Discussion Papers No. 222, July 1998
Statistics Norway, Research Department

John K. Dagsvik

Nonparametric Identification of Discrete Choice Models

Abstract:
In this paper we give simple proofs of identification results in discrete choice models for the case where neither the deterministic part nor the distribution function of the random parts of the utility function is specified parametrically. The regularity conditions imposed are standard, but differ from conditions applied by other researchers, such as Matzkin (1992, 1993).

Keywords: Nonparametric identification, Discrete choice, Random utility models.

JEL classification: C14, C25

Address: John K. Dagsvik, Statistics Norway, Research Department. E-mail: jda@ssb.no

Discussion Papers	comprises research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborated than a usual article by including intermediate calculation and background material etc.

Abstracts with downloadable PDF files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
N-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1. Introduction

As is wellknown, the discrete choice theory was originally developed by psychologists, among others by Thurstone (1927) and Luce (1959), but it is fair to say that it was the seminal work by McFadden (1973) which, to a large extent, made economists aware of the potential of this theory for analyzing a large number of choice settings, such as choice among consumer brands, labor force participation, choice of mode of transportation, etc.

In the last decade the issue of identification of discrete choice models has been addressed by many researchers, cf. Manski (1988), Thompson (1989), Strauss (1979), Yellott (1977) and, most notably, Matzkin (1992, 1993). For example, Yellott (1977) proved that under fairly weak regularity conditions the distribution of the random terms in an additive random utility model are identified apart from a scale and location transform, provided the random terms are i.i.d. Matzkin is particularly concerned with nonparametric identification under different types of regularity conditions. In this paper we provide simple proofs related to identification under fairly general regularity conditions. The regularity conditions we apply are different from the ones applied by Matzkin. It is believed, however, that the assumptions presented in this paper are standard ones which are easy to interpret. Furthermore, under the assumptions maintained in this paper, the results on identification (or rather the "degree" of identification) are rather simple to prove.

2. Identification in discrete choice models

Discrete choice models apply in the context of analyzing behavior when individual agents make choices from finite sets of mutually exclusive alternatives. To describe a standard modelling framework, let S be the universe of m discrete alternatives, let $\Im$ be the collection of subsets from S and let B, $B \in \Im$, be the agent's choice set i.e., the set of alternatives that are feasible to him. Let U_j be the utility assigned to alternative $j \in S$. Since the alternatives are mutually exclusive the agent will choose alternative j if $U_j = \max_{k \in B} U_k$. Alternative j is described by a vector of observable attributes $z_j = (z_{j1}, z_{j2}, \ldots, z_{jK})$. Let $\mathbf{z} = (z_1, z_2, \ldots, z_m)$. To the observing econometrician the utilities, $\mathbf{U} = (U_1, U_2, \ldots, U_m)$, are perceived as random variables. The corresponding model therefore takes the form of a probabilistic choice model, represented by choice probabilities

$$P\left(U_j = \max_{k \in B} U_k \mid \mathbf{z}\right)$$

for $j \in B$, where the expression above is to be interpreted as the probability that the agent will choose j from B, given the attributes **z**. The choice models generated by random utilities are as described above called *random utility models*.

Let us next define identification formally. Let $M_m(D)$ be the family of m-dimensional cumulative distribution functions on D (say) where D is equal to R^m, or a subset of R^m. We say that two distribution functions on D are different if they attain different values on a subset of D with positive Lebesgue measure.

Definition

Let U and $\widetilde{U}$ *be two random utility models with respective conditional c.d.f.* $F(u|z)$ *and* $\widetilde{F}(u|z)$*, that belong to* $M_m(D)$*,* $z \in \boldsymbol{Z} \subset R^{mK}$*. If for all* $j \in B \subset \mathfrak{I}$*,*

$$P\left(U_j = max_{k \in B}\, U_k \mid z\right) = P\left(\widetilde{U}_j = max_{k \in B}\, \widetilde{U}_k \mid z\right),$$

implies that $F(\cdot \mid z) = \widetilde{F}(\cdot \mid z)$ *for* $z \in \boldsymbol{Z}^*$ *where* $\boldsymbol{Z}^*$ *is a subset of* $\boldsymbol{Z}$ *with positive measure, then* $M_m(D)$ *is identified.*

Let V be the family of real-valued continuous functions with domain $T \subset R^K$. Let $M_m^*(D)$ be the family of m-dimensional continuous c.d.f. on D with the property that the corresponding marginal distributions of each component, given the remaining components, are strictly increasing.

Assumption 1

The utility function has the structure

$$U_j = v(z_j) + \varepsilon_j$$

where $v \in V$ *and* $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$ *is a zero mean random vector with distribution function that is independent of* z*,* $z \in R^{mK}$.

Thus, under Assumption 1 the joint c.d.f. of the utilities is determined by the joint distribution of $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$ and the function v.

Assumption 2

For $\mathrm{v} \in V$ *and for any* $x \in R$ *there exists a* $z \in T$ *such that* $\mathrm{v}(z) = x$.

Theorem 1

Let $S = \{1,2\}$, (U_1, U_2) *and* $(\tilde{U}_1, \tilde{U}_2)$ *be two sets of random utilities,* $U_j = \mathrm{v}(z_j) + \varepsilon_j$ *and* $\tilde{U}_j = \tilde{\mathrm{v}}(z_j) + \tilde{\varepsilon}_j$, *such that Assumptions 1 and 2 are satisfied. Assume furthermore that the c.d.f. of* $(\varepsilon_1, \varepsilon_2)$ *and* $(\tilde{\varepsilon}_1, \tilde{\varepsilon}_2)$ *belong to* $M_2^*(R^2)$ *and* $\mathrm{v}, \tilde{\mathrm{v}} \in V$. *If*

(1) $$P(\mathrm{v}(z_1) + \varepsilon_1 > \mathrm{v}(z_2) + \varepsilon_2) = P(\tilde{\mathrm{v}}(z_1) + \tilde{\varepsilon}_1 > \tilde{\mathrm{v}}(z_2) + \tilde{\varepsilon}_2)$$

for $z_1, z_2 \in T$, *it follows that for some constant* $a > 0$, $\tilde{\mathrm{v}}(z) = a\,\mathrm{v}(z)$, $z \in T$, *and* $\tilde{\varepsilon}_1 - \tilde{\varepsilon}_2$ *has the same distribution as* $a(\varepsilon_1 - \varepsilon_2)$.

Proof:

Let G and $\tilde{G}$ be the cumulative distributions of $\varepsilon_2 - \varepsilon_1$ and $\tilde{\varepsilon}_2 - \tilde{\varepsilon}_1$, respectively. It follows that we can express (1) as

(2) $$G(\mathrm{v}(z_1) - \mathrm{v}(z_2)) = \tilde{G}(\tilde{\mathrm{v}}(z_1) - \tilde{\mathrm{v}}(z_2))$$

for $z \in T$. If $F_{12}(x_1, x_2)$ is the c.d.f. of $(\varepsilon_1, \varepsilon_2)$ it follows that

$$G(y) = \int_R F_{12}(dx, x + y),$$

and consequently

(3) $$G(y_2) - G(y_1) = \int_R \left(F_{12}(dx, x + y_2) - F_{12}(dx, x + y_1)\right)$$

for $y_1, y_2 \in R$. From Assumption 1 it follows that the integrand in (3) is positive when $y_2 > y_1$, which implies that G, and similarly $\tilde{G}$, are strictly increasing. Hence, G and $\tilde{G}$ are continuous and invertible.

Let $\varphi = \tilde{G}^{-1} G$. From (2) it follows, with $z_1, z_2, z_3 \in T$, that

(4) $$\tilde{\mathrm{v}}(z_1) - \tilde{\mathrm{v}}(z_2) = \varphi(\mathrm{v}(z_1) - \mathrm{v}(z_2)),$$

(5) $$\tilde{v}(z_1) - \tilde{v}(z_3) = \varphi\left(v(z_1) - v(z_3)\right),$$

and

(6) $$\tilde{v}(z_3) - \tilde{v}(z_2) = \varphi\left(v(z_3) - v(z_2)\right).$$

By combining (4), (5) and (6) we obtain

(7) $$\varphi\left(v(z_1) - v(z_2)\right) = \tilde{v}(z_1) - \tilde{v}(z_3) + \tilde{v}(z_3) - \tilde{v}(z_2) = \varphi\left(v(z_1) - v(z_3)\right) + \varphi\left(v(z_3) - v(z_2)\right).$$

Let $x = v(z_1)$, $y = v(z_3)$, and assume that z_2 is such that $v(z_2) = 0$. From (7) we get

(8) $$\varphi(x) = \varphi(x - y) + \varphi(y),$$

for $x, y \in R$, or equivalently

(9) $$\varphi(x + y) = \varphi(x) + \varphi(y),$$

for $x, y \in R$. In particular, with $x = y = 0$, we get

(10) $$\varphi(0) = 2\,\varphi(0)$$

so that $\varphi(0) = 0$. The only continuous solution of (8) with $\varphi(x) = 0$ is $\varphi(x) = ax$, for some positive a, cf. Aczél (1966). Since $\varphi(x) = G^{-1}(G(x))$, this implies that

$$G(x) = \tilde{G}(ax).$$

Since by assumption, $v(z_2) = 0$, it follows from (3) that $\tilde{v}(z) = av(z)$.

Q.E.D.

Let us now compare the result of Theorem 1 with Matzkin's results. Matzkin (1992) considers a more general setting in which the function v depends on the respective alternatives, and she therefore needs additional conditions to obtain identification compared to the conditions needed in this paper. Perhaps the result that is closest in spirit to the result obtained above is provided by Matzkin (1993), Theorem 1. However, in this theorem Matzkin considers a more general setting where the distribution of the random disturbances may depend on the attributes, and for that reason she needs stronger conditions than the ones needed above. Matzkin also allows the function v to be

dependent on individual characteristics. The results obtained in this paper can also be interpreted at the individual level, i.e., v can be interpreted as an agent specific function.

Above we have demonstrated that by considering the binary case we find that the structural term v(z) is identified apart from a scale transform. It remains to study identification of the joint distribution of the random terms of the utility function in the multinomial case. We can now prove the following result which is similar to the result of Theorem 1 in Strauss (1979), p.p. 39-40.

Analogous to (1) we wish to investigate the implications from

$$(11) \qquad P\left\{v\left(z_j\right)+\varepsilon_j=\max_{k\in B}\left(v\left(z_k\right)+\varepsilon_k\right)\right\}=P\left\{\tilde{v}\left(z_j\right)+\tilde{\varepsilon}_j=\max_{k\in B}\left(\tilde{v}\left(z_k\right)+\tilde{\varepsilon}_k\right)\right\},$$

for $j\in B\in\Im$, where $\left\{v\left(z_j\right)+\varepsilon_j, j\in S\right\}$ and $\left\{\tilde{v}\left(z_j\right)+\tilde{\varepsilon}_j, j\in S\right\}$ are two sets of random utilities.

We have the following result.

Theorem 2

Assume that (11) and Assumptions 1 and 2 hold. Assume furthermore that $\{1,2\}\in\Im$. *Then there is an* $a>0$ *such that*

$$(12) \qquad a\sum_{k=1}^{m} s_k\,\varepsilon_k \quad \text{and} \quad \sum_{k=1}^{m} s_k\,\tilde{\varepsilon}_k$$

have the same distribution for all $(s_1,s_2,\ldots,s_m)\in R^m$, *for which* $\sum_{k=1}^{m} s_k=0$.

Proof:

Observe that (10) is equivalent to

$$(13) \qquad P\left\{\bigcap_{k\in B}\left(\varepsilon_k-\varepsilon_j\le v\left(z_j\right)-v\left(z_k\right)\right)\right\}=P\left\{\bigcap_{k\in B}\left(\tilde{\varepsilon}_k-\tilde{\varepsilon}_j\le\tilde{v}\left(z_j\right)-\tilde{v}\left(z_k\right)\right)\right\}.$$

By letting $B=\{1,2\}$ it follows from Theorem 1 there exists a positive constant a such that $\tilde{v}(z)=av(z)$ for $z\in T$. Consequently, $\left(a\left(\varepsilon_1-\varepsilon_j\right),a\left(\varepsilon_2-\varepsilon_j\right),\ldots,a\left(\varepsilon_m-\varepsilon_j\right)\right)$ has the same distribution as $\left(\tilde{\varepsilon}_1-\tilde{\varepsilon}_j,\tilde{\varepsilon}_2-\tilde{\varepsilon}_j,\ldots,\tilde{\varepsilon}_m-\tilde{\varepsilon}_j\right)$. To realize this, recall that due to Assumption 2,

$\left(v(z_j)-v(z_1), v(z_j)-v(z_2),\ldots, v(z_j)-v(z_m)\right)$, which is $m-1$ dimensional, can take any value in R^{m-1}. But then these distributions must have the same characteristic functions, i.e.

$$(14)\qquad E\exp\left(iat\sum_{k=1}^{m}\lambda_k\left(\varepsilon_k-\varepsilon_j\right)\right)=E\exp\left(it\sum_{k=1}^{m}\lambda_k\left(\tilde{\varepsilon}_k-\tilde{\varepsilon}_j\right)\right)$$

for $t\in R$, which means that

$$a\sum_{k=1}^{m}\lambda_k\left(\varepsilon_k-\varepsilon_j\right)\ \text{and}\ \sum_{k=1}^{m}\lambda_k\left(\tilde{\varepsilon}_k-\tilde{\varepsilon}_j\right)$$

must have the same distribution for any $(\lambda_1,\lambda_2,\ldots,\lambda_m)\in R^m$. Provided the random terms are normalized to have zero mean this is equivalent to the statement that

$$a\sum_{k=1}^{m}s_k\,\varepsilon_k\ \text{and}\ \sum_{k=1}^{m}s_k\,\tilde{\varepsilon}_k$$

have the same distribution provided $\sum_{k=1}^{m}s_k=0$.

Q.E.D.

In the formulation of the discrete choice model above it is assumed that $\mathfrak{I}$ contains a binary choice set. We shall now consider the case when $\mathfrak{I}$ only contains the maximal set S.

Corollary 1

Assume that (11) and Assumptions 1 and 2 hold. Assume furthermore that $\mathfrak{I}=\{S\}$. Then the conclusion of Theorem 2 holds.

Proof:

Consider (13) with $B=S$ and $j=1$. Due to Assumption 2 it is possible to choose **z** such that $v(z_j)-v(x_k)=\tilde{v}(z_j)-\tilde{v}(x_k)=\infty$, for $k=3,4,\ldots,m$. In this case the model thus reduces to a binary choice model and by Theorem 1, $\tilde{v}(z)=av(z)$ for $z\in T$. The rest of the proof is the same as the last part of the proof of Theorem 2.

Q.E.D.

Corollary 2

Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$, $m \geq 3$, *are independent with nonvanishing characteristic functions, and similarly for* $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_m$. *Then, under the assumptions of Theorem 2, it follows that for each j, there is an* $a > 0$ *such that* $a\tilde{\varepsilon}_j + b$ *and* ε_j *have the same distribution, where* $b \in R$ *is arbitrary.*

The result of Corollary 2 has been proved by Strauss (1979). The next result is immediate.

Corollary 3

Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ *and* $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_m$, $m \geq 3$, *are i.i.d.. Then, under the assumptions of Theorem 2 it follows that for each j, there is an* $a > 0$ *such that* $a\tilde{\varepsilon}_j + b$ *and* ε_j *have the same distribution, where* $b \in R$ *is arbitrary.*

The result of Corollary 3 has been proved by Yellott (1977).

Example

In this example we apply the result of Theorem 2 to the multinomial probit model. In this case ε and $\tilde{\varepsilon}$ are multinormal random variables. Since the multinormal distribution is completely characterized by the mean and the covariance matrix it follows that (12) holds if and only if

$$a^2 \operatorname{Var}\left(\sum_{k=1}^{m} s_k \varepsilon_k\right) = \operatorname{Var}\left(\sum_{k=1}^{m} s_k \tilde{\varepsilon}_k\right)$$

for all $s_1, s_2, \ldots, s_m$, for which $\sum_{k=1}^{m} s_k = 0$. This implies that

$$a^2 \operatorname{Var}\left(\sum_{k=1}^{m} s_k \left(\varepsilon_k - \varepsilon_j\right)\right) = \operatorname{Var}\left(\sum_{k=1}^{m} s_k \left(\tilde{\varepsilon}_k - \tilde{\varepsilon}_j\right)\right) \tag{15}$$

for any $s_1, s_2, \ldots, s_m$. From (15) it follows that

$$a^2 \left(\sigma_{kr} - \sigma_{kj} - \sigma_{rj} + \sigma_{jj}\right) = \tilde{\sigma}_{kr} - \tilde{\sigma}_{kj} - \tilde{\sigma}_{rj} + \tilde{\sigma}_{jj} \equiv b_{krj} \tag{16}$$

where $\sigma_{kr} = E\,\varepsilon_k\,\varepsilon_r$ and $\tilde{\sigma}_{kr} = E\,\tilde{\varepsilon}_k\,\tilde{\varepsilon}_r$. It is easily verified that

$$2\,b_{krj} = b_{kkj} + b_{rrj} - b_{kkr}\,.$$

Hence, the restrictions implied by (16) can be represented by

$$a^2\left(\sigma_{kk}-2\sigma_{kj}+\sigma_{jj}\right)=b_{kkj} \tag{17}$$

for all $k \neq j$. For given $\left\{b_{kkj}\right\}$, (17) represents the necessary conditions for identification. Evidently, these restrictions are not sufficient. To achieve identification one may set $\sigma_{jj}=1$ for all j and $\sigma_{12}=0$. Then a and $\sigma_{kj},\ k \neq j,$ are determined by $2a^2=b_{221}$ and $2a^2\left(1-\sigma_{kj}\right)=b_{kkj}$.

3. Concluding remarks

It is difficult to interpret what the implication of (12) in Theorem 2 means in general. In other words, it would be interesting to obtain a characterization of the class of joint distribution functions of $\left(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_m\right)$ for which the characteristic function of $\left(\varepsilon_1-\varepsilon_j,\varepsilon_2-\varepsilon_j,\ldots,\varepsilon_m-\varepsilon_j\right)$ is given for each j, cf. (14). To this end it is intriguing that the GEV class is sufficiently large so that any random utility model can be approximated arbitrarily closely by a GEV model. This result was proved by Dagsvik (1995). Recall that the models in the GEV class (Generalized Extreme Value Model) are generated by utilities that are multivariate extreme value distributed, see McFadden (1981). The result is somewhat surprising since it is generally believed that the multivariate extreme value distributions are restrictive. In particular, the correlation between bivariate extreme value distributed variables is always non-negative.

Falmagne (1978) has considered identification of completely general random utility models, i.e., models which are derived from maximization of a general random utility index with no observed attributes associated with the alternatives. Falmagne has in fact established necessary and sufficient conditions that characterize random utility models. Furthermore, he has also considered the issue of how much information about the joint distribution of the utilities can be recovered from knowledge of the choice probabilities for every possible choice set. Since we only use utilities to rank order the alternatives it is clear that one can at most recover the ordinal structure of the original random utility variables. Falmagne demonstrates that although the complete ordinal structure cannot be recovered, a great deal of it can.

References

Dagsvik, J.K. (1995): How Large is the Class of Generalized Extreme Value Random Utility Models? *Journal of Mathematical Psychology*, **39**, 90-98.

Falmagne, J.C. (1978): A Representation Theorem for Finite Random Scale Systems. *Journal of Mathematical Psychology*, **18**, 52-72.

Luce, R.D. (1959): *Individual Choice Behavior*. Wiley, New York.

Manski, C. (1988): Identification of Binary Response Models. *Journal of the American Statistical Association*, **83**, 729-738.

Matzkin, R.L. (1992): Nonparametric and Distribution-free Estimation of the Threshold Crossing and Binary Choice Models. *Econometrica*, **60**, 239-270.

Matzkin, R.L. (1993): Nonparametric Identification and Estimation of Polychotomous Choice Models. *Journal of Econometrics*, **58**, 137-168.

McFadden, D. (1973): Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Eds.), *Frontiers in Econometrics*. Academic Press, New York.

Strauss, D. (1979): Some Results on Random Utility Models. *Journal of Mathematical Psychology*, **20**, 35-52.

Thompson, T.S. (1989): Identification of Semiparametric Discrete Choice Models. *Discussion Paper* no. 249. Center for Economic Research, University of Minnesota, Minneapolis, MN.

Thurstone, L.L. (1927): A Law of Comparative Judgment. *Psychological Review*, **34**, 273-286.

Yellott, J.I. (1977): The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment and the Double Exponential Distribution. *Journal of Mathematical Psychology*, **15**, 109-146.