The mean-squared errors of the logit are sometimes smaller (0.021 versus 0.039 in the first row), probably because the logit assumes additional structure that makes continuous choice-probability predictions, rather than just discrete yes-or-no predictions about choices in the maximum-score case. However, any additional precision of the logit does not contribute much to accuracy, as the logit is precisely estimating a biased coefficient.

7. Summary

■ A researcher may have data on a subset of choices, or a researcher may only wish to use data on a few choices because itemizing all choices is computationally burdensome. An important difficulty in estimating parametric multinomial choice models is that maximum likelihood is typically inconsistent when estimation uses covariate data on a subset of choices. An exception is that pseudo-maximum-likelihood estimators using the class of block-additive GEV errors studied by McFadden (1978) and Bierlaire, Bolduc, and McFadden (2006) are consistent using data on a subset of choices.

This article shows that any multinomial maximum-score estimator can consistently estimate the parameters in a linear index over observable characteristics using data on a subset of choices. A maximum-score estimator is semiparametric and does not require assuming that the econometrician knows the exact shape of the error distribution. There has been almost no work on multinomial maximum score since the original paper by Manski (1975). I modernize the conditions on error terms and covariates needed for consistency. I improve the statistical properties of multinomial maximum-score estimators by introducing a pairwise maximum-score estimator that emphasizes comparisons between two choices at a time. The pairwise maximum-score estimator uses more of the restrictions from the rank-order property underlying consistency than some earlier specifications. The estimator does not require sorting algorithms. Also, the estimator includes only one-dimensional functions of the unknown parameters and so is amenable to the smoothing procedure of Horowitz (1992).

Identification and consistency rely on the property that choice probabilities are rank ordered by the deterministic payoffs. I show that rank ordering holds for a subset of choices. A sufficient condition for the rank-order property is exchangeability in the density of the errors. Exchangeability places strong restrictions on the form of correlation and heteroskedasticity across choices for a given agent, but not across agents. To weaken the importance of the restrictions on the error terms, Bajari, Fox, and Ryan (2006) show how to estimate in the presence of agent-specific fixed effects over nests of alternatives.

My own personal experience and Monte Carlo studies show multinomial maximum-score estimation is a tractable and useful procedure for applied work. I hope other researchers consider

estimates should be kept. However, there is no curse of dimensionality in executing an optimization routine multiple times. The execution time of both programs would be improved by coding them in C or Fortran. Further, hardware-specific optimizations to the exp function and various logit-specific optimizations will speed up the logit.

using maximum score to reduce the computational speed of estimation, to handle missing data on choices, and to make identification more robust and transparent.

Appendix

■ The proof of Theorem 1 follows.

Proof of Theorem 1. The proof of consistency involves verifying the conditions of the consistency Theorem 2.1 in Newey and McFadden (1994). The theorem has four conditions:

(i) The probability limit of the maximum-score objective function, $Q_n^K(\beta)$, has a unique global maximum at the true parameter vector, β^0 (constructive identification).
(ii) The parameter space Θ is compact. This is Assumption 4.
(iii) The probability limit of the objective function, $Q_\infty^K(\beta)$, is continuous in β.
(iv) The objective function converges uniformly in probability to its limit.

Newey and McFadden treat binary-choice maximum score as an important example of their general consistency theorem, and go over the regularity conditions needed in some detail. The subset maximum-score objective function is similar.

□ **Constructive identification.** This section will focus on proving Condition (i), that the probability limit of $Q_\infty^K(\beta)$ has a unique global maximum at β^0. This is also a constructive proof of semiparametric identification.

The probability limit of the objective function. To preview how to compute a probability limit of the objective function, I will first compute the probability limit of an important term. Recall that the underlying population random variables generating the data are the vector x of covariate data for all J choices, and the vector ε of error terms for all J choices. In some applications, the number of choices J is itself random, which is fine in what follows. Recall that I am conditioning on the subset of choices K, so the n observations come from agents who select choices in K. By a law of large numbers and the law of iterated expectations,

$$\operatorname{plim}\left(\frac{1}{n}\sum_{a=1}^{n} 1[ai]\right) = E_{x,\varepsilon\,|\,K}\{1[i \text{ optimal}]\,|\,\beta^0\} = E_{x\,|\,K}\{E_{\varepsilon\,|\,K}\{1[i \text{ optimal}]\,|\,x,\beta^0\}\}$$
$$= E_{x\,|\,K}\left\{P_K(i\,|\,x,J,\beta^0)\right\}.$$

The last equality uses the decision rule for a selection, (1), and the notion of a conditional (on K) choice probability. The final expectation is taken over the conditional on K distribution of all J x's. Conditioning on K is also important for the inner expectation over the ε's.

I apply the same tools to the task of finding the probability limit of the maximum-score objective function. Again by a law of large numbers and the law of iterated expectations (over the ε's and the x's), the probability limit of $Q_n^K(\beta)$ is

$$Q_\infty^K(\beta) = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} E_{x,\varepsilon\,|\,K}\left\{1[i \text{ optimal}]\cdot 1[x_i'\beta > x_j'\beta] + 1[j \text{ optimal}]\cdot 1[x_j'\beta > x_i'\beta]\,|\,\beta^0\right\}$$
$$= \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} E_{x\,|\,K}\left\{P_K(i\,|\,x,J,\beta^0)\cdot 1[x_i'\beta > x_j'\beta] + P_K(j\,|\,x,J,\beta^0)\cdot 1[x_j'\beta > x_i'\beta]\right\}.$$

The above derivation uses the fact that the indicator functions of the form $1[x_i'\beta > x_j'\beta]$ do not depend on the ε's, and can be factored out of the expectation over the ε's, temporarily holding the x's constant by the law of iterated expectations.

The true parameter vector β^0 is a global maximum. I prove that $Q_\infty^K(\beta)$ has a global maximum at β^0 by showing the result for a given x, the covariate data on all J choices (even though not all of the choices are observed to the econometrician). If the integrand (the portion of the objective function inside the expected value operator) is maximized at β^0 for an arbitrary point x, the integral (the expectation over the observable covariates x) itself must be maximized at β^0.

First consider the case where there is a unique maximum among the two choice probabilities $P_K(i\,|\,x,\,J,\,\beta^0)$ and $P_K(j\,|\,x,\,J,\,\beta^0)$. Note that the inequalities in the two indicator functions $1[x_i'\beta > x_j'\beta]$ and $1[x_j'\beta > x_i'\beta]$ are mutually exclusive. Only one of the two indicators may be equal to 1 for a given β and x. So, for x and a trial β, one of the two probabilities of the form $P_K(i\,|\,x,\,J,\,\beta^0)$ will multiply 1, and the other will multiply 0. If there is a parameter value that always has the maximum of two choice probabilities multiply the 1 and the minimum multiply the 0, that value is a global maximum. Lemma 1 states that the conditional (on K) choice probabilities are ordered by the deterministic payoffs. Therefore, Lemma 1 shows that β^0 will implement the assignment, and globally maximize $Q_\infty^K(\beta)$. In notation, when $x_i'\beta^0 > x_j'\beta^0$, we know the integrand of $Q_\infty^K(\beta)$ at point x is maximized by the value $P_K(i\,|\,x,J,\beta^0)\cdot 1 + P_K(j\,|\,x,J,\beta^0)\cdot 0$.

Now I show that the set of points where the maximum of the two choice probabilities is not unique,

$$P_K(i\,|\,x,J,\beta^0) = P_K(j\,|\,x,J,\beta^0),$$

has probability 0, and these points do not contribute positively to the probability limit of the objective function. Ties in choice probabilities happen when $x_i'\beta^0 = x_j'\beta^0$. I use Assumption 3, particularly the condition requiring at least one continuously and freely varying covariate $w = x_{1i} - x_{1j}$ for choices i and j, to prove that tied probabilities are a probability 0 event. To see this, define the random variable $z = \tilde{x}_i'\tilde{\beta} - \tilde{x}_j'\tilde{\beta}$, where for choice i, $\tilde{x}_i$ is the vector of the last $d-1$ covariates, in other words, the covariates other than $x_{1,i}$. Consider the parameter normalization where $\beta_1 = 1$. The case where $\beta_1 = -1$ is symmetric. The event $P_K(i \mid x, J, \beta^0) = P_K(j \mid x, J, \beta^0)$ happens with probability

$$\int_{-\infty}^{\infty}\int_{z}^{z} h(w \mid z)m(z)\,dw\,dz,$$

where m is the derivable density of z and h is the derivable density of w. Both h and m can be derived from the distribution G_K in Assumption 3. Notice that the upper and lower limits of the integrand are the same, so the event $z = w$ has 0 measure unless there is a point mass at $w = z$, which is ruled out by Assumption 3. Therefore, the set of tied choice probabilities has probability 0.

Also, the inequalities in $Q_\infty^K(\beta)$ are strict, so $\beta = 0$ is a global minimum to the objective function, not a global maximum.

The global maximum β^0 is unique. There is still the matter of showing that the global maximum β^0 is unique. The uniqueness argument follows standard semiparametric discrete-choice arguments about point identification, and again relies on the assumption of a freely varying continuous covariate for each pair of choices w_{ij}. Consider an alternative parameter vector in Θ, β^-, that for the sake of a proof by contradiction is a global maximum of $Q_\infty^K(\beta)$. If, for values of x with positive probability, β^- and β^0 give different values of the two inequalities of the form $1[x_i'\beta > x_j'\beta]$, then β^- will make choice probabilities that are not the maxima of the two choice probabilities enter $Q_\infty^K(\beta)$, and there will be a contradiction. In other words, $Q_\infty^K(\beta^-)$ will be lower than $Q_\infty^K(\beta^0)$, because at some points, for example, $P_K(j \mid x, J, \beta^0)$ will enter the objective function when $P_K(i \mid x, J, \beta^0)$ is larger.

The question is whether the set of points where β^0 and β^- make different predictions about the rank ordering of the two choice probabilities, from the point of view of Lemma 1, has positive measure. The set of points where β^0 and β^- make different predictions is

$$\begin{aligned} S(\beta^0, \beta^-) = &\left\{x \mid ((x_i' - x_j')\beta^0 < 0 < (x_i' - x_j')\beta^-)\right\} \\ &\cup \left\{x \mid ((x_i' - x_j')\beta^0 > 0 > (x_i' - x_j')\beta^-)\right\}. \end{aligned}$$

A tilde ($\sim$) refers to all the elements of a vector other than the first element. I will show the argument for $\beta_1 = +1$; the argument for $\beta_1 = -1$ is symmetric. Note that the normalization implies $\beta_1^0 = \beta_1^- = +1$, as the normalization holds for both the true and alternative parameter vectors. By separating out the first element of the covariates vector (with its parameter β_1 normalized to +1), and remembering that $w_{ij} = x_{1i} - x_{1j}$,

$$\begin{aligned} S(\beta^0, \beta^-) = &\left\{x \mid (\tilde{x}_i' - \tilde{x}_j')\tilde{\beta}^0 < -w_{ij} < (\tilde{x}_i' - \tilde{x}_j')\tilde{\beta}^-\right\} \\ &\cup \left\{x \mid (\tilde{x}_i' - \tilde{x}_j')\tilde{\beta}^0 > -w_{ij} > (\tilde{x}_i' - \tilde{x}_j')\tilde{\beta}^-\right\}. \end{aligned}$$

If $S(\beta^0, \beta^-)$ has a positive probability, then the alternative β^- will make incorrect predictions (relative to the true data-generating process) with positive probability, and therefore $Q_\infty^K(\beta^-)$ will be lower than $Q_\infty^K(\beta^-)$. Indeed, the set $S(\beta^0, \beta^-)$ has positive probability because, by Assumption 3 above, w_{ij} has continuous support and is freely varying, and thus has positive mass over any given interval, specified by the values $\tilde{x}_i$ and $\tilde{x}_j$ of the other covariates. The only other possibility is that

$$(\tilde{x}_i' - \tilde{x}_j')\tilde{\beta}^0 = (\tilde{x}_i' - \tilde{x}_j')\tilde{\beta}^-$$

for all $\tilde{x}_i$ and $\tilde{x}_j$, which is ruled out because Assumption 3 says the support of the covariates does not lie in a proper linear subspace of $\mathbb{R}^d$. These arguments show that uniqueness and point identification are proved.

Note that the above arguments do not rely on itemizing the entire choice set, or on working with probabilities or data on choices outside of K, although the expectations in the probability limits are over covariates and error terms for all J choices. Therefore, including covariate data on more than $K = 2$ choices is not required for consistency of a properly specified maximum-score estimator.

□ **Continuity of the limiting objective function and uniform convergence.** Conditions (iii) and (iv) are satisfied if the more primitive conditions of Newey and McFadden's (1994) Lemma 2.4 are satisfied. Lemma 2.4 can be used to prove continuity of $Q_\infty^K(\beta)$, Condition (iii), as well as uniform in probability convergence of $Q_n^K(\beta)$ to $Q_\infty^K(\beta)$, which is Condition (iv). Binary-choice maximum score is an important example in Newey and McFadden, so it is not surprising that Lemma 2.4's primitive conditions are satisfied for the subset multinomial maximum-score estimator as well.

The conditions of Lemma 2.4 are that the data across agents are i.i.d., that the objective function terms for each agent are continuous with probability 1 in β, and that the terms for each agent are bounded by a function whose mean is not infinite. The underlying population random variables generating the data are x and ε. The data are i.i.d. across agents.

The maximum-score objective function $Q_n^K(\beta)$ is not continuous in β because β enters indicator functions in $Q_n^K(\beta)$. Condition (iii) is that the probability limit of the objective function, $Q_\infty^K(\beta)$, is continuous in β. Lemma 2.4 only requires

that the objective-function terms for each agent are continuous with probability 1 in β. Although the terms for each agent are not continuous in β because of the indicator functions, they are continuous with probability 1 by the support condition on the covariates, Assumption 3. The points where the objective-function contribution for an agent is not continuous in β are the points where $x_i'\beta = x_j'\beta$. As the continuous covariate w_{ij} is freely varying conditional on the other covariates, $x_i'\beta = x_j'\beta$ with probability 0 by a previous argument.

The other condition we need to verify to apply Lemma 2.4 is that the agent-specific terms in $Q_n^K(\beta)$ are bounded by a function with a noninfinite mean. As each agent makes one choice, the maximum score of correct predictions that can be entered into the objective function is 1 for each agent and pair of choices. Therefore, each agent's contribution to the objective function is bounded by the function $K - 1$.

References

Abrevaya, J. "Rank Estimation of a Generalized Fixed-Effects Model." *Journal of Econometrics,* Vol. 95 (2000), pp. 1–23.

——— and Huang, J. "On the Bootstrap of the Maximum Score Estimator." *Econometrica,* Vol. 73 (2005), pp. 1175–1204.

Amemiya, T. *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press, 1985.

Bajari, P. and Fox, J.T. "Measuring the Efficiency of an FCC Spectrum Auction." Working Paper, University of Minnesota, 2007.

———, ———, and Ryan, S. "Evaluating Wireless Carrier Consolidation Using Semiparametric Demand Estimation." Working Paper, University of Minnesota, 2007.

Bayer, P., McMillan, R., and Reuben, K. "An Equilibrium Model of Sorting in an Urban Housing Market." Working Paper, Duke University, 2004.

Berry, S.T., Levinsohn, J., and Pakes, A. "Automobile Prices in Market Equilibrium." *Econometrica,* Vol. 63 (1995), pp. 841–890.

Bierlaire, M., Bolduc, D., and McFadden, D.L. "The Estimation of Generalized Extreme Value Models from Choice-Based Samples." Working Paper, Ecole Polyechnique Federale de Lausanne, 2006.

Briesch, R.A., Chintagunta, P.C., and Matzkin, R.L. "Semiparametric Estimation of Brand Choice Behavior." *Journal of the American Statistical Association,* Vol. 97 (2002), pp. 973–982.

Chevalier, J. and Goolsbee, A. "Price Competition Online: Amazon versus Barnes and Noble." *Quantitative Marketing and Economics,* Vol. 1 (2003), pp. 203–222.

Delgado, M.A., Rodríguez-Poo, J.M., and Wolf, M. "Subsampling Inference in Cube Root Asymptotics with an Application to Manski's Maximum Score Estimator." *Economics Letters,* Vol. 73 (2001), pp. 241–250.

Goeree, M.S. "Advertising in the US Personal Computer Industry." Working Paper, University of Southern California, 2005.

Goeree, J.K., Holt, C.A., and Palfrey, T.R. "Regular Quantal Response Equilibrium." Working Paper, California Institute of Technology, 2004.

Horowitz, J.L. "A Smoothed Maximum Score Estimator for the Binary Response Model." *Econometrica,* Vol. 60 (1992), pp. 505–551.

———. *Semiparametric Methods in Econometrics,* Vol. 131 of *Lecture Notes in Statistics*. New York: Springer, 1998.

Kim, J. and Pollard, D. "Cube Root Asymptotics." *Annals of Statistics,* Vol. 18 (1990), pp. 191–219.

Manski, C.F. "Maximum Score Estimation of the Stochastic Utility Model of Choice." *Journal of Econometrics,* Vol. 3 (1975), pp. 205–228.

———. "Semiparametric Analysis of Discrete Response." *Journal of Econometrics,* Vol. 27 (1985), pp. 313–333.

———. "Identification of Binary Response Models." *Journal of the American Statistical Association,* Vol. 83 (1988), pp. 729–738.

Matzkin, R.L. "Nonparametric Identification and Estimation of Polychotomous Choice Models." *Journal of Econometrics,* Vol. 58 (1993), pp. 137–168.

McFadden, D.L. "Modelling the Choice of Residential Location." In A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds., *Spatial Interaction Theory and Planning Models,* Vol. I. Amsterdam: North Holland, 1978.

Mehta, N., Rajiv, S., and Srinivasan, K. "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation." *Marketing Science,* Vol. 22 (2003), pp. 58–84.

Newey, W.K. and McFadden, D. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics,* ed. Robert F. Engle and Daniel L. McFadden Vol. IV. Amsterdam: Elsevier, 1994.

Politis, D.N., Romano, J.P., and Wolf, M. *Subsampling*. New York: Springer, 1999.

Storn, R. and Price, K. "Differential Evolution—A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces." *Journal of Global Optimization,* Vol. 115 (1997), pp. 341–359.

Train, K.E., McFadden, D.L., and Ben-Akiva, M. "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices." *RAND Journal of Economics,* Vol. 18 (1987), pp. 109–123.