# A comparative analysis of implicit and explicit methods to model choice set generation

**Michel Bierlaire**
**Ricardo Hurtubia**
**Gunnar Flötteröd**

**STRC 2009** **September 2009**


STRC 2009

# A comparative analysis of implicit and explicit methods to model choice set generation

Michel Bierlaire
Transport and Mobility Laboratory
EPFL
1015 Lausanne
phone: +41 21 6932537
fax: +41 21 6938060
michel.bierlaire@epfl.ch

Ricardo Hurtubia
Transport and Mobility Laboratory
EPFL
1015 Lausanne
phone: +41 21 6939329
fax: +41 21 6938060
ricardo.hurtubia@epfl.ch

Gunnar Flötteröd
Transport and Mobility Laboratory
EPFL
1015 Lausanne
phone: +41 21 6939329
fax: +41 21 6938060
gunnar.floetteroed@epfl.ch

September 2009

## Abstract

In this paper, we compare two methods to model the formation of choice sets in the context of discrete choice models. The first method is the probabilistic approach proposed by Manski (1977), who explicitly models the choice set generation process by expressing the choice as the joint probability of selecting a choice set and an alternative from this set. This approach is theoretically sound and unbiased, but it is hard to implement due to the complexity that arises from the combinatorial number of possible choice sets. The second method, known as the Constrained Multinomial Logit (Martinez *et al.*, 2009), models the choice set generation process implicitly through elimination of alternatives. This approach is easier to implement because it does not require to enumerate the possible choice sets, allowing to deal with large choice sets, but can only be understood as an approximation of Manski's approach.

An experimental analysis and comparison of both methods in presented. Results based on synthetic data show that the Constrained Multinomial Logit may be a poor approximation of Manski's model, with some clear exceptions which are identified and analyzed.

## Keywords

discrete choice set generation

# 1 Introduction

In standard choice models, it is assumed that the alternatives considered by the decision maker can be deterministically specified by the analyst. The choice set is characterized by deterministic rules based on the characteristics of the decision maker and the choice context. For example, single-room apartments are not considered by families with children in a house choice context, car is not considered as a possible transportation mode if the traveler has no travel license, or no car.

There are, however, many situations where the deterministic choice set generation procedure is not satisfactory, or even possible. Data may be unavailable (the number of children in the household is unknown to the analyst), or rules are fuzzy by nature. For instance, train is not considered as a transportation mode if it involves a long walk to reach the train station. But how long is a "long walk"?

Modeling explicitly the choice set generation process involves a combinatorial complexity, which makes the models intractable except for some specific instances. Manski (1977) defines the theoretical framework in a two stage process, where the probability that decision maker $n$ chooses alternative $i$ is given by

$$P_n(i) = \sum_{\mathcal{C}_m \subseteq \mathcal{C}} P_n(i|\mathcal{C}_m) P_n(\mathcal{C}_m) \tag{1}$$

where $P_n(i|\mathcal{C}_m)$ is the probability for individual $n$ to choose alternative $i$ conditional to the choice set $\mathcal{C}_m$ and $P_n(\mathcal{C}_m)$ is the probability for individual $n$ to consider choice set $\mathcal{C}_m$. The sum runs on every possible subset $\mathcal{C}_m$ of the universal choice set $\mathcal{C}$.

Swait and Ben-Akiva (1987) and Ben-Akiva and Boccara (1995) build on this framework and use explicit random constraints to determine the choice set generation probability. The probability of considering a choice set $\mathcal{C}_m$ is a function of the consideration of the different alternatives in the universal choice set:

$$P_n(\mathcal{C}_m) = \frac{\prod_{i \in \mathcal{C}_m} \phi_{in} \prod_{j \notin \mathcal{C}_m} (1 - \phi_{jn})}{1 - \prod_{k \in \mathcal{C}} (1 - \phi_{kn})} \tag{2}$$

where $\phi_{in}$ is the probability that alternative $i$ is considered by user $n$, which may be modeled by a binary logit model that depends on the alternative's attributes. Note that 2 assumes independence of the consideration probabilities across alternatives.

Swait (2001) proposes to model the choice set generation as an implicit part of the choice process in a multivariate extreme value (MEV) framework, requiring no exogenous information. Here, choice sets are not separate constructs but another expression of preferences. The proba-

bility of considering a choice set is defined as the probability for that choice set to correspond to the maximum expected utility for an individual $n$:

$$P_n(\mathcal{C}_m) = \frac{e^{\mu I_{n,\mathcal{C}_m}}}{\sum_{\mathcal{C}_k \subseteq \mathcal{C}} e^{\mu I_{n,\mathcal{C}_k}}} \tag{3}$$

where $\mu$ is the scale parameter for the higher level decision (choice set selection) and $I_{n,\mathcal{C}_m}$ is the inclusive value (the "logsum" or expected maximum utility) of choice set $\mathcal{C}_m$ for decision maker $n$:

$$I_{n,\mathcal{C}_m} = \frac{1}{\mu_m} \ln \sum_{j \in \mathcal{C}_m} e^{\mu_m V_{nj}}. \tag{4}$$

Here, $\mu_m$ is the scale parameter and $V_{nj}$ is the deterministic utility of alternative $i$ for decision maker $n$. Swait's probabilistic choice set generation approach does not require additional assumptions by the analyst about which attributes affect an alternative's availability.

Clearly, these methods are hardly applicable to medium and large scale choice problems due to the computational complexity that arises from the combinatorial number of possible choice sets. If the number of alternatives in the universal choice set is $J$, the number of possible choice sets is $(2^J - 1)$.

Therefore, various heuristics have been proposed in the literature that derive tractable models by approximating the choice set generation process.

In the context of route choice, Frejinger *et al.* (forthcoming) assume that all decision makers consider the universal choice set, so that $P_n(\mathcal{C}_m) = 0$ when $\mathcal{C}_m \neq \mathcal{C}$, and only one term remains in (1). However, this may not be appropriate in other contexts.

The most promising heuristics are based on the use of penalties of the utility functions, and have been proposed by Cascetta and Papola (2001) (the Implicit Availability/Perception (IAP) model) and expanded by Martinez *et al.* (2009) (the Constrained Multinomial Logit (CMNL) model). In the next section, we briefly describe the CMNL model and provide its theoretical background in the context of choice set generation. In Section 3, we compare the CMNL with the theoretical framework (1), first through a simple example and, second, by estimating both models on synthetic data. Section 4 concludes the paper and identifies possible further work.

# 2 Choice set generation with the CMNL model

Assuming that $\mathcal{C}_n$ is the choice set that the decision maker is actually considering, the choice model is given by

$$P_n(i|\mathcal{C}_n) = \Pr\left(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n\right), \tag{5}$$

where $U_{in}$ is the random utility associated with alternative $i$ by decision maker $n$. If $\mathcal{C}_n$ is known to the analyst, it can be characterized by indicators of the consideration of each alternative by the decision maker:

$$A_{in} = \begin{cases} 1 & \text{if alternative } i \text{ is considered by individual } n, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The choice model can be equivalently written as

$$\begin{aligned} P_n(i|\mathcal{C}_n) &= \Pr\left(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n\right) \\ &= \Pr\left(U_{in} + \ln A_{in} \geq U_{jn} + \ln A_{jn}, \forall j \in \mathcal{C}\right). \end{aligned} \tag{7}$$

For an unconsidered alternative, this adds $\ln 0 = -\infty$ to its utility, so that the choice probability is 0, whereas the addition of $\ln 1 = 0$ has no effect on the utility of a considered alternative.

In the case of a logit model, the choice probabilities are

$$P_n(i) = \frac{e^{V_{in} + \ln A_{in}}}{\sum_{j \in C} e^{V_{jn} + \ln A_{jn}}}. \tag{8}$$

The heuristics proposed by Cascetta and Papola (2001) and Martinez *et al.* (2009) consist in replacing the indicators $A_{in}$ by the probability $\phi_{in}$ that individual $n$ considers alternative $i$.

Cascetta and Papola (2001) introduce the IAP model as a way to incorporate awareness of paths into route choice modeling without requiring an explicit choice set generation step. A similar approach that penalizes the utilities of "dominated" alternatives is proposed by Cascetta *et al.* (2007).

Martinez *et al.* (2009) expand the IAP idea and propose the CMNL model. The functional form for $\phi_{in}$ is assumed to be a binary logit, considering that the availability of an alternative is related with bound constraints on its attributes. For example, if $X_{ink}$ is the $k$th variable of alternative $i$ for decision maker $n$ that influences the consideration of $i$, we have

$$\phi^u_{in}(X_{ink}; u_k, \omega_k) = \frac{1}{1 + \exp(\omega_k(X_{ink} - u_k))} \tag{9}$$

where the $u_k$ parameter is the value at which the constraint is most likely to bind, and $\omega_k$ is the scale parameter of the binary logit. For instance, $X_{ink}$ may be the walking distance to the train station, and $u_k$ may be the maximum distance that individual $n$ is willing to walk. Both $u_k$ and $\omega_k$ are to be estimated. The intuition is that when the attribute $X_{ink}$ exceeds $u_k$, the consideration probability $\phi^u_{in}$ tends to zero, while this availability tends to one when the value of the attribute is below $u_k$.

Expression 9 represents an upper value cut-off, where $u_k$ represents the maximum value that the attribute $X_{ink}$ can have in order for alternative $i$ to be considered. To model a lower value cut-off, we only need to invert the sign of the scale parameter $\omega_k$:

$$\phi^\ell_{in}(X_{ink}; \ell_k, \omega_k) = \frac{1}{1 + \exp(-\omega_k(X_{ink} - \ell_k))}. \tag{10}$$

Functions 9 and 10 can be generalized to account for more than one constraint:

$$\phi_{in}(X_{in}; \ell, u, \omega) = \prod_k \phi^u_{in}(X_{ink}; u_k, \omega_k)\phi^\ell_{in}(X_{ink}; \ell_k, \omega_k). \tag{11}$$

The CMNL approach has an operational advantage over Manski's framework since it does not require enumerating the choice sets, which makes it easier to specify and estimate. However, the CMNL model is a heuristic that is based on convenient assumptions about the functional form of the utility function. The CMNL model can thus be understood as an approximation. The next section evaluates the quality of this approximation.

# 3 Comparison of CMNL with Manski's model

This section compares the CMNL model with Manski's model. For this, we first present a simple example where we analytically analyze the difference between the choice probabilities obtained using both models. Second, we estimate the CMNL model and Manski's model over synthetic data and compare the results. For notational simplicity, we subsequently omit the index $n$ for the decision maker.

## 3.1 Simple example

Consider a logit model with only 2 alternatives, where alternative 1 is always considered ($\phi_1 = 1$) and alternative 2 has probability $\phi_2$ of being considered by the decision maker. Figure 1 shows the structure of Manski's framework if we consider every possible combination of

alternatives as a choice set. This simple situation corresponds to a case where the decision maker is captive to alternative 1 with probability $1 - \phi_2$ (see also the captivity logit model proposed by Gaudry and Dagenais (1979)).

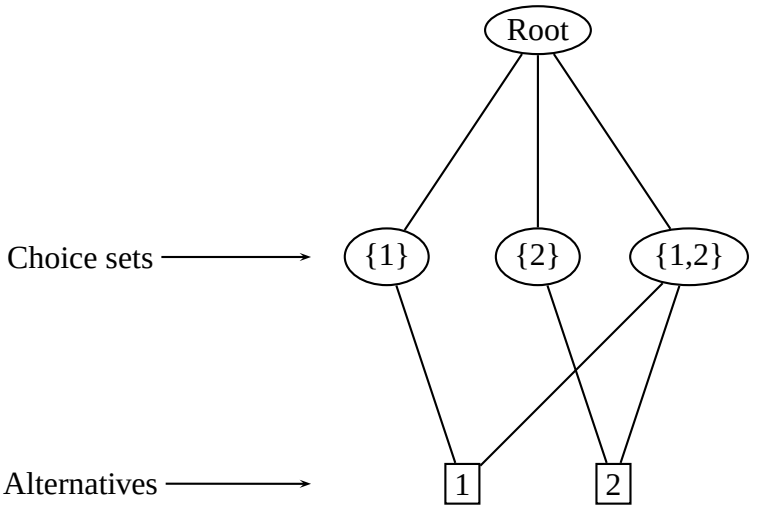



Figure 1: Example of a model in Manski's framework

The CMNL model defines the probability of choosing alternative 1 as

$$P(1) = \frac{e^{V_1}}{e^{V_1} + e^{V_2 + \ln \phi_2}}. \tag{12}$$

Manski's model (1) defines the probability of choosing alternative 1 as

$$P\left(1\right) = P(\{1\})\frac{e^{V_1}}{e^{V_1}} + P(\{1,2\})\frac{e^{V_1}}{e^{V_1} + e^{V_2}} \tag{13}$$

where $P(\{1\})$ is the probability of considering the choice set composed only of alternative 1 and $P(\{1,2\})$ is the probability of considering the choice set containing both alternatives. According to (2), the choice set probabilities are

$$P(\{1\}) = \frac{\phi_1(1-\phi_2)}{1-(1-\phi_1)(1-\phi_2)} = 1 - \phi_2 \tag{14}$$

and

$$P(\{1,2\}) = \frac{\phi_1\phi_2}{1-(1-\phi_1)(1-\phi_2)} = \phi_2. \tag{15}$$

The probability of considering choice set $\{2\}$ is zero because alternative 1 is always be avail-

able. Therefore, (13) becomes

$$P(1) = (1 - \phi_2) + \phi_2 \frac{e^{V_1}}{e^{V_1} + e^{V_2}} \tag{16}$$

In the deterministic limit ($\phi_2 = 0$ or $\phi_2 = 1$), both models are equivalent. However, this is not the case anymore when $\phi_2$ takes values between zero and one. The resulting choice probabilities are shown in Figure 2, assuming the same utility level $V_1 = V_2$ for both alternatives.

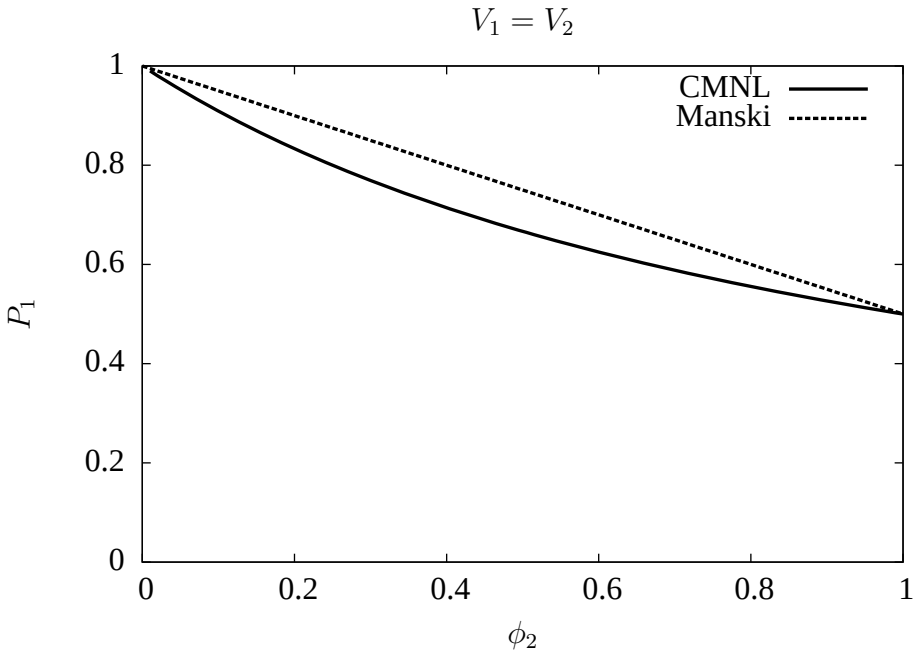



Figure 2: Choice probability of alternative 1 ($V_1 = V_2$)

This figure shows that the CMNL is a good approximation of Manski's model only when $\phi_2$ is close to either zero or one, but it underestimates the probability of alternative 1 elsewhere. If the utility for alternative 1 is larger than the utility for alternative 2 (Figure 3), the approximation improves. This makes sense since the more an alternative is dominated, the less important it is to know if it really belongs to the choice set.

However, as the utility of alternative 1 becomes smaller and smaller compared to the utility of alternative 2, the CMNL becomes a poorer and poorer approximation of Manski's model for intermediate $\phi_2$ values, which is demonstrated in Figures 4 and 5.

These results can be interpreted as an unwanted compensatory effect in the CMNL model. The constraint is enforced by modifying the utility of the constrained alternative. However, when the utility of this alternative is high, it compensates the penalty. We analyze the performance of the CMNL on synthetic data in the next section.

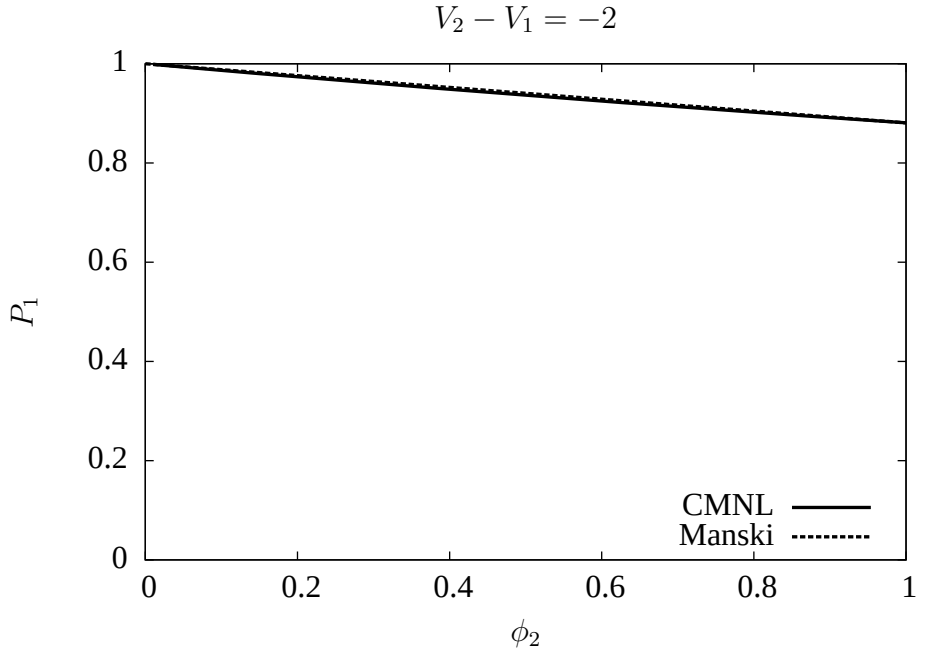



Figure 3: Probability of alternative 1 $(V_1 > V_2)$

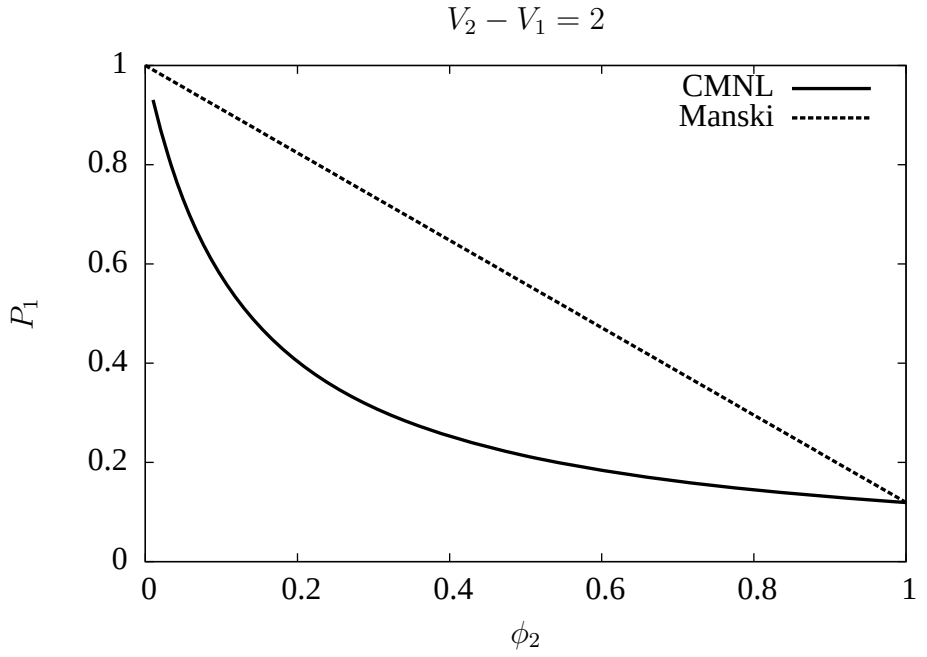



Figure 4: Choice probability of alternative 1 $(V_1 < V_2)$

## 3.2 Synthetic data

This section describes a series of controlled experiments where some of the data is synthetically generated. We start from a real stated preference data set that was collected for the analysis of a hypothetical high speed train in Switzerland (Bierlaire *et al.*, 2001). The alternatives are:

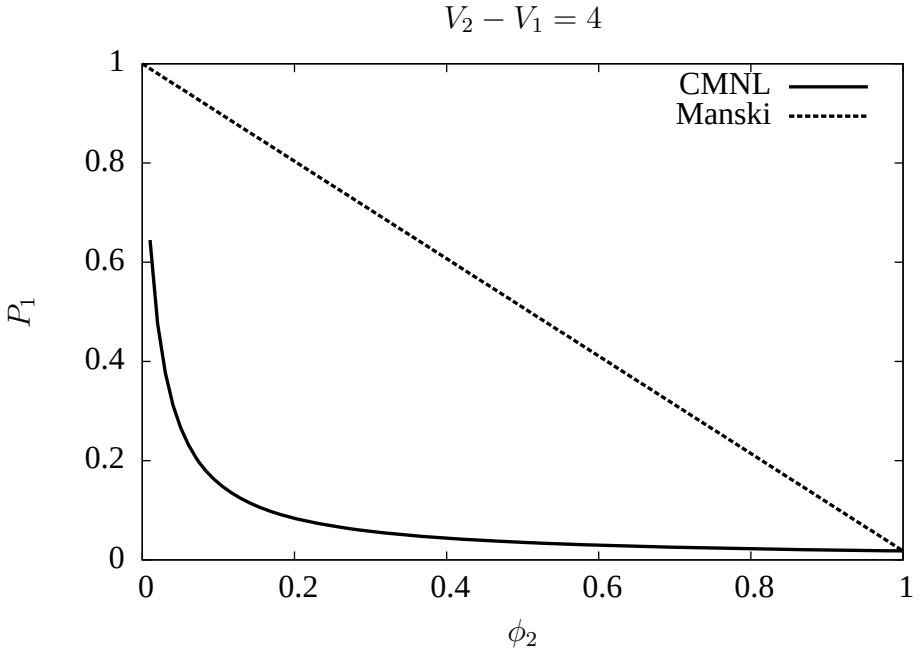



Figure 5: Choice probability of alternative 1 ($V_1 < V_2$)

1. Driving a car (CAR)
2. Regular train (TRAIN)
3. Swissmetro, the future high speed train (SM)

From this data set, which consists of 5607 observations, we use the attributes of the alternatives and simulate synthetic choices based on a postulated "true" model. It is a logit model with linear-in-parameters utility functions. The specification table as well as the "true" values of the parameters are reported in Table 1. The values have been obtained by estimating the model on real choices, and by rounding the estimates.

It is assumed that the TRAIN and the SM alternatives are always considered, whereas the consideration of the CAR alternative depends on the travel time according to

$$\phi_{\text{CAR}} = \frac{1}{1 + \exp(\omega(TT_{\text{CAR}}/60 - a))}, \tag{17}$$

which states that the probability of considering CAR as an available alternative decreases with the travel time $TT_{CAR}$, in minutes, and that this probability is 0.5 when the availability threshold $a$, in hours, is reached.

This implies that, depending on the availability of the CAR alternative, there are two possible choice sets: the full choice set and the choice set containing only the TRAIN and the SM alternative. The random constraints approach (Ben-Akiva and Boccara, 1995) defines the

| Parameter | Value | Car | Train | Swissmetro |
|---|---|---|---|---|
| $\text{ASC}_{\text{CAR}}$ | 0.3 | 1 | 0 | 0 |
| $\text{ASC}_{\text{SM}}$ | 0.4 | 0 | 0 | 1 |
| $\beta_{cost}$ | -0.001 | Cost (CHF) | Cost (CHF) | Cost (CHF) |
| $\beta_{tt}$ | -0.001 | In veh. travel time (minutes) | In veh. travel time (minutes) | In veh. travel time (minutes) |
| $\beta_{he}$ | -0.005 | 0 | Headway (minutes) | Headway (minutes) |
| $a$ | 3 | Consideration threshold of car (hours) | | |
| $\omega$ | 1,2,3,5,10 | Consideration dispersion of car | | |

Table 1: Parameter descriptions and values

probability of each choice set as follows:

$$\begin{aligned} P(\{\text{TRAIN, SM}\}) &= \frac{\phi_{\text{TRAIN}}\phi_{\text{SM}}(1-\phi_{\text{CAR}})}{1-(1-\phi_{\text{CAR}})(1-\phi_{\text{TRAIN}})(1-\phi_{\text{SM}})} \\ &= 1-\phi_{\text{CAR}} \end{aligned} \tag{18}$$

and, accordingly,

$$P(\{\text{CAR, TRAIN, SM}\}) = \phi_{\text{CAR}}. \tag{19}$$

The synthetic choices are generated by (i) simulating a choice set for each decision maker according to (18) and (19), and (ii) simulating a choice for each decision maker using the "true" model specified in Table 1.

100 choice data sets are simulated for each value of $\omega$. These values generate constraints with different levels of uncertainty. Figure 6 shows the shape of these constraint functions. Estimation results for both the Manski and the CMNL model are given in Tables 2 and 3. For each parameter $\beta$, the average value $\bar{\beta}$ and the standard error $\sigma$ over 100 simulations are computed. In the tables, both $\bar{\beta}$ and the t-statistic $(\bar{\beta}-\beta)/\sigma$ are reported, the latter value being used to test if the estimated value is significantly different from the true one.

The estimates of Manski's model are unbiased. We cannot reject the hypothesis that the true value of any parameters is equal to the postulated value, at 95% level. Several estimates of the CMNL model are biased (marked with *), the hypothesis that the true value of the parameter is equal to the postulated value being rejected at the 95% level. The quality of the CMNL estimates improves with decreasing dispersion (increasing $\omega$). This is consistent with the findings of Section 3.1.

Table 2: Estimation results for Manski's model

| real $\omega$ value | | 1 | | 2 | | 3 | | 5 | | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| parameter | real value | estimate | t-test | estimate | t-test | estimate | t-test | estimate | t-test | estimate | t-test |
| $ASC_{CAR}$ | 0.3 | 0.304 | 0.027 | 0.288 | 0.113 | 0.300 | 0.010 | 0.301 | 0.012 | 0.314 | 0.184 |
| $ASC_{SM}$ | 0.4 | 0.396 | 0.044 | 0.399 | 0.010 | 0.405 | 0.053 | 0.401 | 0.017 | 0.410 | 0.151 |
| $\beta_{cost}$ | -0.01 | -0.010 | 0.283 | -0.010 | 0.001 | -0.010 | 0.179 | -0.010 | 0.052 | -0.010 | 0.012 |
| $\beta_{he}$ | -0.005 | -0.005 | 0.241 | -0.005 | 0.010 | -0.005 | 0.048 | -0.005 | 0.082 | -0.005 | 0.078 |
| $\beta_{time}$ | -0.01 | -0.01 | 0.074 | -0.010 | 0.050 | -0.010 | 0.049 | -0.010 | 0.003 | -0.010 | 0.001 |
| $a$ | 3 | 2.963 | 0.019 | 3.008 | 0.118 | 3.000 | 0.100 | 2.998 | 0.081 | 3.002 | 0.101 |
| $\omega$ | see top | 1.003 | 0.028 | 2.014 | 0.079 | 3.066 | 0.210 | 5.095 | 0.170 | 10.523 | 0.353 |

Table 3: Estimation results for CMNL model

| real $\omega$ value | | 1 | | 2 | | 3 | | 5 | | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| parameter | real value | estimate | t-test | estimate | t-test | estimate | t-test | estimate | t-test | estimate | t-test |
| $ASC_{CAR}$ | 0.3 | 0.503 | 0.950 | 0.421 | 1.153 | 0.406 | 1.365 | 0.380 | 0.988 | 0.326 | 0.313 |
| $ASC_{SM}$ | 0.4 | 0.565 | 2.013 * | 0.550 | 2.375 * | 0.536 | 1.804 | 0.506 | 1.485 | 0.463 | 0.872 |
| $\beta_{cost}$ | -0.01 | -0.008 | 4.825 * | -0.008 | 3.580 * | -0.009 | 2.309 * | -0.009 | 1.182 | -0.010 | 0.613 |
| $\beta_{he}$ | -0.005 | -0.005 | 0.202 | -0.005 | 0.151 | -0.005 | 0.071 | -0.005 | 0.120 | -0.005 | 0.090 |
| $\beta_{time}$ | -0.01 | -0.007 | 3.929 * | -0.008 | 3.645 * | -0.008 | 2.813 * | -0.009 | 2.316 * | -0.009 | 1.523 |
| $a$ | 3 | 2.186 | 1.753 | 2.656 | 3.073 * | 2.773 | 3.762 * | -2.869 | 3.305 * | 2.948 | 1.864 |
| $\omega$ | see top | 1.043 | 0.239 | 2.094 | 0.403 | 3.118 | 0.431 | 5.238 | 0.424 | 12.146 | 3.149 * |

(* indicates an insignificant parameter)

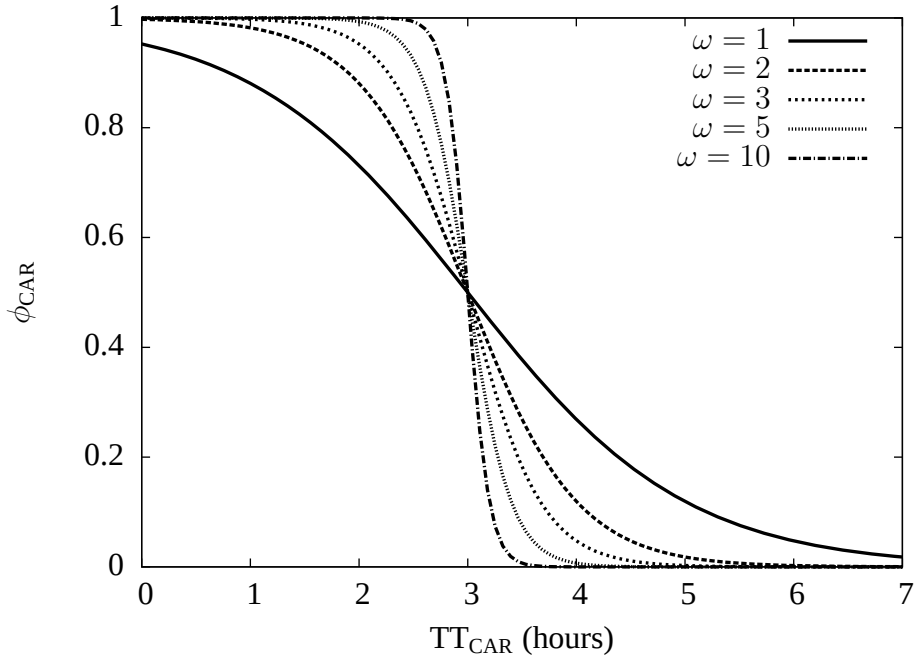



Figure 6: Shape of the constraint for different values of $\omega$

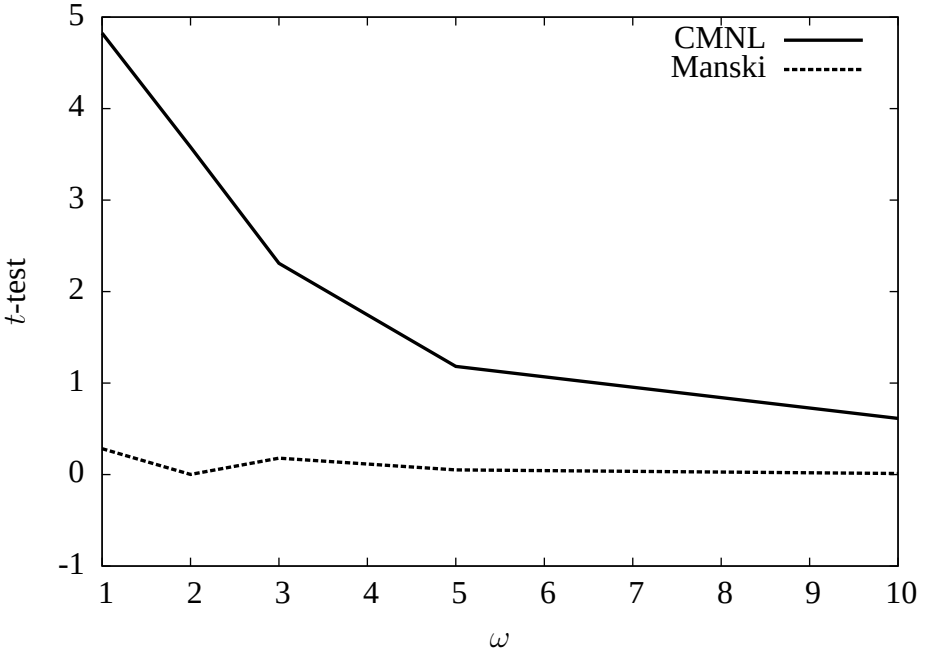
Figure 7 shows the t-statistics for the cost and travel time parameter over different $\omega$ values for Manski's model and the CMNL model. The quality of the estimates is constant across different values of $\omega$ for Manski's model. The quality of the CMNL estimates increases with $\omega$, and their t-statistics reach acceptable values when the constraint function becomes steep.

# 4 Conclusions and further work

We have shown on simple examples that the Constrained Multinomial Logit (CMNL) model is not adequate to model the choice set generation process consistently with Manski's framework. Consequently, the CMNL model should be considered as a model on its own, derived from semi-compensatory assumptions as described by Martinez *et al.* (2009), but not as a way to capture the choice set generation process. Its complexity is linear with the number of alternatives, while Manski's framework exhibits an exponential complexity.

We have started to investigate if a modified version of the CMNL could approximate better Manski's framework, but have been unsuccessful so far. The derivation of a good approximation of Manski's model with the complexity of the CMNL would be particularly useful to handle models with a large number of alternatives.

(a) Cost




(b) Travel time

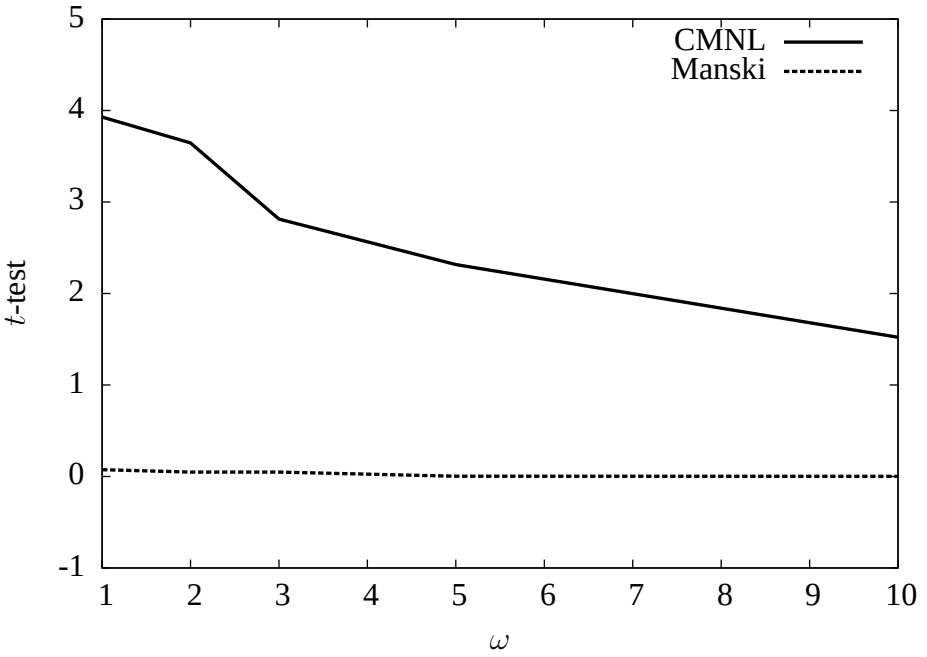



Figure 7: $t$-statistics for the cost and time parameter over $\omega$

# References

Ben-Akiva, M. E. and B. Boccara (1995) Discrete choice models with latent choice sets, *International Journal of Research in Marketing*, **12**, 9–24.

Bierlaire, M., K. Axhausen and G. Abay (2001) Acceptance of modal innovation: the case of the Swissmetro, paper presented at *Proceedings of the 1st Swiss Transportation Research Conference*, Ascona, Switzerland. Www.strc.ch.

Cascetta, E., F. Pagliara and K. Axhausen (2007) The use of dominance variables in choice set generation, paper presented at *Proceedings of the 11th World Conference on Transport Research*, University of California at Berkeley.

Cascetta, E. and A. Papola (2001) Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand, *Transportation Research Part C*, **9**, 249–263.

Frejinger, E., M. Bierlaire and M. Ben-Akiva (forthcoming) Sampling of alternatives for route choice modeling, *Transportation Research Part B: Methodological*. Accepted for publication.

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

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

Martinez, F., F. Aguila and R. Hurtubia (2009) The constrained multinomial logit model: A semi-compensatory choice model, *Transportation Research Part B*, **43**, 365–377.

Swait, J. (2001) Choice set generation within the generalized extreme value family of discrete choice models, *Transportation Research Part B*, **35** (7) 643–666, August 2001.

Swait, J. and M. Ben-Akiva (1987) Incorporating random constraints in discrete models of choice set generation, *Transportation Research Part B*, **21** (2).