The initial condition problem with complete history dependency in learning models for travel choices

C. Angelo GUEVARA
Universidad de Chile
Department of Civil Engineering
University of Chile
Blanco Encalada 2002
Santiago
Phone: +562-29784380
E-mail: crguevar@ing.uchile.cl

Yue TANG
Department of Civil and Environmental Engineering
University of Massachusetts, Amherst
141 Marston Hall
130 Natural Resources Road
Amherst, MA 01003, USA
Phone: +1-413-9927491
Email: yuet@umass.edu

Song GAO (corresponding author)
Department of Civil and Environmental Engineering
University of Massachusetts Amherst
214C Marston Hall
130 Natural Resources Road
Amherst, MA 01003
Phone: +1-413-5452688
Fax: +1-413-5459569
Email: sgao@umass.edu

6148 words + 4 tables
total 7148

June 8, 2017

ABSTRACT

Learning-based models that capture travelers' day-to-day learning processes in repeated travel choices could benefit from ubiquitous sensors such as smartphones, which provide individual-level longitudinal data to help validate and improve such models. However, the common problem of missing initial observations in longitudinal data collection can lead to inconsistent estimates of perceived value of attributes in question, and thus inconsistent parameter estimates. In this paper, the stated problem is addressed by treating the missing observations as latent variables. The proposed method is implemented in practice as maximum simulated likelihood (MSL) correction with two sampling methods in an instance-based learning model for travel choice, and the finite sample bias and efficiency of the estimators are investigated. Monte Carlo experimentation based on synthetic data shows that both the MSL with random sampling (MSLrs) and MSL with importance sampling (MSLis) are effective in correcting for the endogeneity problem in that the percent error and empirical coverage of the estimators are greatly improved after correction. Compared to the MSLrs method, the MSLis method is superior in both effectiveness and computational efficiency. Furthermore, MSLis passes a formal statistical test for the recovery of the population values up to a scale with a large number of missing observations, while MSLrs systematically fails due to the curse of dimensionality. The impacts of sampling size in MSLrs and number of high probability choice sequences in MSLis on the methods' performances are investigated. The methods are applied to an experimental route-choice dataset to demonstrate their empirical application. Hausman-McFadden tests show that the estimators after correction are statistically equal to the estimators of the full dataset without missing observations, confirming that the proposed methods are practical and effective for addressing the stated problem.

1. INTRODUCTION

Learning-based models for travel choice capture travelers' learning process in repeated choices (e.g., Ben-Elia and Shiftan, 2010; Lu et al., 2014; Tang and Gao, conditionally accepted). In a learning model, a traveler' s perception of an alternative's attribute (e.g., travel time) evolves over time based on all her past experience with the alternative. When forming the perception, each past experience with the alternative takes a weight in memory and the perception is a weighted average of all past experience. The weighting scheme of past experience is specific to the learning model in use. Compared to non-learning models where the perception of an alternative is static over time, estimation of a learning model requires data of travelers' complete past experience with the alternatives. Longitudinal data collection in real life, however, inevitably starts midstream, and rarely includes subjects' complete choice histories. Specialized data collection targeted at newcomers (e.g., new employees or students) to a region might provide the needed data, but such efforts are difficult to implement. In the case of incomplete data, the missing initial observations can lead to biased estimate of the perceived value of the attribute in question, and thus inconsistent parameter estimates. Note that the majority of empirical studies on learning models for travel choice are based on experimental data in a laboratory, where subjects make choices from "day" and thus the stated problem does not exist.

An econometric model is said to suffer from endogeneity when the systematic part of the utility is correlated with the error term. The variables that cause the correlation are called the endogenous variables. Endogeneity can lead to inconsistent estimation of model parameters, since changes in the error term are misinterpreted as changes of the endogenous variable. Endogeneity is common in discrete choice models (e.g., probit, logit, nested logit) as the assumption that the explanatory variables are independent from the error term is often violated. Guevara (2010) classifies endogeneity into three types based on their causes: (1) Omission of the variables that are correlated with some observed variables; (2) Simultaneous determination of multiple variables; and (3) The propagation of measurement errors in explanatory variables to the error term. Several correction methods have been developed to solve endogeneity problems (e.g., Berry et al., 1995; Brownstone, 1991; Fernández-Antolín et al., 2016; Guevara, 2010; Guevara and Polanco, 2016; Heckman, 1978; Schenker and Welsh, 1988) . The endogeneity problem this paper tackles can be classified within the third group, a special case in which endogeneity arises because the researcher has an incorrect measure of the attributes of the alternatives perceived by the decision makers.

Solving the initial observation problem for dynamic panel data discrete choice models is known to be a difficult task. Most existing studies deal with first-order Markov process where the dependent variable is only lagged once. The major focus of these studies is that the initial condition is not exogenous due to correlation of error terms over time. Therefore, if there is no serial correlation, first-order Markov process model would not suffer from the problem. For example, Heckman (1981) and Lee (1997) examined the problem of initial conditions in a time-discrete data stochastic process when serially correlated unobservable variables generate the process. Correction methods were proposed and tested with Monte Carlo experiments. More of such studies can be found in the reference list (e.g. Blundell and Bond, 1998; Carro, 2007; Honore and Kyriazidou, 2000; Wooldridge, 2005). In the learning models for travel choice, a current decision depends on the entire history of past experience, defined as a Polya process in Heckman (1981a). The complete history dependence makes the initial condition problem more challenging than those in the existing studies. The model will suffer from the initial observation problem even without serial correlation. To the best of our knowledge, no solution has been developed to date.

In this paper, the proposed method is based on noting that the likelihood function of this problem can be written as a sequence of integrals over the conditional distribution of the possible choices on the missing days. This multifold integral is then maximized using a variation of the maximum simulated likelihood (MSL), which is described in detail by Train (2009). The MSL numerical estimation method has reached great popularity in the past 15 years, thanks to the significant improvement in computational power. This method has been mainly used for the estimation of Logit Mixture models aimed to account for random coefficients or different error component. The application of the method in this paper is different from the usual ones, although all the conditions for consistency described in Train (2009) are extendable, e.g., the need for having the number of draws growing faster than the square root of the sample size. Despite its popularity, the MSL is not exempt from drawbacks. For example, MSL estimators have a downward bias for a finite number of draws, and they may suffer from empirical identification problems, both in the form of false empirical identification and lack of empirical identification. More importantly for this application, MSL may suffer from the problem known as the curse of dimensionality, which in this case implies that the number of draws required for estimation grows exponentially with the number of missing days, quickly making estimation impractical. This problem is shared by all estimation methods based on simulation. This issue will be illustrated and investigated with Monte Carlo experimentation.

Two sampling methods are proposed for the correction. The MSL random sampling (MSLrs) method ran-

domly draws a set of missing choice sequences following the learning model and a simple average of the simulated choice probabilities is used in the simulated likelihood. This sampling method is expected to suffer from the curse of dimensionality as the number of missing days grows. To overcome this limitation, the MSL importance sampling (MSLis) method is proposed. It can be seen as a variation of the kernel conditional density nonparametric estimator proposed by Rosenblatt (1969) and enhanced by Hyndman et al. (1996). In this case, instead of randomly simulating a large enough number of missing choice sequences to evaluate the Logit Kernel function, a small number of sequences with high probability of occurrence are sampled and the kernel, conditioning on the said probability are evaluated.

The main contribution of this paper is that a practical and theoretically sound correction method is developed and assessed to address the endogeneity problem due to missing initial observations in learning models with complete history dependency. To the best of our knowledge, the stated problem is tackled for the first time. Two sampling methods are proposed for the correction method, with the aim of avoiding the problem of the curse of dimensionality that arises as the number of missing days grows. The sample bias and effectiveness of the proposed method is investigated using a learning model proposed in recent literature (Tang and Gao, conditionally accepted) where perceived attribute values are non-linear functions of a memory decay parameter. The suitability of the proposed method is confirmed using Monte Carlo experimentation on synthetic data, and its applicability is demonstrated using a laboratory experimental dataset.

The remainder of the paper is organized as follows. The next section introduces the instance-based learning (IBL) model for travel choice and presents the endogeneity problem due to missing initial observations. The MSL method and two sampling methods are then proposed. The effectiveness and applicability of MSL with the two sampling methods are demonstrated using both synthetic data and empirical data. Lastly, conclusions are presented and future research directions are discussed.

2. AN INSTANCE-BASED LEARNING MODEL FOR TRAVEL CHOICE

The IBL model developed by Tang and Gao (conditionally accepted) is utilized to investigate the finite sample bias and effectiveness of the proposed methods in correcting the endogeneity problem due to missing initial observations, since: (1) The model is developed based on mainstream psychological findings of the power law of forgetting and reinforcement and is shown to be able to capture various psychological effects that reside in travelers' repeated choice behaviors (Anderson and Schooler, 1991; Gonzalez et al., 2003; Newell and Rosenbloom, 1981; Rubin and Wenzel, 1996; Wickelgren, 1976). (2) Learning in the IBL model resides in the nonlinear memory decay parameter and is based on complete history. The complexity of the model presents challenges to the effectiveness and efficiency of the proposed method. For illustrative purpose, in this study the IBL model is introduced within a repeated binary route-choice context.

A traveler n chooses one alternative from a choice set with two alternatives on each day t from day 1 to K. Each alternative has an underlying random travel time whose realizations are independent from day to day, and independent across alternatives. The traveler experiences the realized travel times of the chosen alternative on a given day, and has no knowledge of the realized travel time on un-chosen alternatives. An instance is defined as a past experience of a chosen alternative i on day t' and its associated outcome (realized travel time), $x_i(t')$. The realized travel time is not indexed by traveler n, since it is sampled from the nature's process and does not differ depending on who is experiencing it. The index of a past day t' ranges from 0 to $t-1$, where $t'=0$ is a special time index of the traveler's initial perception of the alternative prior to her first experience. The traveler's initial perception is unobserved and reasonable assumptions can be made to represent its value, e.g., free flow travel time or a personal trip planner's information (such as Google Maps). An instance is stored in the declarative memory of the traveler, and its activation decays over time following a power law. Specifically, on day t, its activation is $(t-t')^{-d}$, where the decay parameter d captures the rate of forgetting in that a smaller d value translates into higher activation in memory and $t-t'$ measures the recency of the experienced travel times (smaller $t-t'$ values represent higher recency). The weight of an instance in the traveler' s memory is directly related to its activation.

Eq.(1) shows the weight of the experience from a past day t' for traveler n, where the denominator is the summation of activations over all past experiences on alternative i. The binary indicator $a_{ni}(t)$ indicates whether traveler n chose alternative i on day t. By definition $a_{ni}(0)=1, \forall i$. The weight function shows that recency and frequency jointly define the weight, i.e. more recent and frequent experienced travel times are more active in memory.

$$w_{ni}(t',t)=\frac{a_{ni}(t')(t-t')^{-d}}{\sum_{\tau=0}^{t-1} a_{ni}(\tau)(t-\tau)^{-d}} \tag{1}$$

where

t : index of the current day, $t = 1, \ldots, K$
t' : index of a previous day, $t' = 0, \ldots, t-1$
$w_{ni}(t', t)$: weight of the experienced travel time on day t' for the perceived travel time on day t for alternative i, traveler n
d : decay parameter that captures the rate of forgetting, $d > 0$
$a_{ni}(t')$: a binary indicator. It is 1 if traveler n chose alternative i on day t' and 0 otherwise

On day t, the perceived travel time of alternative i is the weighted average of experienced travel times of all past days when alternative i is experienced, shown in Eq. (2). It depends on the entire choice history $\{a_{ni}(1), \ldots, a_{ni}(t-1)\}$.

$$b_{ni}(t) = \sum_{t'=0}^{t-1} w_{ni}(t', t) x_i(t') \tag{2}$$

where

$b_{ni}(t)$: perceived travel time of alternative i on day t for traveler n
$x_i(t')$: realized travel time of alternative i on day t', $t' = 0, \ldots, t-1$

Eq. (3) shows the utility function with the parameter vector $\boldsymbol{\phi} = \{d, \beta_{time}, \boldsymbol{\alpha}\}$, where the random residual ε is assumed to be i.i.d. Gumbel distributed. The systematic utility is linear in the perceived travel time $b_{ni}(t)$ that varies from day to day and other attributes z_i of the alternative that are constant over time (e.g., bus fare and number of traffic lights). It is straightforward to extend the utility function to include other attributes that vary from day to day, such as perceived fuel consumption or perceived crowdedness of public transit. Eq. (4) and Eq. (5) specify the choice probability of choosing path 1 and log-likelihood of observing all travelers' choice sequences from day 1 respectively. The binary network assumption can be generalized to a larger number of alternatives, where issues such as overlapping alternatives and choice set generation need to be addressed.

$$U_{ni}(t; \phi) = V_{ni}(t; \phi) + \varepsilon_{ni}(t) = \beta_{time} b_{ni}(t) + \alpha' \boldsymbol{z_i} + \varepsilon_{ni}(t) \tag{3}$$

where

$U_{ni}(t)$: random utility of alternative i for traveler n on day t
$V_{ni}(t)$: systematic utility of alternative i for traveler n on day t
β_{time} : coefficient to perceived travel time
$\boldsymbol{z_i}$: explanatory variables for alternative i and traveler n that do not vary from day to day
α : a vector of coefficients for attributes $\boldsymbol{z_i}$
$\boldsymbol{\phi}$: parameter vector, $\boldsymbol{\phi} = \{d, \beta_{time}, \boldsymbol{\alpha}\}$
$\varepsilon_{ni}(t)$: random residuals that are i.i.d. Gumbel distributed at location 0 and scale 1

$$P_n(1|t, \{1, 2\}) = \frac{e^{V_{n1}(t)}}{e^{V_{n1}(t)} + e^{V_{n2}(t)}} \tag{4}$$

where

$P_n(1|t, \{1, 2\})$: choice probability of path 1 for traveler n on day t

$$\ell_{N1} = \sum_{n=1}^{N} \sum_{t=1}^{K} log\Big[P_n(1|t, \{1, 2\})^{a_{n1}(t)} \big(1 - P_n(1|t, \{1, 2\})\big)^{1 - a_{n1}(t)}\Big] \tag{5}$$

3. ENDOGENEITY DUE TO MISSING INITIAL OBSERVATIONS

Suppose the data are collected from day C. It is likely that the travelers have already accumulated some experience with the alternatives prior to day C. In such cases, the dataset only contains observations from day C to day K, while those from day 1 to day $C-1$ are missing. In this paper, the dataset without missing observations is referred as the full dataset, and that with missing observations is referred as the cutoff dataset. Variables in the cutoff dataset are all denoted with asterisks (*), while those of the full dataset are denoted without asterisks. In this section, the cause of endogeneity due to missing initial observations is derived and the estimation biases are demonstrated.

3.1 Cause of endogeneity

The true likelihood of the cutoff dataset is the one shown in Eq. (6). However, this likelihood is impractical to compute because the true perceived travel time $b_{ni}(t)$ cannot be calculated. Recall that for the full dataset, the perceived travel time of alternative i on day t is the weighted average of all past instances (Eq. (2)). Instead of $b_{ni}(t)$, a curtailed version $b^*_{ni}(t)$ could be used, resulting in the modified likelihood shown in Eq. (7), where maximization will not retrieve consistent estimators of the model parameters.

$$\ell_{NC} = \sum_{n=1}^{N}\sum_{t=C}^{K} log\Big[P_n(1|t,\{1,2\})^{a_{n1}(t)}\big(1-P_n(1|t,\{1,2\})\big)^{1-a_{n1}(t)}\Big] \tag{6}$$

$$\ell^*_{NC} = \sum_{n=1}^{N}\sum_{t=C}^{K} log\Big[P^*_n(1|t,\{1,2\})^{a_{n1}(t)}\big(1-P^*_n(1|t,\{1,2\})\big)^{1-a_{n1}(t)}\Big] \tag{7}$$

To illustrate the problem, consider that the perceived travel time can be written as the sum of the weighted average of the perceived travel time derived from the instances from day 0 to day $C-1$ and the perceived travel time derived from the instances from day C to day $t-1$ as in Eq. (8).

$$b_{ni}(t) = \sum_{t'=0}^{C-1} w_{ni}(t',t)x_i(t') + \sum_{t'=C}^{t-1} w_{ni}(t',t)x_i(t') \tag{8}$$

In the cutoff dataset, an initial perception b_i^{IP} is assumed to happen on day $C-1$ to approximate the perceived travel time prior to day C (it is effectively assumed zero if experiences prior to day C are simply ignored), and the perceived travel time at day t is the weighted average of the initial perception and instances happened from day C to day $t-1$ (Eq. (9)). The absolute value of activation of an observed instance (that occurs on or after day C) stays the same as in the full dataset, however it is normalized over a smaller set of instances including the assumed initial perception on day $C-1$, as shown in Eq. (10). Therefore, the weights of the observed instances are scaled up compared to their true weights in the full dataset.

$$b^*_{ni}(t) = w^*_{ni}(C-1,t)b_i^{IP} + \sum_{t'=C}^{t-1} w^*_{ni}(t',t)x_i(t') \tag{9}$$

$$w^*_{ni}(t',t) = \frac{a_{ni}(t')(t-t')^{-d}}{(t-C+1)^{-d} + \sum_{\tau=C}^{t-1} a_{ni}(\tau)(t-\tau)^{-d}} \tag{10}$$

where

$w^*_{ni}(t',t)$: weight of the experienced travel time on day t' for the perceived travel time on day t for alternative i traveler n in the cutoff dataset
$w^*_{ni}(C-1,t)$: weight of initial perception on day $C-1$ for the perceived travel time on day t for alternative i for traveler n in the cutoff dataset
$b^*_{ni}(t)$: perceived travel time of alternative i on day t for traveler n in the cutoff dataset
b_i^{IP} : initial perception of alternative i

TABLE 1 : Endogeneity due to missing initial observations

# of missing obs	β_{time} (-0.4)				β_{cost} (-1.2)				VOT (0.333)			
	Average	Percent error	p-value	Empirical coverage	Average	Percent error	p-value	Empirical coverage	Average	Percent error	p-value	Empirical coverage
0	-0.399	0.0433	0.873	96	-1.12	0.0521	0.780	94	0.333	0.00168	0.944	99
1	-0.339	15.2	<1e-05	59	-1.17	2.61	<1e-05	91	0.390	12.9	<1e-05	66
5	-0.273	31.9	<1e-05	1	-1.17	2.31	<1e-05	92	0.233	30.2	<1e-05	3
10	-0.241	39.8	<1e-05	0	-1.18	1.39	<1e-05	97	0.204	38.9	<1e-05	0
15	-0.216	46.0	<1e-05	0	-1.18	1.58	<1e-05	96	0.183	45.1	<1e-05	0
20	-0.200	49.9	<1e-05	0	-1.18	1.41	<1e-05	96	0.170	49.1	<1e-05	0
25	-0.183	54.3	<1e-05	0	-1.18	1.92	<1e-05	93	0.156	53.3	<1e-05	0
30	-0.167	58.2	<1e-05	0	-1.17	2.59	<1e-05	94	0.143	57.1	<1e-05	0
35	-0.151	62.2	<1e-05	0	-1.16	3.25	<1e-05	91	0.130	60.9	<1e-05	0
40	-0.131	67.3	<1e-05	0	-1.15	4.39	<1e-05	92	0.114	65.7	<1e-05	0

*Estimation results are based on 100 repetitions. The nominal value of empirical coverage is 95%. P-values are calculated against true values.

The discrepancy between the perceived travel time in the cutoff dataset $b^*_{ni}(t)$ and that of the full dataset $b_{ni}(t)$ is propagated to the error term, such that the error term in the utility function of the cutoff dataset $e_{ni}(t) = \varepsilon_{ni}(t) + \beta_{time}(b_{ni}(t) - b^*_{ni}(t))$ is correlated to the systematic part of the utility function. Thus, the perceived travel time is the endogenous variable, and the model that omits the missing initial observations can be seen as a model that suffers from a special case of endogeneity due to measurement error.

3.2 Experiments based on synthetic data

The impact of the endogeneity problem on parameter estimates is illustrated using synthetic datasets. Since VOT has important policy indication, toll price is included in the utility function as an attribute that is constant over time to exemplify travel cost. VOT is calculated based on the perceived travel time coefficient β_{time} and toll coefficient β_{cost}. The estimator of VOT is used to investigate the effectiveness of the correction method. The true value of the decay parameter d follows its conventional value of 0.5, and the true values of the perceived travel time coefficient β_{time} and toll coefficient β_{cost} are postulated at -0.4 and -1.2 respectively. The underlying travel time distributions are generated following truncated normal distribution.

100 datasets are generated following the true model. For each dataset, 200 sets of 50-day observations are generated. For each set of observations, the travel time of Path 1 follows a normal distribution with the mean uniformly sampled between 10 and 50 and the standard deviation uniformly sampled between 0.1 and 0.3 times the mean. The mean travel time of Path 2 is uniformly sampled between 0.8 and 1.2 times the corresponding mean travel times of Path 1, and the standard deviation uniformly sampled between 0.1 and 0.3 times the mean. The travel time distributions of both paths are truncated at half of its mean travel time to mimic a distribution with a lower bound set by the free flow travel time. The toll price of both paths is uniformly sampled between \$0 to \$10. Without any other information, the mean travel time of an alternative is the best approximation one can find to use as the initial perception. For simplicity, the decay parameter d is fixed at its true value and only the travel time coefficient β_{time} and toll coefficient β_{cost} are estimated. Unreported Monte Carlo experiments show that the decay parameter d can be retrieved in the full dataset, and in Section 5.3 d is estimated in the empirical dataset. The software R-3.2 is used for both data generation and estimation throught the paper, and Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is used for likelihood maximization.

Table 1 shows the estimation results of the full dataset and cutoff datasets with a variety of number of missing initial observations. The average, percent error from the true value, p-value against the true value, and empirical coverage of each estimate are reported. The empirical coverage is calculated as the percent of the tests among the 100 repetitions where the null hypothesis that the estimator is equal to its true value is accepted with 95% confidence. For the full dataset, the percent errors of the model parameters and VOT are all very small. Both the empirical coverages and p-values suggest the retrieval of the true values with 95% confidence. For the curtailed model, however, all the metrics suggest that the null hypothesis of the retrieval of the true value is rejected even when only 1 observation is missing. Thus, it is concluded that the missing initial observations can cause the endogeneity problem in a learning model and this problem gets more severe as the number of missing observations increases.

4. MAXIMUM SIMULATED LIKELIHOOD (MSL) METHOD

Realized travel times are assumed observable, since traffic monitoring devices are generally available to obtain travel time measurements. Therefore, the choice histories prior to day C are the only latent variables. The MSL method uses

simulation to integrate out the latent variables. The likelihood function of the IBL model with missing observations can be written as a sequence of integrals over the conditional distribution of the possible missing choices. The multivariate integration is carried out numerically through simulation, and an iterative algorithm is utilized to find the maximum simulated likelihood. At each iteration, the log-likelihood function needs to be evaluated for a given trial values of the parameters. A set of choice sequences prior to day C is obtained based on a specific sampling method for the log-likelihood function. The total probability theorem is used to obtain an estimator of the log-likelihood corresponding to the trial values of the paramenters. The consistency of this method can be demonstrated using an approach equivalent to the one described in Train (2009). The algorithm is described in detail below. As it occurs with other methods to correct for endogeneity in discrete choice models, the proposed MSL method will consistently recover the linear utility coefficients, in general, only up to a scale (Guevara and Ben-Akiva, 2012). For example, if the utility considers travel time and travel cost of each route, then only the ratio of their coefficients, i.e., the VOT, will be consistently recovered with the proposed method, but not the individual coefficients. Conversely, the decay parameter should be fully recovered because of the nonlinear way in which it defines the normalized weights in Eq.1.

Maximum Simulated Likelihood Algorithm with Random Sampling or Importance Sampling

Given the initial trial values ϕ_0, which could be gathered, e.g., from the estimators of the curtailed model.

Iteration $k = 1$

1. Obtain a choice sequence set H_n for the missing days $t = 1$ to $t = C - 1$ for each traveler n following the IBL model. For random sampling, H_n is sampled at each iteration. For importance sampling, H_n is fixed over iterations.
2. For each choice sequence $h_n \in H_n$
   i. For each day $t \geq C$, calculate the perceived travel time $b_{ni}(t)$ using the weights $w(t', t)$ and the sampled choices from h_n, that is, $a_{ni}(\tau) = a_{ni}^{h_n}(\tau)$ for $\tau < C$.
   ii. Calculate the choice probabilities for the current choice sequence h_n for each day $t \geq C$ as $P_n(i|t, \{1, 2\}, h_n)$
3. Based on the chosen sampling method, the choice probability $\hat{P}_n(i|t, \{1, 2\})$ to be considered in the likelihood function is calucated based on $P_n(i|t, \{1, 2\}, h_n), \forall h_n \in H_n$.
4. Find new trial values ϕ_k to maximize the following simulated likelihood to retrieve the estimators $\hat{\phi}$:

$$\tilde{\ell}_{NC}^{MSL} = \sum_{n=1}^{N} \sum_{t=C}^{K} log\left[\left(\hat{P}_n(1|t, \{1, 2\}))\right)^{a_{n1}(t)} \left(\hat{P}_n(2|t, \{1, 2\}))\right)^{1-a_{n1}(t)}\right]$$

$k = k + 1$ to repeat steps 1-4 till convergence. For importance sampling, the set of choice sequences H_n can be re-sampled after enough number of iterations.

The random sampling and importance sampling approaches are proposed to implement the MSL method. The random sampling method follows the simulation approach described by Train (2009) for the Logit Mixture model. It sequentially simulates the missing choice sequences prior to day C following the IBL model with given trial values of ϕ. Sample R times to form the choice sequence set H_n. For each simulated choice sequence h_n, the likelihood of observing choices starting from day C is calculated as $P_n(i|t, \{1, 2\}, h_n)$. Due to the nature of random sampling, the simulated log-likelihood is thus

$$\hat{P}_n(i|t, \{1, 2\}) = \frac{1}{R} \sum_{h_n \in H_n} P_n(i|t, \{1, 2\}, h_n) \tag{11}$$

The importance sampling approach can be better described if the complete enumeration method, a special case of importance sampling that quickly becomes impractical as the number of missing days grows, is reviewed first. The complete enumeration method finds the set H_n by enumerating each possible choice sequence that could have been chosen prior to day C by each traveler n. The probability of occurrence of each possible choice sequence, π_{h_n}, is the product of the sequence of conditional choice probabilities, shown in Eq. (12). Based on the total probability theorem, the choice probability to be considered in the likelihood function can be calculated as a weighted average of the conditional choice probabilites and their respective probability of occurence π_{h_n} as in Eq. (13).

$$\pi_{h_n} = \prod_{t=1}^{C-1} P_n(i|t, \{1, 2\}, h_n^1, h_n^2, ..., h_n^{t-1}) \tag{12}$$

$$\hat{P}_n(i|t,\{1,2\}) = \sum_{h_n \in H_n} P_n(i|t,\{1,2\},h_n)\pi_{h_n} \tag{13}$$

The complete enumeration method becomes quickly impractical as the set of unique choice sequences grows exponentially in the number of missing observations. To avoid this limitation, the importance sampling method defines the choice sequence set H_n by keeping a subset of the full choice sequence set with high probability of occurrence. Based on the total probability theorem, the choice probability to be considered in the likelihood function is calculated as in Eq. (14). The choice sequences in H_n are sampled by simulating the missing choices prior to day $C-1$ sequentially following the IBL model with given trial values of ϕ. If a choice sequence for a given traveler n is drawn twice, the second draw is discarded to keep the sequence set unique. The sequence set H_n is fixed over MSL iterations, and can be re-sampled after a certain number of MSL iterations. In practice, the choice of sampling size shall depend on the number of missing observations and number of high probablity choice sequences. Since the sampling process is independent of the estimation procedure and once a choice sequence is sampled, it can be reused for any given number of high probability sequences, the rule of thumb is to sample a large number of times to cover the sampling distribution of the choice sequences as much as possible.

$$\hat{P}_n(i|t,\{1,2\}) = \frac{\sum_{h_n \in H_n} P_n(i|t,\{1,2\},h_n)\pi_{h_n}}{\sum_{h_n \in H_n} \pi_{h_n}} \tag{14}$$

5. COMPUTATIONAL EXPERIMENTS

The MSLis method can be seen as a variation of the kernel conditional density nonparametric estimator proposed by Rosenblatt (1969) and enhanced by Hyndman et al. (1996). In this case, instead of drawing a large number of choice sequences with potentially very low probability of occurrence, the effort is concentrated on drawing a small number of choice seuqences with large probability of occurrence and evaluating the kernel, conditioning on the said probability. Monte Carlo evidence provided in the following section shows that this modification is critical to avoid the problem of the curse of dimensionality as the number of missing observations grows, achieving a full recovery of the model parameters up to a scale with feasible estimation time.

5.1 Monte Carlo experimentation based on synthetic data

The effectiveness of the MSL using the two sampling methods, i.e. MSLrs and MSLis, is investigated using the same cutoff datasets as in Section 3.2. The experimentation was conducted using the Massachusetts Green High Performance Computing Center (MGHPCC)* clusters. For the reported results in Table 2, 1,400 jobs (100 datasets $\times$ 7 different numbers of missing observations $\times$ 2 methods) were submitted to the center specifying 4GB of memory per job. The estimates before correction given the specific number of missing observations are used as the starting values for all experiments.

Table 2 reports the estimation results before and after applying correction. For the MSLrs method, the choice sequence is sampled 2000 times. It should be noted that this does not necessarily mean that 2000 draws will be enough for a general case, neither even for the synthetic problem at hand. Because of the curse of dimensionality, the number of draws is a dimension of the problem that needs to be investigated in a case by case basis. After correction, the percent error of VOT is generally more than 5 times better than before correction. The empirical coverage is greatly improved although it is still below the nominal value of 95%. The null hypothesis that the estimator is statistically equal to the true parameter value is rejected at all numbers of missing observations. This result is interpreted as evidence that, although consistency is achieved with the proposed correction method, the curse of dimensionality precludes formal recovery of the population parameters for the finite sample size. The MSLis method is proposed as a potential cure for the curse of dimensionality issue for this particular problem. The empirical results suggests that, for the problem at hand, the issue is satisfactorily resolved.

For the MSLis method, the complete enumeration sampling method is used for up to 5 missing days (32 unique choice sequences), and the importance sampling method is used when the number of missing days is 10 (1024 unique choice sequences) and above. The choice sequence set is generated by simulating the missing choices using random numbers following the IBL model. If a choice sequence for a given traveler is drawn twice, the second draw is discarded, since the set must contain unique choice sequences. For 10 and 15 missing days, the choice sequence is sampled 1000 and 2000 times respectively, and 20 and 100 high probability sequences are used in the estimation

*http://www.mghpcc.org

TABLE 2 : Monte Carlo experimentation results

# of missing obs	Para-meter	No correction					MSLrs (2000 draws)					MSLis					
		Average	Percent error	p-value	Empirical coverage	Run-time (sec)	Average	Percent error	p-value	Empirical coverage	Run-time (min)	Average	Percent error	p-value	Empirical coverage	Run-time (min)	Sampling method
1	β_{time}	-0.339	15.2	<1e-05	59	12.9	-0.407	1.89	<1e-05	93	278	-0.403	0.102	<1e-05	96	0.541	Complete enumeration: 2 choice sequences
	β_{cost}	-1.17	2.16	<1e-05	91		-1.21	0.899	0.00005	94		-1.21	1.67	<1e-05	89		
	VOT	0.290	12.9	<1e-05	66		0.337	0.991	0.00123	93		0.332	0.232	0.676	97		
2	β_{time}	-0.313	21.7	<1e-05	23	12.4	-0.411	2.76	<1e-05	90	299	-0.406	1.72	<1e-05	94	0.952	Complete enumeration: 4 choice sequences
	β_{cost}	-1.17	2.87	<1e-05	93		-1.22	1.27	<1e-05	90		-1.22	2.05	<1e-05	84		
	VOT	0.269	19.3	<1e-05	31		0.338	1.48	0.00003	90		0.332	0.296	0.447	95		
3	β_{time}	-0.296	29.5	<1e-05	7	12.2	-0.416	3.90	<1e-05	82	320	-0.410	2.44	<1e-05	93	1.87	Complete enumeration: 8 choice sequences
	β_{cost}	-1.17	2.84	<1e-05	90		-1.22	1.68	<1e-05	86		-1.233	2.76	<1e-05	84		
	VOT	0.254	27.4	<1e-05	13		0.341	2.21	<1e-05	90		0.332	0.271	0.549	96		
4	β_{time}	-0.281	29.5	<1e-05	2	11.8	-0.420	5.01	<1e-05	75	332	-0.413	3.15	<1e-05	90	2.85	Complete enumeration: 16 choice sequences
	β_{cost}	-1.17	2.84	<1e-05	91		-1.23	2.20	<1e-05	88		-1.24	3.58	<1e-05	74		
	VOT	0.242	27.4	<1e-05	2		0.343	2.77	<1e-05	88		0.332	0.363	0.476	97		
5	β_{time}	-0.273	31.9	<1e-05	1	11.8	-0.424	6.10	<1e-05	76	345	-0.414	3.42	<1e-05	92	6.70	Complete enumeration: 32 choice sequences
	β_{cost}	-1.17	2.31	<1e-05	92		-1.23	2.48	<1e-05	85		-1.25	4.12	<1e-05	69		
	VOT	0.233	30.2	<1e-05	3		0.345	3.56	<1e-05	88		0.331	0.675	0.220	94		
10	β_{time}	-0.241	39.8	<1e-05	0	9.18	-0.436	8.89	<1e-05	71	744	-0.402	0.607	0.216	94	23.3	Importance sampling sampling size:1000 choice sequence:20
	β_{cost}	-1.18	1.39	0.00303	97		-1.25	3.79	<1e-05	82		-1.21	0.743	0.00613	91		
	VOT	0.204	38.9	<1e-05	0		0.350	4.92	<1e-05	85		0.333	0.111	0.809	95		
15	β_{time}	-0.216	46.0	<1e-05	0	7.56	-0.444	11.1	<1e-05	66	1209	-0.396	1.05	0.0503	93	149	importance sampling sampling size:2000 choice sequences:100
	β_{cost}	-1.18	1.58	<1e-05	96		-1.27	6.12	<1e-05	79		-1.20	0.338	0.253	96		
	VOT	0.183	45.1	<1e-05	0		0.354	6.27	<1e-05	83		0.331	0.689	0.168	95		

* Estimation results are based on 100 repetitions. The nominal value of empirical coverage is 95%. P-values are calculated against true values. A p-value greater than 0.05 indicates no statistical difference. Runtime is the average of one repetition.

procedure. It should be noted that the setting does not necessarily mean that it will be enough for a general case, and the choice of sampling size and number of high probability choice sequences shall depend on the specific setting of the problem. The impact of number of high probability choice sequences on the effectiveness of the correction methods is preliminarily investigated in Section 5.2.2. After correction, the percent error of VOT is consistently below 1% and the empirical coverage of VOT is almost always above the nomial value of 95%. The p-values (>0.05) suggest that the null hypothesis that the estimator is statistically equal to the true value is not rejected. Thus, for the problem at hand, the curse of dimensioanlity issue is satisfactorily resolved.

In the experimentation, the runtime of the MSLis is significantly shorter than that of the MSLrs. For the MSLis, since the full choice sequence set grows exponentially in the number of missing observations, the sampling size and number of high probability sequences required to statistically retrieve the true value of VOT is also expected to grow rapidly. Therefore, the runtime of larger numbers of missing observaitons is significantly longer but still 10 times smaller than that of the MSLrs, which also does not fully recover the population parameters.

The box-plots of VOT in Fig.1 show the sampling distributions of VOT before and after the corrections with 10 missing initial observations. The red diamonds are the means of the estimators. It is shown that the population value of the VOT (0.333) is not covered by any point of the whole empirical distribution of the model without correction, not even by its outliers. The result is substantially improved after the MSLrs correction, in that not only the mean and median are much closer to the true population value but also the population value falls within the upper and lower (25%) quartiles. After the MSLis correction, the mean and median of the estimators are almost equal to the true population value and the population value falls within the upper and lower (25%) quartiles, confirming again that the proposed MSLis can retrieve the population parameters.

5.2 Sensitivity analysis to other simulation assumptions

5.2.1 Sampling size in random sampling

In Section 5.1, the choice sequence is sampled 2000 times for the MSLrs method. Fig.2 investigates the impact of sampling size on the percent error of VOT using 500, 1000, and 2000 draws. The estimators based on 2000 draws are generally better than those based on 500 and 1000 draws, but the improvement is not significant. Theoretically, as the sampling size increases, the quality of the estimators should also increase, however, this can be very computationally expensive.

5.2.2 Number of high probability choice sequences in importance sampling

For the MSLis method, the impacts of the size of high probability choice sequence set on the percent error of VOT and runtime are investigated for 10 missing observations. In this experiment, the choice sequence is sampled 1000 times to represent an adequate sampling size. In Fig. 3, as the number of high probability sequences increases from 2 to 50, the percent error decreases from close to 5% to below 1%. When the number of high probability sequences is greater than 20, the percent error increases slightly. The hypothesis is that the inclusion of the choice sequences with very low probability of occurrence may cause numerial issues in the estimation procedure. It should be noted that not

FIGURE 1 : Box-plots of VOT before and after corrections with 10 missing intial observations.

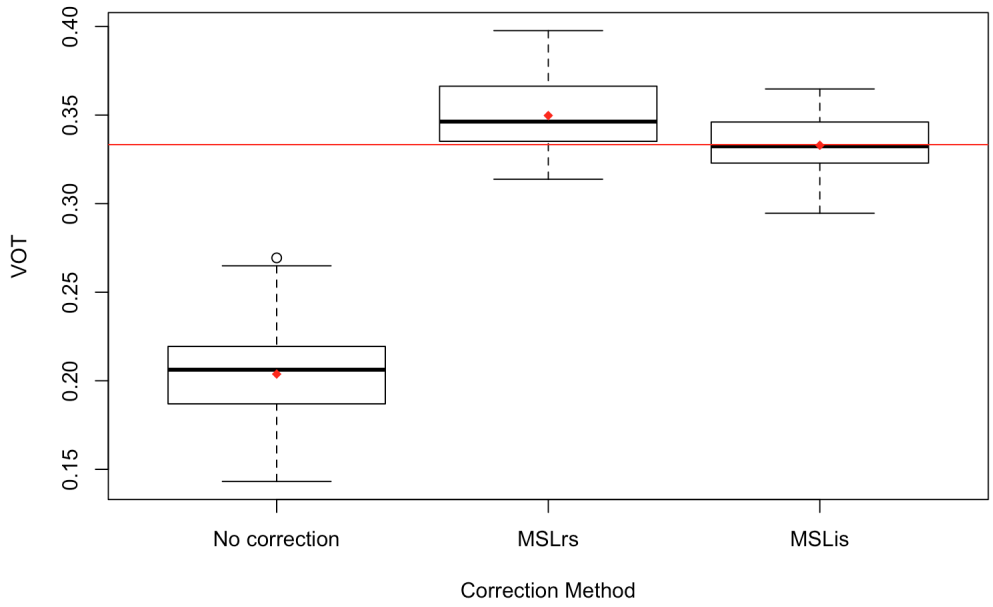

FIGURE 2 : Impact of Number of Draws on Percent Error of VOT.

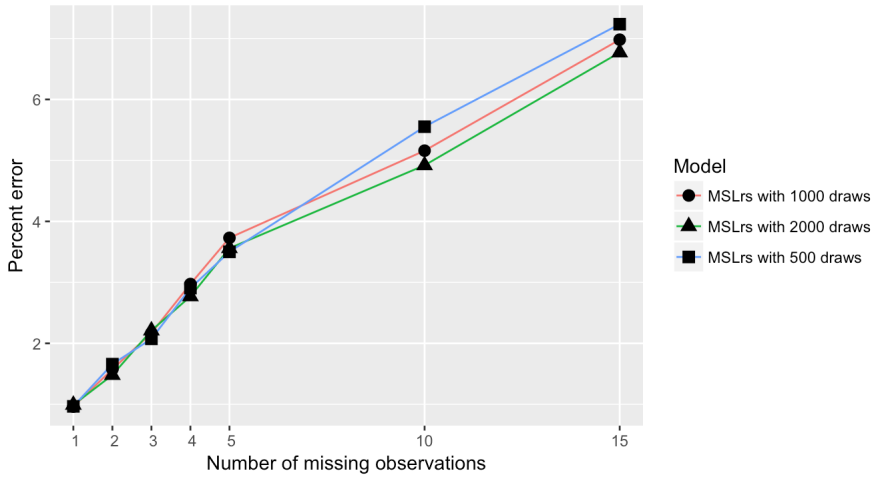

FIGURE 3 : Impact of Number of High Probability Choice Sequences on Percent Error of VOT and Runtime with 10 Missing Observations.

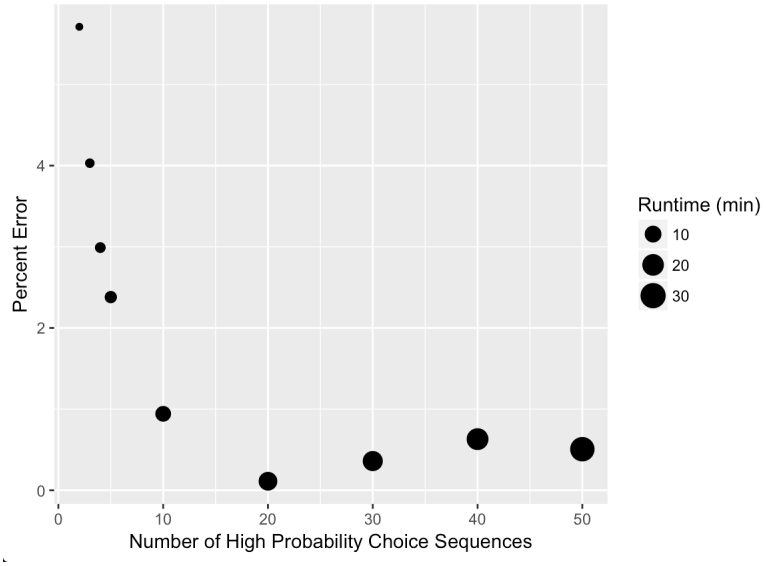

TABLE 3 : Hypothetical travel time scenarios of the experimental dataset

Scenario	Travel time ranges (minutes)	
	Route F - 25 min.	Route S - 30 min.
Fast & Safe	± 5	± 15
Fast & Risky	± 15	± 5
Low-Risk	± 5	± 5

all numbers of high probability choice sequences can statistically retrieve (i.e., p-value>0.05) the true VOT value. The runtime increases with the number of high probability sequences. When computational efficiency is a major concern, it is recommended to reduce the number of high probability sequences for large number of missing observations.

5.3 Computational experiments based on empirical data

To confirm the applicability of the proposed methods, the IBL model is estimated using the experimental dataset described in Ben-Elia and Shiftan (2010). For illustrative purpose, only the data of the informed group is used for the estimations. In the experiment, twenty-four participants were faced with three scenarios of binary route-choice as presented in Table 3. A small degree of variation was programmed (± 5 or ± 15 min around the mean) to simulate a simple variable message sign (VMS). Each scenario included 100 choices so in total each participant completed 300 trials. For each choice situation the participants received real-time information about the travel time range (the minimum and maximum travel times) for each of the two routes. Following the choice, a feedback was received regarding the "actual" travel time on the chosen route but not of the un-chosen one. This travel time was randomly drawn from the distribution of the travel time range.

A simplified version of the IBL model developed in Tang and Gao (conditionally accepted) is specified. Eq.(15) shows the utility functions for the two paths. On a given day t, T_{SLOW}, T_{FAST} are the perceived travel times for the two paths respectively and are non-linear functions of the decay parameter d. Note that d is estimated in the experimental dataset, as opposed to fixed in the Monte Carlo tests. The cumulative weighted average (CWA) of the preceding choices is used to reflect travelers' trends to repeat past choices. See Ben-Elia and Shiftan (2010) for specification. Sensitivity to variability of the travel times is represented using dummy variables for the scenarios, LRISK for Low-Risk and FRISKY for Fast & Risky. Although the model is a simplification of that developed in Tang and Gao (conditionally accepted), the perceived travel time and CWA of the preceding choices that have complete history dependency are kept in the model to assess the proposed correction methods.

TABLE 4 : Estimation results based on empirical data

Experiment	Metric	d	β_{TIMEF}	β_{TIMES}	β_{LRISK}	β_{FRISKY}	β_{CWA}
Full dataset	Estimate	1.28	-0.198	-0.086	0.864	0.405	5.71
	Std. error	0.137	0.015	0.013	0.144	0.120	0.192
	t-test (against 0)	9.31	-13.2	-6.62	6.00	3.38	29.7
10 missing observations, no correction	Estimate	1.19	-0.205	-0.0697	0.805	0.276	6.74
	Std. error	0.140	0.0176	0.0156	0.168	0.132	0.245
	Hausman-McFadden test	22.7					
10 missing observations, MSLrs correction	Estimate	1.180	-0.212	-0.0843	0.782	0.300	6.35
	Std. error	0.129	0.016	0.014	0.152	0.126	0.227
	Hausman-Mcfadden test	9.96					
10 missing observations, MSLis correction	Estimate	1.20	-0.196	-0.0852	0.842	0.312	6.24
	Std. error	0.149	0.0144	0.0164	0.213	0.161	0.213
	Hausman-McFadden test	8.56					

$$\begin{cases} U_{SLOW}(t) = \beta_{TIMES} T_{SLOW}(t) \\ U_{FAST}(t) = \beta_{TIMEF} T_{FAST}(t) + \beta_{LRISK} LRISK(t) + \beta_{FRISKY} FRISKY(t) + \beta_{CWA} CWA(t) \end{cases} \tag{15}$$

The cutoff dataset is generated by removing the first 10 observations for each participant. The IBL model is estimated using the full dataset and the estimates are assumed to be the "true" parameter values. The cutoff dataset is used to estimate the IBL model before correction and after the MSLrs and MSLis correction methods. The sampling size for the MSLrs method is 1000, and the sampling size and high probability sequences are set to 1000 and 20 respectively for the MSLis method. With the correction methods applied, the estimates obtained from the cutoff dataset are used as the priors to mimic real-life practice. The Hausman-McFadden test (Hausman and McFadden, 1984) is used to exam whether the estimators of the cutoff dataset are statistically equal to the estimators of the full dataset. Table 4 presents the estimation results. For the full dataset, all estimates are statistically significant according to the t-test against 0. For the cutoff dataset, before correction the null hypothesis of Hausman-McFadden test that the estimators are statistically equal to the estimators of the full dataset model is rejected (95% confidence, degree of freedom of 6, and critical value of 12.59). After applying the correction methods, the null hypothesis of the Hausman-McFadden test is accepted, meaning the estimators are statistically equal to the estimators of the full dataset model. The difference between the estimators of the curtailed model and corrected models is expected to be larger as the number of missing initial observations increases. Finally, note that the finite sample bias of the MSLis correction is notably smaller than that of the curtailed and the MSLrs models.

6. CONCLUSIONS AND FUTURE DIRECTIONS

Learning-based models that capture travelers' day-to-day learning process in repeated travel choice can suffer from the common problem of missing initial observations in longitudinal data collection that leads to inconsistent estimate of the perceived value of the attribute in question, and thus inconsistent parameter estimates. In this paper the MSL with two sampling methods is developed and assessed to address the endogeneity problem due to missing initial observations in learning models with complete history dependency. An IBL model in recent literature is used for its capability of precisely capturing travelers' learning process in repeated choice and model complexity.

Monte Carlo experimentation based on synthetic data shows that the proposed method drastically reduces the finite sample bias of the estimators compared to the curtailed model. For the MSLrs method, a size distortion that reflects in p-values against the true VOT value is detected, which suggests the inefficiency of the sampling method makes the method suffer from the curse-of-dimensionality problem. In contrast, the MSLis method can retrieve the true VOT value. Moreover, the computational efficiency of the MSLis is significantly better than the MSLrs method. The impacts of the sampling size in the MSLrs method and number of high probability choice sequences in MSLis are investigated. Empirical results suggest that when the number of missing observations is large, the number of high probability sequences in MSLis should be limited for computational efficiency. The two methods are also applied to empirical data to demonstrate their applicabilities. Estimation results show that the estimators after correction are statistically equal to the estimators of the full dataset model.

The research can be extended in the following directions. First, since the runtime of the MSLis increases

as the number of missing observations grows, we would like to investigate the possibility of limiting the number of missing initial observations to be simulated. Due to the nature of the model that more recent and frequent outcomes take larger weights in memory, only the omission of recent instances will cause estimation biases for practical purposes. Thus, the hypothesis is that only a certain number of unobserved instances prior to the first observation needs to be simulated to improve the estimators up to a desired threshold. Second, we would like to explore alternative correction methods. For example, the Multiple Imputation (MI) principle proposed by Little and Rubin (1987) can be used to develop a correction method using importance sampling. In this method, instead of simulating the likelihood as with MSL, each simulated choice sequence is used to estimate the model parameters via maximum likelihood estimation. The vectors of estimators obtained from all choice sequences are used to build the sampling distribution of the estimators with standard complete-data methods. Other alternative methods similar to the method proposed by Guevara and Ben-Akiva (2013a,b) may also be explored.

ACKNOWLEDGMENTS

This research was partially funded by project FONDECYT N°1150590 and the Complex Engineering Systems Institute, ISCI (ICM-FIC: P05-004-F, CONICYT: FB0816), and the US Department of Transportation through the New England University Transportation Center. The authors thank Eran Ben-Elia for providing the route-choice experimental data.

REFERENCES

Anderson, J. R. and Schooler, L. J. (1991). Reflections of the environment in memory, *Psychological Science* **2**(6): 396–408.

Ben-Elia, E. and Shiftan, Y. (2010). Which road do I take? A learning-based model of route-choice behavior with real-time information, *Transportation Research Part A* **44**: 249–264.

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium, *Econometrica* **63**(4): 841–890.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models, *Journal of Econometrics* **87**(1): 115–143.

Brownstone, D. (1991). Multiple imputations for linear regression models, *University of California Transportation Center* .

Carro, J. (2007). Estimating dynamic panel data discrete choice models with fixed effects, *Journal of Econometrics* **140**(2): 503–528.

Fernández-Antolín, A., Guevara, C. A., de Lapparent, M. and Bierlaire, M. (2016). Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data, *Journal of Choice Modelling* **20**: 1–15.

Gonzalez, C., Lerch, J. F. and Lebiere, C. (2003). Instance-based learning in dynamic decision making, *Cognitive Science* **27**: 591–635.

Guevara, C. A. (2010). *Endogeneity and Sampling of Alternatives in Spatial Choice Models (Doctoral dissertation)*, PhD thesis, MIT.

Guevara, C. A. and Ben-Akiva, M. E. (2012). Change of scale and forecasting with the control-function method in logit models, *Transportation Science* **46**(3): 425–437.

Guevara, C. A. and Ben-Akiva, M. E. (2013a). Sampling of alternatives in Logit Mixture models, *Transportation Research Part B: Methodological* **58**: 185–198.

Guevara, C. A. and Ben-Akiva, M. E. (2013b). Sampling of alternatives in multivariate extreme value (MEV) models, *Transportation Research Part B: Methodological* **48**: 31–52.

Guevara, C. A. and Polanco, D. (2016). Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution, *Transportmetrica A: Transport Science* **12**(5): 458–478.

Hausman, J. and McFadden, D. (1984). A specification test for the multinomial logit model, *Econometrica* **46**: 1219–1240.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system, *Technical report*, National Bureau of Economic Research.

Heckman, J. (1981). *Structural Analysis of Discrete Data and Econometric Applications*, Cambridge: The MIT Press, chapter The incidental parameters problem and the problem of initial condition in estimating a discrete time-discrete data stochastic process, pp. 179–195.

Honore, B. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables, *Econometrica* **74**: 611–629.

Hyndman, R. J., Bashtannyk, D. M. and Grunwald, G. K. (1996). Estimating and visualizing conditional densities, *Journal of Computational and Graphical Statistics* **5**(4): 315–336.

Lee, L. (1997). Simulated maximum likelihood estimation of dynamic discrete choice statistical models: some Monte Carlo results, *Journal of Econometrics* **82**(1): 1–35.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York" John Wiley and Sons.

Lu, X., Gao, S., Ben-Elia, E. and Pothering, R. (2014). Travelers' day-to-day route choice behavior with real-time information in a congested risky network, *Mathematical Population Studies* **21**: 205–219.

McFadden, D. and Train, K. (2000). Mixed MNL models for discrete response, *Journal of applied Econometrics* pp. 447–470.

Newell, A. and Rosenbloom, P. (1981). *Mechanisms of skill acquisition and the law of practice, In J.R. Anderson*, Hillsdale, NJ: Earlbaum, chapter Cognitive skills and their acquisition, pp. 1 – 55.

Rosenblatt, M. (1969). Conditional probability density and regression estimators, *Multivariate analysis II* **25**: 31.

Rubin, D. and Wenzel, A. (1996). One hundred years of forgetting: A quantitative description of retention, *Psychological Review* **103**: 734 – 760.

Schenker, N. and Welsh, A. (1988). Asymptotic results for multiple imputation, *Annals of Statistics* **16**: 1550 – 1566.

Tang, Y. and Gao, S. (conditionally accepted). An exploratory study of instance-based learning for route choice with random travel times, *Journal of Choice Modelling. Under 2nd review* .

Train, K. (2009). *Discrete Choice Methods with Simulation*, Cambridge University Press, chapter 10.

Wickelgren, W. (1976). *Handbook of Learning and Cognitive Processes*, Potomac, Maryland: Lawrence Erlbaum Associates.

Wooldridge, J. (2005). Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity, *Journal of Applied Econometrics* **20**(1): 39–54.