TOPIC 15
TRAVEL CHOICE AND
DEMAND MODELLING

A HYBRID PROBABILISTIC CHOICE SET MODEL WITH COMPENSATORY AND NONCOMPENSATORY CHOICE RULES

TAKA MORIKAWA
Department of Civil Engineering
Nagoya University
Chikusa-ku, Nagoya, 464-01 JAPAN

Abstract

This paper proposes a discrete choice model with probabilistic choice sets that is computationally tractable even with a large number of alternatives. The choice set formation process is modelled by a random constraint model with non-compensatory nature. The model is applied to destination choice analysis of vacation trips.

INTRODUCTION

Discrete choice models have been successfully applied to various aspects of transport demand such as mode choice, destination choice, route choice, and car-ownership (Ben-Akiva and Lerman 1985). Most of the successful applications are for regular trips (eg commuting) in the choice contexts with a relatively small number of available alternatives. The most successful application, for instance, is the mode choice behavior of commuting with a few available modes.

However, there are many choice contexts in which the number of available alternatives is large and/or the choice set considered by the decision maker is uncertain to the analyst. Such choice contexts are typically found in the destination choice of non-regular trips such as vacation trips.

The choice context with uncertain choice sets can be modeled by the probabilistic choice set (PCS) models that explicitly consider uncertainty of individual choice sets. The most general PCS models consider all the combinations of potential choice sets. The number of combinations, however, increases exponentially with the increase in the number of alternatives. Therefore, it is practically impossible to estimate the general PCS models with a large number of alternatives (eg more than four). The Dogit model (Gaudry and Dagenais 1979) is a special form of the PCS models which requires a strong assumption on the possible choice sets.

The model proposed in this paper follows the basic PCS model paradigm in which the choice set formation model at the first stage and the discrete choice model at the second. The choice set formation model is a random constraint model that has non-compensatory nature among multiple constraints. More specifically, an alternative is included in the choice set if and only if all the latent conditioning measures of the alternative satisfy the criteria. The paper shows that this type of choice set formation process can be modeled by pairwise comparison of alternatives, which dramatically reduces the computational load of the PCS models. This enables one to apply the PCS models to the choice contexts with a large number of alternatives. The proposed method estimates the choice set formation model and the discrete choice model simultaneously using only the information of actual choices.

A case study of destination choice of vacation trips is presented. It shows that the proposed PCS model is better fitted to the data than the ordinary discrete choice model with deterministic choice sets (DCS models). It is concluded that choice contexts with a large number of alternatives in many cases require explicitly modeling the choice set formation process and that the proposed method is applicable and effective to such contexts.

PROBABILISTIC TWO STAGE CHOICE PARADIGM

Concepts of two stage models

The general form of the PCS models presented by Manski (1977) is as follows:

$$P_n(i) = \sum_{C \in G} P_n(i|C)Q_n(C|G) \tag{1}$$

where

$P_n(i)$ = probability of individual n choosing i ($i \in M$); M is the master set of alternatives;

$P_n(i|C)$ = probability of individual n choosing i given choice set C;

G = set of all non-empty subsets of M; and

$Q_n(C|G)$ = probability of individual n's choice set being C.

Conceptually the above model represents two stages of choice behavior. The first stage is the choice set formation process, of which model produces the probability $Q_n(C|G)$. The choice set termed here is the set of alternatives that the individual considers for choice in a particular situation. The second stage is the choice behavior given the choice set. This behavior is modeled by the discrete choice model that produces the choice probability $P_n(i|C)$.

The ordinary discrete choice analysis assumes that the analyst can find the true choice set for each individual by some deterministic rule. It implies in the above formulation:

$Q_n(C|G) = 1$ for a specific C_n; 0 otherwise

then, the equation (1) is reduced to:

$$P_n(i) = P_n(i|C_n). \tag{2}$$

Equation (2) is modeled by ordinary discrete choice models such as logit and probit models.

Typical deterministic rules to specify the individual choice set in the mode choice context are, for example, i) an individual owning no driver's license or no car does not have the "drive alone" alternative in the choice set, ii) an individual living further than half a mile from the nearest bus stop does not have the "bus" alternative, and so on.

Deterministic rules, however, are appropriate only for very limited contexts because the choice set is formed not only by objective physical constraints but also by informational and psychological restrictions. The informational restriction is particularly important when the master choice set consists of a large number of alternatives. Choices of destinations, routes, and residential locations are examples of such choice contexts where individuals are not likely to compare numerous alternatives by examining trade-offs among various attributes to search for the best alternative. When the deterministic rules are not applicable, the choice set will be probabilistic to the analyst and the model for the first stage should be formulated to explicitly evaluate $Q_n(C|G)$.

The process of choice set formation can be viewed as the process of examining the restrictions, or equivalently satisfying various constraints. In this case, it is natural to consider that multiple constraints do not "compensate" one another. More specifically, when one of the constraints is not satisfied for an alternative, that alternative cannot be included in the choice set even if the other constraints are completely satisfactory. The choice set formation, therefore, bears the non-compensatory nature among multiple constraints.

The second stage may be formulated by a discrete choice model with the ordinary compensatory utility function. In the compensatory utility function, an "unattractive" attribute is compensated by an "attractive" attribute.

A practical problem of the probabilistic choice set model (PCS model) formulated by equation (1) arises from the number of potential choice sets, or the elements of G. G is the set of all non-empty subsets of the master set. If there are three alternatives in the master set, then G consists of,

$G = \{ \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,3\}, \{1,2,3\} \}$

In general, if the master set consists of J alternatives, G has 2^J-1 elements (-1 represents exclusion of the empty set). It implies that the choice context with 10 potential alternatives requires the summation of 1,023 terms in equation (1). This seems the main reason that general PCS models have not been practically applied although they are theoretically attractive.

This paper develops an alternative derivation of a PCS model that alleviates the computational difficulty mentioned above.

Choice set formation model

Suppose that there are K independent constraints in the choice context of interest. Here, "independent" means K constraints have non-compensatory nature and statistical independence of each single constraint, as described below (Swait and Ben-Akiva 1987):

$$q_n(i) = \prod_{k=1}^{K} q_{kn}(i) \tag{3}$$

where $q_n(i)$ is the probability of alternative i being included in the choice set of individual n and $q_{kn}(i)$ is the probability of alternative i satisfying the k-th constraint for individual n.

Assume each constraint has the following conceptual structure: When a latent variable representing the desirability of the constraining condition exceeds a threshold value, then the constraint is satisfied. The latent variable can be expressed by,

$$E_{kn}(i) = \alpha_k' w_{kin} - \zeta_{kin} \quad k=1, \cdots, K \tag{4}$$

where

α_k = vector of unknown parameters;

w_{kin} = vector of variables affecting the restriction; and

ζ_{kin} = disturbance.

Denoting the threshold for the k-th constraint by μ_k, we can express $q_{kn}(i)$ by,

$$\begin{aligned} q_{kn}(i) &= \text{Prob}\left[E_{kn}(i) \geq \mu_k\right] \\ &= \text{Prob}\left[\alpha_k' w_{kin} - \zeta_{kin} \geq \mu_k\right] \\ &= \text{Prob}\left[\zeta_{kin} \leq \alpha_k' w_{kin} - \mu_k\right] \qquad k=1,\cdots,K \end{aligned} \tag{5}$$

If we assume ζ_{kin} to be logistically distributed, $q_{kn}(i)$ is given by,

$$q_{kn}(i) = \frac{1}{1+e^{-(\alpha_k' w_{kin} - \mu_k)}} \tag{6}$$

which is analogous to the binary logit model. From equations (3) and (6) we obtain,

$$q_n(i) = \prod_{k=1}^{K} \frac{1}{1+e^{-(\alpha_k' w_{kin} - \mu_k)}} \tag{7}$$

Then, the probability of individual n's choice set being C given the master set is expressed by,

$$Q_n C I G) = \frac{1}{1-Q_n(\varnothing)} \prod_{i \in M} \left[q_n(i)^{d_{iC}} \{1-q_n(i)\}^{1-d_{iC}}\right] \tag{8}$$

where

$Q_n(\varnothing)$ = probability of the random constraint model yielding the empty choice set; and

d_{iC} = 1 if alternative i is an element of choice set C; 0 otherwise

Choice probability of PCS model

The second choice stage is modeled by an ordinary discrete choice model with the compensatory utility function:

$$U_{in} = \beta' X_{in} + \varepsilon_{in} \equiv V_{in} + \varepsilon_{in} \tag{9}$$

where

β = vector of unknown parameters;

X_{in} = vector of explanatory variables; and

ε_{in} = disturbance.

Assuming the disturbances to be i.i.d. Gumbel, the multinomial logit model is derived to express the choice probability of the second stage:

$$P_n(i|C) = \frac{e^{V_{in}}}{\sum_{h\in C} e^{V_{hn}}} \tag{10}$$

Accordingly, the marginal choice probability of the two stage PCS model is expressed by substituting equations (8) and (10) into (1):

$$\begin{aligned} P_n(i) &= \sum_{C\in G} P_n(i|C)Q_n(C|G) \\ &= \frac{1}{1-Q_n(\emptyset)}\sum_{C\in G}\left[\frac{e^{V_{in}}}{\sum_{h\in C} e^{V_{hn}}}\prod_{j\in M}\left[q_n(j)^{d_{jC}}\{1-q_n(j)\}^{1-d_{jC}}\right]\right] \end{aligned} \tag{11}$$

If the number of alternatives in the master set is small enough (eg less than five), equation (11) may be directly evaluated to calculate the likelihood. However, when the number of alternatives increases, the number of possible choice set, or the number of elements in G, increases exponentially and the direct evaluation of equation (11) becomes virtually impossible.

ALTERNATIVE DERIVATION OF TWO STAGE PCS MODEL

Deriving choice probability

Ordinary discrete choice models can be derived by pairwise conditions of alternatives in terms of utility as shown below,

$$P(i) = \text{Prob}\left[U_i \geq U_1,\ U_i \geq U_2,\ \cdots, U_i \geq U_J\right] \tag{12}$$

Similarly, the PCS model can also be derived by pairwise comparison of alternatives. The situation of alternative i being *preferred* to alternative j, in this case, includes the following two possible cases: i) both alternatives i and j are included in the choice set and alternative i has a greater utility value than alternative j; or ii) alternative i is included in the choice set but alternative j is eliminated at the first stage of choice set formation. Equivalently, the marginal choice probability can be given in the following expression:

$$\begin{aligned} P_n(i) &= \frac{1}{1-Q_n(\emptyset)} \text{ x Prob}(i\in C_n) \text{x Prob}\left[\begin{array}{c} \{(1\in C_n)\cap(U_{in}\geq U_{1n})\}\cup\{1\notin C_n\} \\ \text{and} \\ \{(2\in C_n)\cap(U_{in}\geq U_{2n})\}\cup\{2\notin C_n\} \\ \text{and} \\ \vdots \\ \text{and} \\ \{(J\in C_n)\cap(U_{in}\geq U_{Jn})\}\cup\{J\notin C_n\} \end{array}\right] \\ &= \frac{1}{1-Q_n(\emptyset)} \text{ x } q_n(i) \text{x Prob}\left[\bigcap_{j\in M,\, j\neq i}\{(j\in C_n)\cap(\varepsilon_{jn}\leq U_{in}-U_{jn}+\varepsilon_{in})\}\cup\{j\notin C_n\}\right] \end{aligned} \tag{13}$$

where C_n is the latent true choice set of individual n.

Taking the conditional probability on the random component of utility, ε_{in}, equation (13) can be rewritten as:

$$P_n(i)=\frac{q_n(i)}{1-Q_n(\varnothing)}\int_{-\infty}^{+\infty} f(\varepsilon_{in}) \prod_{j\in M, j\neq i} \{q_n(j)F(V_{in}-V_{jn}+\varepsilon_{in})+(1-q_n(j))\}d\varepsilon_{in} \tag{14}$$

where f(•) and F(•) denote the PDF and CDF of ε's, respectively. If we assume the i.i.d. Gumbel ε's, as usual, we obtain,

$$P_n(i) = \frac{q_n(i)}{1-Q_n(\varnothing)}\int_{-\infty}^{+\infty} e^{-\varepsilon_{in}}\, e^{-e^{-\varepsilon_{in}}} \text{ x } \prod_{j\in M, j\neq I} \left[q_n(j)e^{-e^{-V_{in}+V_{jn}-\varepsilon_{in}}}+\{1-q_n(j)\}\right]d\varepsilon_{in} \tag{15}$$

$$= \frac{q_n(i)}{1-\prod_{j\in M}\{1-q_n(j)\}}\int_{-\infty}^{+\infty} e^{-\varepsilon_{in}}\, e^{-e^{-\varepsilon_{in}}} \text{ x } \prod_{j\in M, j\neq I} \left[q_n(j)e^{-e^{-V_{in}+V_{jn}-\varepsilon_{in}}}+\{1-q_n(j)\}\right]d\varepsilon_{in}$$

Although computation of equation (15) requires a single integration with respect to ε_{in}, it does not need to evaluate the choice probability for all the possible choice sets that sum up to 2^J-1. Therefore, the proposed method is applicable to the choice context with a large number of alternatives. To ensure that equation (15) is equivalent to equation (11), the case of three alternatives is demonstrated below:

$$P_n(1)$$

$$= \frac{q_n(1)}{1-Q_n(\varnothing)}\int_{-\infty}^{+\infty} f(\varepsilon_{1n})\{q_n(2)F(V_{1n}+V_{2n}+\varepsilon_{1n})+(1-q_n(2))\}\{q_n(3)F(V_{1n}-V_{3n}+\varepsilon_{1n})+(1-q_n(3))\}d\varepsilon_{1n}$$

$$= \frac{q_n(1)}{1-Q_n(\varnothing)}\Bigg[q_n(2)\, q_n(3)\int_{-\infty}^{+\infty} f(\varepsilon_{1n})F(V_{1n}-V_{2n}+\varepsilon_{1n})\, F(V_{1n}-V_{3n}+\varepsilon_{1n})d\varepsilon_{1n}$$

$$+q_n(2)\,(1-q_n(3))\int_{-\infty}^{+\infty} f(\varepsilon_{1n})F(V_{1n}-V_{2n}+\varepsilon_{1n})d\varepsilon_{1n}+(1-q_n(2))q_n(3)\int_{-\infty}^{+\infty} f(\varepsilon_{1n})F(V_{1n}-V_{3n}+\varepsilon_{1n})d\varepsilon_{1n}$$

$$+ (1-q_n(2))\,(1-q_n(3))\int_{-\infty}^{+\infty} f(\varepsilon_{1n})d\varepsilon_{1n}\Bigg]$$

$$= \frac{1}{1-Q_n(\varnothing)}\Bigg[q_n(1)q_n(2)\, q_n(3)\frac{e^{V_{1n}}}{e^{V_{1n}}+e^{V_{2n}}+e^{V_{3n}}} + q_n(1)q_n(2)(1-q_n(3))\frac{e^{V_{1n}}}{e^{V_{1n}}+e^{V_{2n}}}$$

$$+ q_n(1)(1-q_n(2))\, q_n(3)\frac{e^{V_{1n}}}{e^{V_{1n}}+e^{V_{3n}}} + q_n(1)(1-q_n(2))(1-q_n(3))\Bigg]$$

$$= \frac{1}{1-Q_n(\varnothing)}[P_n(1|\{1,2,3\})Q(\{1,2,3\}|G) + P_n(1|\{1,2\})Q(\{1,2\}|G)$$

$$+ P_n(1|\{1,3\})Q(\{1,3\}|G) + Q(\{1\}|G)] \tag{16}$$

Estimation methods

The unknown parameters in both choice set formation model and discrete choice model are estimated by the maximum likelihood method. The likelihood of observed choices is given by,

$$L = \prod_{n=1}^{N} P_n(i_n) \tag{17}$$

where i_n is the chosen alternative by individual n and $P_n(i)$ is given by equation (15).

In case the information on individual choice sets is available in addition to the choice data, then, the choice set formation model can be separately estimated using the following likelihood:

$$L_1 = \prod_{n=1}^{N} Q_n(C_n|G) \tag{18}$$

where Q_n is given by equation (8) and C_n is the choice set of individual n given by the data. Then, one can substitute the parameter estimates of the choice set formation model into equation (15) and estimate the parameter of the discrete choice model by maximizing equation (17).

EMPIRICAL ANALYSIS

Description of the case

The proposed model is applied to destination choice of vacation trips. Destination choice is a typical example of choice contexts with a large number of potential alternatives. Furthermore, it is usually very difficult for the analyst to identify individual choice sets for infrequent behavior such as vacation trips.

In the empirical analysis presented in this paper, alternative destinations are defined as 18 partitioned regions of Japan. A survey on domestic multi-day vacation trips was conducted for 600 college students. Surveyed items include characteristics of the trips during the past year, subjective choice sets at the time of the trip, and subjective ratings of the 18 regions with respect to vacation potentials. The subjective choice sets were obtained by asking other regions that the respondent wished to visit at the time of the decision making. Appropriateness of using this question as the indicators of the latent choice sets is discussed later.

Two composite variables, *attractiveness* and *perceptual distance* of each region, are calculated using the subjective rating data. The subjective ratings were obtained with respect to scenic, historical, gastronomic, cultural and athletic attractiveness of each region and perceptual distance of each region from the trip origin. Using these ratings as indicators of the two latent composite variables, the linear structural equation model is applied to estimate the latent composite variables. (The presentation of this analysis is not the scope of this paper. Refer to Morikawa et al. (1990) for explanation of this methodology).

Model

The choice set formation model is composed of two constraints: *information availability* and *minimum attraction*. The probability that region i is included in the choice set of individual n is expressed below:

$$\begin{aligned} q_n(i) &= q_{1n}(i)q_{2n}(i) \qquad (19) \\ &= \frac{1}{1+e^{-(\alpha_1 \mathrm{inf}_{in}-\mu_1)}} \times \frac{1}{1+e^{-(\alpha_2 \mathrm{attrct}_{in}-\mu_2)}} \end{aligned}$$

where

inf_{in} = amount of information on vacation possibility of region i; (this is approximated by the frequency of appearance of the region in travel guidebooks and advertisements of travel agents), and

$attrct_{in}$ = attractiveness of region i (this latent variable is estimated by the above mentioned model).

The utility function of the discrete choice model is specified as below:

$$U_{in} = \beta_1 attrct_{in} + \beta_2 dist_{in} + \beta_3 hotel_i + \beta_4 cost_i + \varepsilon_{in} \tag{20}$$

where

$dist_{in}$ = perceptual distance of region i (estimated by the above mentioned model);

$hotel_i$ = number of rooms of hotels in region i; and

$cost_i$ = average room rate of hotels and inns in region i.

Estimation results

The first column of Table 1 shows the parameter estimates of the two stage PCS model by the sequential method. In the first step the choice set formation model given by equation (8) is estimated by regarding the stated subjective choice set as the indicator of the true choice set (the likelihood is given by equation (18)). Then, the estimated parameters of the choice set formation model are substituted in equation (15) and the parameters of the discrete choice model are estimated by equation (15) (the likelihood is given by equation (17)). The result shows that alternatives are included in the choice set when the amount of information and the attractiveness exceed positive threshold values.

The second column of Table 1 shows the PCS model estimated by the simultaneous method. It uses only the information of "choice" to estimate both the choice set formation and discrete choice models. The parameter estimates have the correct signs, indicating the validity of the proposed method. The attractiveness variable is insignificant in the choice set formation model probably because the variable also appears in the discrete choice model.

Table 1 Estimation results (t-statistics in parentheses)

			PCS modes		DCS models	
	Parameter	Variable	sequential estimation	simultaneous estimation	full choice set	stated choice set
Choice set formation stage	α_1	inf	0.777 (8.0)	0.545 (1.9)		
	μ_1	threshold	2.25 (31.8)	1.49 (2.7)		
	α_2	attrct	0.105 (2.7)	0.0120 (0.1)		
	μ_2	threshold	1.80 (47.8)	0.092 (-0.3)		
Choice stage	β_1	attrct	4.43 (2.3)	0.417 (2.5)	0.384 (1.2)	-0.0606 (0.2)
	β_2	dist	-3.82 (-2.5)	-0.351 (-2.9)	-0.720 (3.0)	-0.434 (-2.0)
	β_3	hotel	2.00 (4.0)	0.245 (4.0)	0.264 (2.3)	0.0293 (0.8)
	β_4	cost	-3.87 (-3.3)	-0.579 (-4.5)	-0.600 (3.0)	

Ordinary MNL models with deterministic choice sets (DCS models) are shown in the third and fourth columns of Table 1. The model of the third column assumes that every individual has the full choice set consisting of 18 alternatives. The model of the fourth column assumes that the stated subjective choice set is the true (deterministic) choice set. Parameter estimates of DCS models are less significant than the PCS models. Particularly, the fourth model yields unacceptable estimation results, which may imply that the stated subjective choice set is not a proper indicator of the true choice set in this data set. In general, it seems difficult to elicit the latent choice set by directly questioning it to the respondent.

Hence, in case that reliable information on individual latent choice sets is not available, the simultaneous estimation method that does not require such information may result in more reliable parameter estimates.

CONCLUSION

This paper proposed a two stage choice model with probabilistic choice sets and its practical estimation method. In the proposed two stage model, each stage employs a different decision making protocol. The first stage, or choice set formation process, is characterized by the non-compensatory rule among the constraints. It represents that the choice set is formed by the alternatives which satisfy every constraint. The second stage is modeled by ordinary discrete choice models with the utility maximization process that is characterized by the compensatory rule.

PCS models that have been proposed thus far cannot virtually be estimable when there are a large number of alternatives (eg more than four) unless very restrictive assumptions on possible choice sets are made (eg Dogit model). The method proposed in this paper compares alternatives in pairwise and derives a computationally tractable form even for a large number of alternatives.

The model was applied to the destination choice of vacation trips that typically has a large number of alternatives and uncertain individual choice sets. According to the empirical analysis, the proposed model yielded reasonable parameter estimates and better fit to the data than the DCS models.

This research is the first attempt of applying the proposed PCS model and, therefore, problems to be solved are left out for further research. They include development of more efficient estimation programs and more persuasive empirical evidence of the effectiveness of the method.

REFERENCES

Ben-Akiva, M. and Lerman, S. (1985) *Discrete Choice Analysis*, MIT Press, Cambridge.

Gaudry, M. and Dagenais, M. (1979) The dogit model, *Transportation Research B* 13B, 105-111.

Manski, C. (1977) The structure of random utility models, *Theory and Decision* 8, 229-254.

Morikawa, T., Ben-Akiva, M. and McFadden, D. (1990) Incorporating psychometric data in econometric demand models, Prepared for *Banff Invitational Symposium on Consumer Decision Making and Choice Behavior*, Banff.

Swait, J. and Ben-Akiva, M. (1987) Incorporating random constraints in discrete models of choice set generation, *Transportation Research B* 21B, 91-102.