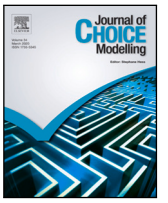
Journal of Choice Modelling 50 (2024) 100467

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm



Research note

Resampling estimation of discrete choice models

Nicola Ortelli [a,b,*], Matthieu de Lapparent [a], Michel Bierlaire [b]

[a] School of Management and Engineering Vaud, HES-SO, University of Applied Sciences and Arts Western Switzerland, Switzerland
[b] Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ARTICLE INFO

Keywords:
Discrete choice models
Maximum likelihood estimation
Dataset reduction
Sample size
Locality-sensitive hashing

ABSTRACT

In the context of discrete choice modeling, the extraction of potential behavioral insights from large datasets is often limited by the poor scalability of maximum likelihood estimation. This paper proposes a simple and fast dataset-reduction method that is specifically designed to preserve the richness of observations originally present in a dataset, while reducing the computational complexity of the estimation process. Our approach, called LSH-DR, leverages locality-sensitive hashing to create homogeneous clusters, from which representative observations are then sampled and weighted. We demonstrate the efficacy of our approach by applying it on a real-world mode choice dataset: the obtained results show that the samples generated by LSH-DR allow for substantial savings in estimation time while preserving estimation efficiency at little cost.

1. Introduction

The technological advancements of the past decades have allowed transforming an increasing part of our daily actions and decisions into storable data. Specifically, the rise of digital communication has led to a radical change in the scale and scope of available data in relation to virtually any object of interest. In the field of discrete choice analysis, such abundance of data has the potential to significantly expand our understanding of human behavior, but this prospect is limited by the poor scalability of discrete choice models (DCMs).

Indeed, when estimating DCMs, the use of ever-larger datasets raises two issues: (i) the number of possible model specifications exponentially grows with the number of candidate explanatory variables, implying that analysts must spend more time searching for appropriate specifications; and (ii) the computational cost of maximum likelihood estimation increases with the number of observations, quickly becoming intractable for any advanced model structure. While the first issue has spurred great interest,[1] the second has received much less attention: in order to deal with the increased computational cost associated with large datasets, effort has mostly been dedicated to improving the optimization methods used to estimate DCMs (Lederrey et al., 2021; Rodrigues, 2022) or to enhancing their implementation (Molloy et al., 2021; Arteaga et al., 2022).[2]

This study moves beyond traditional techniques by exploring a less common approach, which consists in reducing the size of datasets by sampling. Because the most frequently used algorithms for maximum likelihood estimation compute the log likelihood

* Corresponding author at: School of Management and Engineering Vaud, HES-SO, University of Applied Sciences and Arts Western Switzerland, Switzerland.
E-mail addresses: nicola.ortelli@heig-vd.ch, nicola.ortelli@epfl.ch (N. Ortelli), matthieu.delapparent@heig-vd.ch (M. de Lapparent), michel.bierlaire@epfl.ch (M. Bierlaire).

[1] The recent literature is rich in studies that seek to mitigate the need for presumptive structural assumptions in DCMs. We refer the reader to van Cranenburgh et al. (2021) for a review and discussion.

[2] Another important computational challenge related to modern-day datasets arises when those include numerous choice alternatives. While this topic falls out of the scope of this paper, interested readers may consult Guevara and Ben-Akiva (2013a,b), Bierlaire and Krueger (2020) and Tsoleridis et al. (2022).

https://doi.org/10.1016/j.jocm.2023.100467
Received 30 March 2023; Received in revised form 19 December 2023; Accepted 20 December 2023
Available online 3 January 2024

function and its gradient across the whole dataset at each iteration, considering fewer observations effectively reduces their computational burden. Removing observations from a dataset is usually advised against by choice modelers, but has nevertheless become common practice when training machine learning models on large amounts of data: given the iterative nature of model specification, the use of a smaller subsample allows for early modeling decisions to be taken significantly faster (Park et al., 2019). Moreover, in the case of maximum likelihood estimation for DCMs, subsamples may additionally be used to obtain good starting values for estimation on the whole dataset after the definitive model specification is reached.

We propose a fast dataset-reduction technique that is designed to alter the parameter estimates of the model of interest as little as possible. We diverge from the common premise that all datasets contain some fraction of less relevant observations; instead, our method aims at preserving the diversity of observations originally present in the dataset, while reducing its size. Our proposed approach leverages locality-sensitive hashing to create clusters of similar observations, from which "representative" observations are sampled. The observations obtained in such way are then given weights that are proportional to the sizes of the clusters they represent, so as to mimic the full dataset during the model estimation process. As argued in the following sections, we believe that a carefully selected *and weighted* subsample of observations is capable of providing close-to-identical estimation results while being, by definition, less computationally demanding. The source code is currently being consolidated and integrated into the Biogeme software (Bierlaire, 2023).

The remainder of this document is organized as follows: Section 2 provides an overview of the pertinent literature; Section 3 introduces the concept of locality-sensitive hashing and then proceeds to describe our proposed algorithm; Section 4 presents and discusses the results obtained by applying our method to a real-world mode choice dataset; finally, Section 5 summarizes the findings of this study and identifies the future steps of this research.

2. Literature review

Instance selection and prototype generation are dataset-reduction tasks that consist in producing a smaller representative set of data points from a given dataset, respectively by sampling said points or by creating new, artificial ones. While shrinking datasets, instance selection and prototype generation techniques typically aim at minimizing information or performance loss by discarding redundant data points; as such, they have been shown to be particularly beneficial to instance-based methods whose performances rely on specific data points, such as support vector machines or k-nearest neighbors algorithms (Olvera-López et al., 2010; Alexandropoulos et al., 2019).

The recent literature offers a variety of instance selection and prototype generation methods. Among those, a prevalent approach consists in using clustering algorithms to identify groups of similar data points, from which some are either sampled or merged into prototypes. Methods based on k-means and its variations are particularly popular, despite becoming computationally heavy when applied to large datasets (Ougiaroglou and Evangelidis, 2016; Ren and Yang, 2019; Castellanos et al., 2021; Chang et al., 2021; Ougiaroglou et al., 2021; Saha et al., 2022). To circumvent this limitation, another stream of research makes use of locality-sensitive hashing (LSH) to cluster similar data points together (Arnaiz-González et al., 2016; Aslani and Seipel, 2020; Zhang and Liu, 2023). While k-means is known to be superior in terms of accuracy and reliability, LSH is intrinsically faster because its complexity is linear in the number of data points to be hashed (Paulevé et al., 2010). This aspect is crucial when instance selection and prototype generation methods are used to reduce the computational burden of model training; we therefore deem techniques based on LSH as particularly promising.

To the best of our knowledge, there have only been two attempts at developing dataset-reduction methods in the context of discrete choice modeling. We explain this scarcity by the fact that DCMs are generally used to extract behavioral insights from data; any dataset-reduction technique applied prior to estimating a DCM therefore needs to preserve a certain number of characteristics of the full dataset in the smaller sample, as it could otherwise lead to erroneous or biased estimation results. In contrast, the machine learning paradigm solely focuses on maximizing the model's predictive accuracy — or on minimizing any loss due to dataset reduction — without taking into account any aspect related to the data.

The earliest instance selection method for DCMs we could find is proposed by van Cranenburgh and Bliemer (2019): their method scales any dataset down to a predefined fraction of its size while iteratively minimizing an estimate of the D-error, obtained by means of a simplified version of the model of interest.[3] In doing so, they seek to guarantee that the model parameters are estimated as precisely as possible on a sample that is much smaller than the full dataset, but in reality, this only encourages their algorithm to keep observations that are similar among them. As a result, the obtained samples may not be representative of the full dataset and therefore lead to biased parameter estimates. The second direct precedent of this study is described by Schmid et al. (2022) as a pre-processing step applied to a very large dataset. The method consists in dividing the dataset into clusters using k-means; a single observation is then sampled from each cluster and weighted according to the cluster size. While this approach is capable of preserving the characteristics of the full dataset thanks to the weighting scheme it employs, it still suffers from the fact that k-means is computationally heavy, which severely limits its usage.

In this paper, we follow an approach similar to the one presented in Schmid et al. (2022), but based on LSH. Our main contribution is a fast dataset-reduction technique that relies on clustering to sample and weight observations. Jointly, the sampling strategy and weighting schemes used by our method guarantee that the characteristics of the full dataset are preserved. Indeed, our experiments show that the generated subsamples lead to the same behavioral findings as the full dataset, while the computational complexity of model estimation is effectively reduced.

[3] The D-error statistic is a measure of efficiency commonly used in experimental design. It is defined as the determinant of the asymptotic variance–covariance matrix of the estimated model parameters.

3. Methodology

3.1. Intuition

Consider a choice dataset of N observations (x_n, i_n), each consisting of a vector x_n of explanatory variables associated with individual n, together with the observed choice i_n of that same individual among J alternatives. In its simplest form, a discrete choice model $P(i \mid x_n; \theta)$ calculates the probability that individual n chooses any alternative i as a function of x_n and θ, where θ is a vector of model parameters to be estimated from the data.

The values of the model parameters are typically determined through maximum likelihood estimation, which consists in finding the values of θ that maximize the joint probability of replicating all observed choices in the dataset. In practice, the logarithm of the likelihood is usually maximized instead, for numerical reasons. The *log likelihood* function is therefore defined as

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log P\left(i_n \mid x_n; \theta\right). \tag{1}$$

Let us now assume that the dataset contains some observations that are identical in all explanatory variables and in the observed choice. By gathering the observations into $G < N$ groups of identical observations, we may rewrite (1) as

$$\mathcal{L}(\theta) = \sum_{g=1}^{G} N_g \cdot \log P\left(i_g \mid x_g; \theta\right), \tag{2}$$

where N_g denotes the size of group g, and i_g and x_g are the observed choice and explanatory variables shared by all observations in group g, respectively. By definition, the G groups are a partition of the full dataset and therefore verify

$$\sum_{g=1}^{G} N_g = N. \tag{3}$$

(1) and (2) are equivalent and, as such, yield the exact same parameter estimates when maximized. However, since $G < N$, the computational cost of evaluating (2) is smaller, by a ratio of approximately G/N.[4] This is empirically shown in Section 4, Experiment B.

The idea behind our dataset-reduction method is to extend this factorization trick to observations that are nearly identical. In other words, by clustering together not only duplicates, but also "very similar" observations, the intent of our method is to further decrease the number of distinct groups and, in doing so, to effectively reduce the computational time associated with evaluating the log likelihood function and its gradient. Of course, this comes at the cost of degrading the estimation results because part of the information contained in the dataset is lost; still, the use of an adequate clustering scheme limits said degradation while granting our method a certain reliability. The clustering technique designed for this purpose is inspired from locality-sensitive hashing (LSH), which we introduce now.

3.2. Locality-sensitive hashing

LSH is an efficient method for finding similar items in data. As opposed to conventional hashing functions, which allocate items to unique encrypted outputs, LSH seeks to gather "similar" items into clusters — or *buckets*. It achieves this goal by combining the outcomes of several hashing functions, designed in such way that pairs of items are more likely to be hashed to the same bucket if they are close to each other in their original space than if they are far apart. As mentioned earlier, the main advantage of LSH over other clustering techniques is that its computational complexity is linear in the number of items to be hashed.

A *family* of LSH functions $\mathcal{H} = \{h : (M, d) \to \mathbb{Z}\}$ is a collection of functions h that map elements of a metric space (M, d) onto the set of integers $\mathbb{Z}$ (Leskovec et al., 2020). Each integer represents a different bucket, and two data points x_p and x_q belong to the same bucket of function h if and only if $h(x_p) = h(x_q)$. For instance, a well-known family that is suited for Euclidean spaces is based on the function

$$h_{a,b}(x) = \left\lfloor \frac{a \cdot x + b}{w} \right\rfloor, \tag{4}$$

where $\lfloor \cdot \rfloor$ represents the floor function, a is a vector whose entries are independently drawn from a normal distribution $\mathcal{N}(0, 1)$, b is a real value chosen uniformly from the range $[0, w)$ and w is the bucket width (Datar et al., 2004). One may see (4) as a projection of all data points onto a random line whose direction is given by vector a; an offset equal to b is then added to all projected points before the line is discretized into uniform intervals of size w. All data points that fall in the same interval are therefore deemed "equivalent" — contingent on a — and are assigned to the same bucket.

Parameter w plays a crucial role in the effectiveness of LSH, but its value is context-dependent. By changing the bucket width — or discretization step — one can choose an appropriate degree of similarity between data points within buckets: a sufficiently small

[4] One could argue that multiplying $P\left(i_g \mid x_g; \theta\right)$ by N_g adds arithmetic operations that are not required in (1); still, these additional multiplications are negligible in comparison to the number of operations needed to evaluate $P\left(i_n \mid x_n; \theta\right)$ for every n, even in trivial model structures.

w only groups points that are exactly identical, whereas greater values result in fewer buckets that contain larger amounts of increasingly dissimilar points.

Another way of improving the discriminative power of LSH is to combine several hash functions. In the case of the family defined by (4), this corresponds to simultaneously projecting the data onto multiple random lines. Suppose a and b are drawn R times: now, two data points x_p and x_q belong to the same bucket if and only if they are grouped together by all R random projections, *i.e.*:

$$H_{A,B}(x_p) = H_{A,B}(x_q) \iff h_{a_r,b_r}(x_p) = h_{a_r,b_r}(x_q) \quad \forall r = 1, \dots, R, \tag{5}$$

where, for the sake of conciseness, $A = (a_1, \dots, a_R)$ and $B = (b_1, \dots, b_R)$ gather the R realizations of a and b, respectively. Increasing R reduces the joint probability that two data points are grouped together by all projections, which results in more buckets containing fewer items.

3.3. LSH-based dataset reduction (LSH-DR)

Our dataset-reduction algorithm has three main ingredients, namely: (i) an LSH function or a combination of LSH functions capable of partitioning a dataset of size N into buckets that only contain "similar" observations; (ii) a sampling strategy, according to which some observations are selected from each bucket; and (iii) a weighting scheme that assigns a weight N_g to each selected observation (x_g, i_g). The G observations obtained in such a way, together with their associated weights $N_1, \dots, N_G$, constitute the outcome of our dataset-reduction method. A model of interest may then be estimated on the obtained sample — rather than on the whole dataset — by using the log likelihood function of (2), with i_g and x_g now referring to the observed choice and explanatory variables associated with the g-th selected observation, respectively. Fig. 1 illustrates each step of the LSH-DR algorithm by means of a toy dataset.

Clustering

Our method uses the family of LSH functions introduced in (4) with, as parameters, the discretization step w and the number of projections R. Prior to hashing, all variables that are not considered by the model of interest are dropped from the dataset and the remaining variables are either normalized such that their values are between 0 and 1 or encoded using binary indicators, depending on their numerical or categorical nature.

We denote as $\mathcal{Q}_1, \dots, \mathcal{Q}_K$ the K buckets obtained via LSH. The individuals' choices are not taken into account during the hashing; rather, they are used to further partition each bucket $\mathcal{Q}_k$ into J sub-buckets $\mathcal{Q}_{k1}, \dots, \mathcal{Q}_{kJ}$ that only contain observations of the same chosen alternative, where J is the number of alternatives in the choice context. At this point, any pair of observations $\big((x_p, i_p), (x_q, i_q)\big)$ in any sub-bucket $\mathcal{Q}_{ki}$ verifies

$$H_{A,B}(x_p) = H_{A,B}(x_q) \quad \text{and} \quad i_p = i_q = i. \tag{6}$$

Sampling

The exact number of observations to be sampled from each sub-bucket $\mathcal{Q}_{ki}$ is denoted by z_{ki} and computed as

$$z_{ki} = \left\lceil \frac{|\mathcal{Q}_{ki}|}{N_{\max}} \right\rceil, \tag{7}$$

where $\lceil \cdot \rceil$ stands for the ceiling function and $N_{\max}$ is a parameter that represents the maximum weight allowed to be associated to any selected observation. This definition of z_{ki} guarantees that at least one observation is selected from each non-empty bucket.

When choosing a value for $N_{\max}$, one must keep in mind that it also sets a lower bound equal to $N/N_{\max}$ on the size of the subsamples that can be generated by the procedure, where N is the number of observations in the full dataset. For instance, if $N_{\max}$ is set to 5, LSH-DR will not be able to generate subsamples smaller than 20% of N, irrespective of the values of w and R. Still, as opposed to selecting a single observation per sub-bucket, sampling proportionally to their size helps attenuates the loss of information that would otherwise be observed in large sub-buckets; this significantly improves the overall quality of our method's results.

Weighting

Every observation (x_g, i_g) sampled from any sub-bucket $\mathcal{Q}_{ki}$ is given a weight N_g that is equal to the number of observations in $\mathcal{Q}_{ki}$, divided by the total number of observations sampled from $\mathcal{Q}_{ki}$:

$$N_g = \frac{|\mathcal{Q}_{ki}|}{z_{ki}}. \tag{8}$$

By design, the adopted sampling strategy and this weighting scheme jointly guarantee that the sum of all weights in the subsample is equal to the size of the full dataset, as in (3).

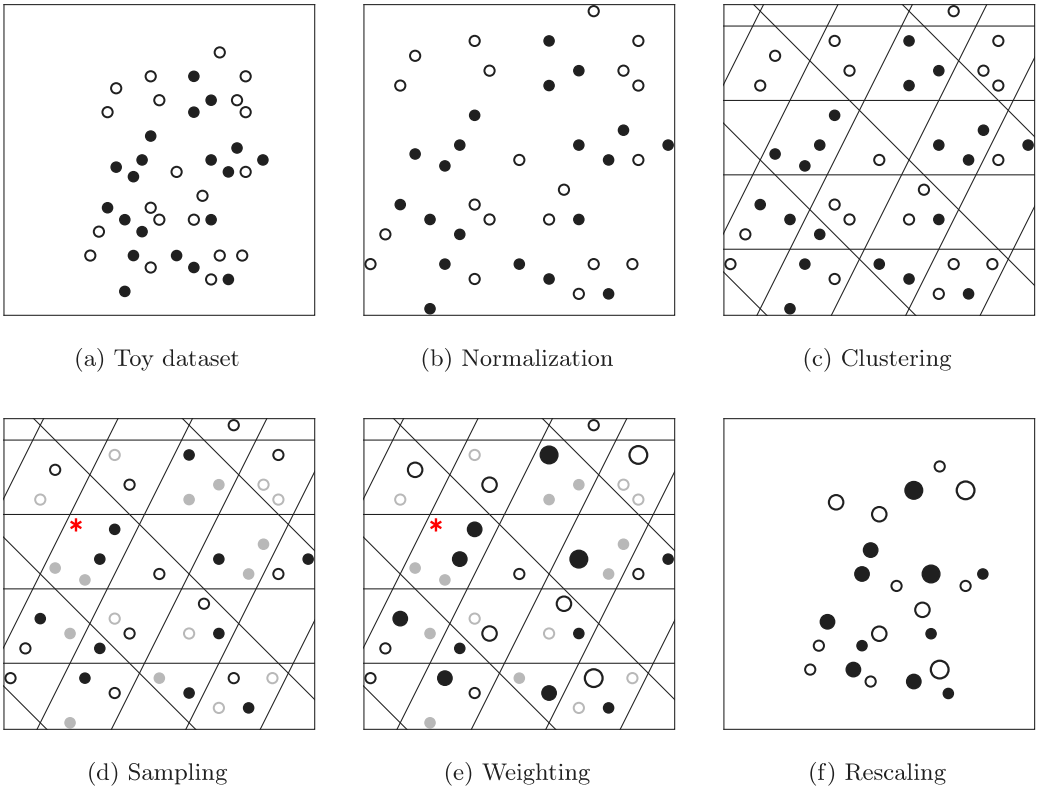


Fig. 1. Illustration of the main steps of the LSH-DR algorithm, as applied to a toy dataset, with $R = 3$ and $N_{\max} = 3$. (a) The dataset is made up of 40 observations split between two alternatives, which are represented here by the color of each dot. Only two explanatory variables are considered; they correspond to the axes of each graph. (b) The LSH-DR algorithm begins by normalizing both explanatory variables. (c) LSH is then used to hash the observations into buckets, which are further divided into sub-buckets that only contain observations of the same chosen alternative; in this representation however, the buckets and their sub-buckets perfectly overlap and are therefore indistinguishable. (d) Since $N_{\max}$ is set to 3, two observations must be sampled from the sub-bucket marked with an asterisk. A single observation is obtained from all other sub-buckets. (e) The weight assigned to each selected observation is equal to the size of its sub-bucket, except for the sub-bucket marked with an asterisk: in that case, the two sampled observations share the total weight equally. (f) The obtained subsample is converted back to the original scale of the data.

Two additional comments on the generated samples

First, it is important to note that the subsamples generated by LSH-DR are endogenously stratified. One may recall that the partition into buckets is entirely determined by the values of the explanatory variables, but since each bucket is further divided into sub-buckets based on the choices, the probability for an observation to be drawn varies both with the endogenous and the exogenous variables. From the perspective of the "classical" discrete choice modeling literature, the weights are therefore of crucial importance in the estimation process because exogenous sampling maximum likelihood (ESML) does not yield consistent parameter estimates on endogenously stratified samples (Manski and McFadden, 1981); the only exception is the logit model, for which Manski and Lerman (1977) and McFadden (1978) have shown that only the alternative-specific constants are not consistently estimated and that they can effortlessly be corrected afterwards. With this in mind, we still recommend using the weighted ESML, *even for logit models*. Admittedly, the theoretical derivations of Manski and Lerman (1977) and McFadden (1978) are based on the assumption that the model of interest is correctly specified, which is rarely the case in practice. Instead, ignoring the weights was shown to exacerbate the problems associated with misspecification and to produce severely biased parameter estimates. Fig. 9 in the Appendix uses one of the experiments presented in Section 4 to illustrate this phenomenon.

Second, we would like to emphasize that the heterogeneity of the full dataset is preserved — to the extent possible — because LSH guarantees that dissimilar observations have low chances of being hashed to the same bucket. As a result, any uncommon observation is likely to end up alone in a sub-bucket, which ensures that it is selected for inclusion in the subsample. A drawback of this approach is that the buckets do not offer any behavioral interpretation, simply because they are the result of a combination of random projections. They only represent an arbitrary way of segmenting the space in which the observations live.

4. Experiments

The efficacy of our method is demonstrated by means of a series of experiments based on the London passenger mode choice (LPMC) data (Hillel et al., 2018). The dataset consists of more than 81'000 trip records collected over three years, combined with systematically matched trip trajectories alongside their corresponding mode alternatives. Four modes are distinguished: walk, cycle, ride public transport and drive.[5] We divide the dataset into two parts: the first two years of data — 54'766 observations — are used for model estimation whilst the final year — 26'320 observations — is set aside for out-of-sample validation.

In the experiments, the data is used to train two multinomial logit models that we borrow from Hillel (2019). We refer to those as "MNL-S" and "MNL-L" and provide their specifications in Table 1 in the Appendix. The MNL-S includes 10 continuous variables and 13 associated parameters, whereas the MNL-L considers 11 continuous variables, 8 categorical variables encoded using binary indicators and 53 associated parameters. In addition, we create a nested logit model that has the exact same utility specifications as the MNL-S, but gathers the cycle, ride and drive alternatives within a nest. We refer to this third model as "NL-S". Upon inspection of the results presented in this section, one may point out that all three models yield very high values of time, but this is a known limitation of the models — whose specification we do not endorse — rather than a specific issue of our study. While these unrealistic values would normally warrant particular attention, the emphasis of these experiments remain on demonstrating the algorithm's efficacy in comparison to existing methods.

All model estimations are performed using the Biogeme package for Python (Bierlaire, 2023) on a 2.4 GHz 36-core cluster node with 256 GB of RAM. On an average laptop all computational times are approximately five times larger.[6]

Experiment A: random sampling

We begin by illustrating the relation between sample size, estimation time and quality of the results. For this purpose, we estimate the MNL-S on random subsamples of the LPMC dataset and report the following quantities: (i) the execution time, which consists of the sampling and estimation times; (ii) the normalized out-of-sample log likelihood (OSLL), *i.e.*, the log likelihood yielded by the estimated model on the validation data, normalized by the number of observations; (iii) the mean absolute percentage error (MAPE) of the parameter estimates; and (iv) the value of time for the "drive" alternative, computed as the ratio between the estimates of the parameters associated with travel time and with cost. The subsamples range from 100% to 1% of the full dataset size — *i.e.*, 54'766 — and 100 repetitions are performed at each sample size. Fig. 2 presents the obtained results by means of moving-window boxplots displaying the 5^{th}, 25^{th}, 50^{th}, $75^{\text{th}}$and 95^{th} percentiles.

The first subfigure illustrates that the model estimation time decreases linearly with the number of observations in the sample, and so does the range of values obtained across repetitions. The normalized OSLL stays reasonably close to its maximum value down to 30% of the full dataset size, but seriously declines for smaller samples. Similarly, the MAPE slowly increases until the sample size reaches approximately 20%, then degrades at a much faster pace. Finally, the median of the value of time appears to be relatively stable, but its accuracy deteriorates considerably as the sample size decreases. Overall, Fig. 2 shows that random sampling is a decent strategy when the model of interest is small; still, the next experiments show that significantly better results can be achieved with a negligible increase in execution time.

Experiment B: LSH-DR

We now move to estimating the MNL-S on samples generated by our proposed method. We apply the LSH-DR algorithm on the LPMC data 10'000 times, with $N_{\max} = 10$, $R = 4$ and w ranging from 0.02 to 0.2. Only the ten continuous variables relevant to the MNL-S are fed to LSH-DR. The obtained weighted samples range from 48'206 to 6'365 unique observations, that is, from 88% to 12% of the full dataset. The MNL-S is then estimated on these samples; the collected results are shown in Figs. 3 and 4. For comparative purposes, we also report the outcomes of the previous experiment.

Fig. 3 demonstrates that LSH-DR is capable of producing substantially better samples than random sampling, for a small increase in execution time: indeed, the samples generated by LSH-DR yield smaller MAPEs of the parameters and more accurate estimates of the value of time. Fig. 4 further shows that our method also has a beneficial effect on individual parameter estimates. As the sample sizes decrease, the precision of the estimates obtained on random samples deteriorates faster than on the samples generated by LSH-DR. One should note that the estimate of the parameter associated with rail in-vehicle time degenerates at a much faster pace than the other parameters for both sampling methods, but this is likely due to the fact that the variable is zero in almost 70% of the observations.

[5] The "drive" alternative also includes car passenger, taxi, van and motorbike.

[6] This was tested using a 2.3-GHz 8-core processor with 16 GB of RAM.

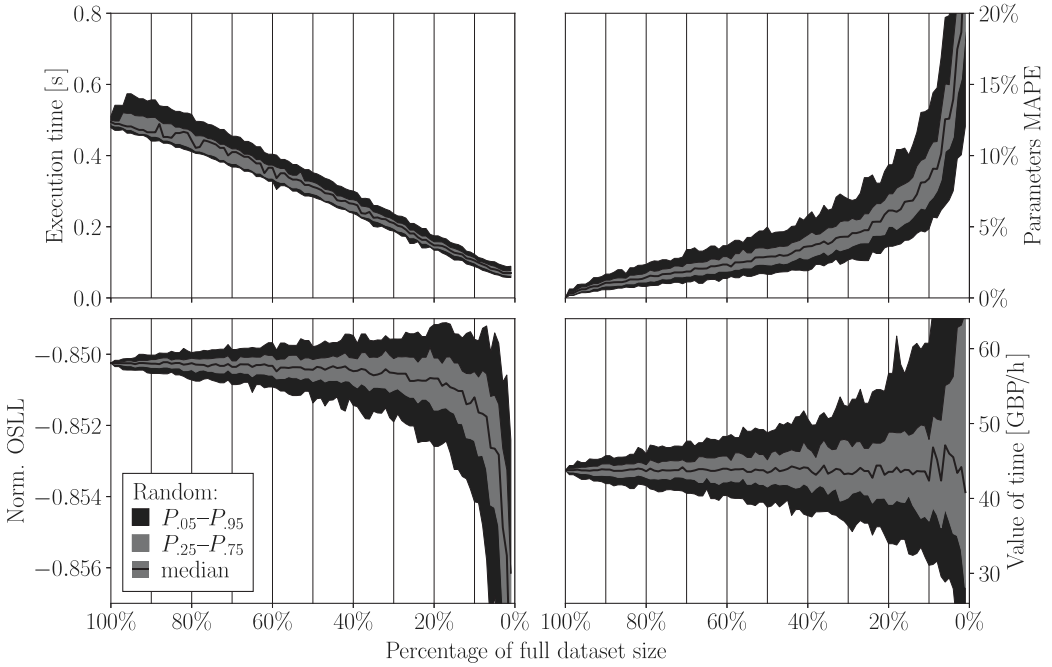


Fig. 2. Estimation of the MNL-S model on random samples of the LPMC dataset.

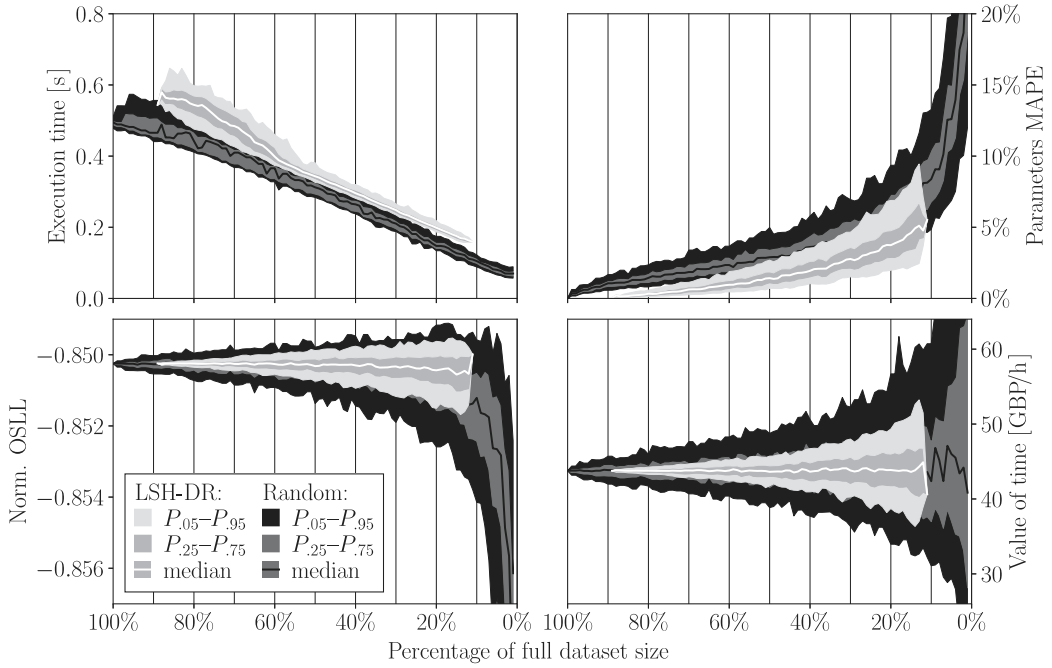


Fig. 3. Estimation of the MNL-S model on samples generated by LSH-DR. The results obtained on random samples are also reported, for comparative purposes.

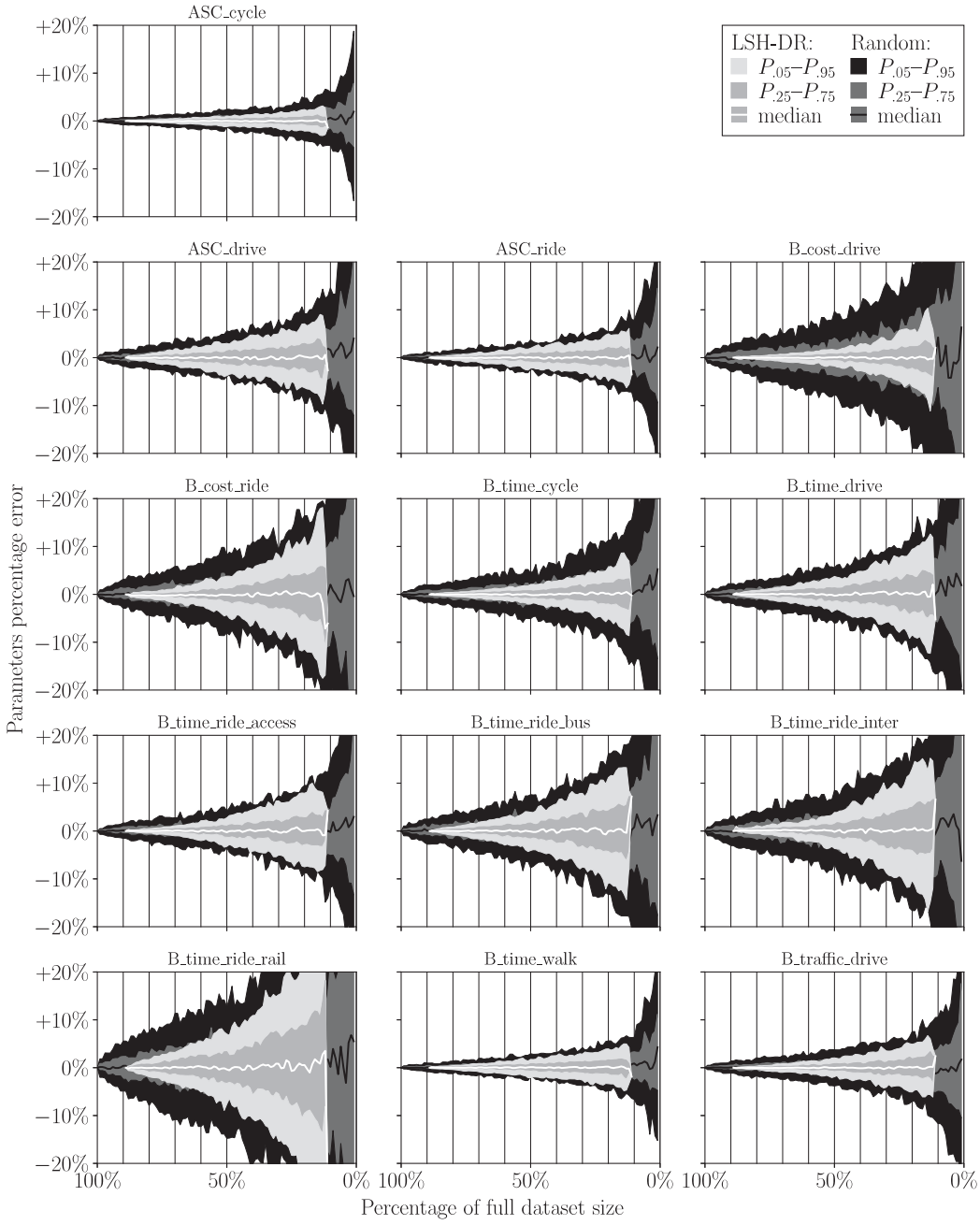


Fig. 4. Percentage error of the MNL-S parameter estimates on samples generated by LSH-DR. The results obtained on random samples are also reported, for comparative purposes.

Experiment C: comparison with state-of-the-art methods

In this experiment, we compare the performance of our method with three other dataset-reduction techniques, namely: (i) random sampling; (ii) k-means clustering;[7] and (iii) sampling of observations (SoO), as proposed by van Cranenburgh and Bliemer (2019).[8] We proceed as follows: a certain percentage of the full dataset size is chosen and we retrieve from Experiment–B the 100 samples of

[7] Our approach consists in dividing the dataset based on the choices and running k-means on each part separately. The number of clusters in each part is chosen proportionally to the number of observations it contains, such that the total number of clusters is equal to the required sample size. A single observation is then randomly selected from each cluster and associated with a weight that corresponds to the size of its cluster. As opposed to Schmid et al. (2022), who

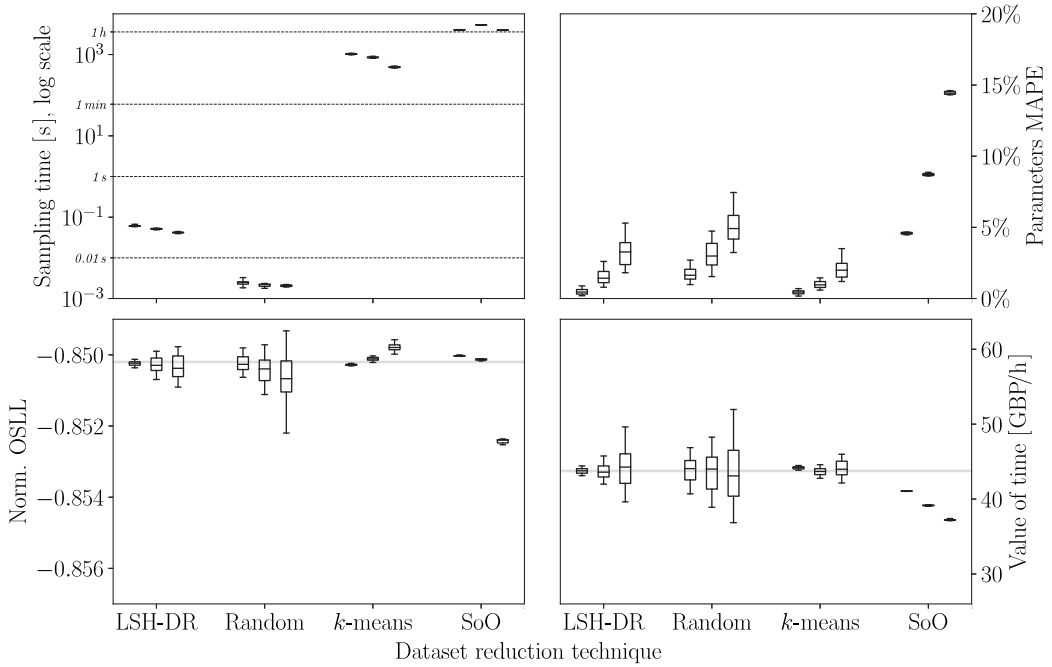


Fig. 5. Comparison of dataset-reduction techniques at approximately 75%, 50% and 25% of the full dataset size. The boxplot whiskers indicate the 5th and 95th percentiles, whereas the gray horizontal lines represent the OSLL and value of time obtained on the full dataset.

size closest to that percentage; the three benchmark dataset-reduction techniques are then used to generate samples of those exact sizes, which are finally used to train the MNL-S model. Fig. 5 reports the sampling time, normalized OSLL, parameters MAPE and value of driving time for 75%, 50% and 25% of the full dataset. The size of the samples retrieved from Experiment B ranges from 40'881 to 41'275, from 27'191 to 27'582 and from 13'542 to 13'860, respectively.

Overall, Fig. 5 illustrates that the samples producing the most accurate results are obtained via k-means. For all chosen percentages of the full dataset size, k-means generates the best samples in terms of OSLL, MAPE and value of time; only for the largest size does LSH-DR generate samples of comparable quality. Still, despite its superiority, k-means is practically unusable because of its runtime: it takes from 8 up to 23 min to obtain a sample from a relatively small dataset, *on a dedicated cluster node*. That is between 9'000 and 24'000 times longer than LSH-DR and up to 800'000 times longer than random sampling. As regards SoO, the method is shown to provide the worst results in terms of OSLL, MAPE and value of time, while also displaying the largest runtimes. This is due to the fact that SoO is designed to maximize the efficiency of the parameter estimates rather than their precision or the model's predictive accuracy.

Experiment D: more complex models

Finally, we also estimate the NL-S and MNL-L models on samples generated by LSH-DR to demonstrate that our method may also be beneficial to more complex or larger models. The NL-S can be directly estimated on the subsamples generated for Experiment B, but the MNL-L requires the LSH-DR to be run again with all additional variables that the larger model considers. To this end, our algorithm is again run 10'000 times, with $N_{\max} = 10$, $R = 4$ and w ranging from 0.1 to 1. Note that the MNL-L model includes several discrete explanatory variables: those are not treated differently by the LSH-DR algorithm. The generated samples range from 51'674 to 7'305 observations in size, that is, from 94% to 13% of the full dataset. Figs. 6 and 7 display the results obtained for the NL-S model and Fig. 8 shows those for the MNL-L model. For the sake of comparison, the results obtained on random samples are also shown on the three figures.

apply k-means only once and on their entire dataset, this approach guarantees that the market shares in the subsample match the original ones. We use the implementation of k-means available in the scikit-learn package for Python (Pedregosa et al., 2011).

[8] In SoO, we use the MNL-S model estimated on the full dataset as the sampling model. All individual Fisher information matrices are pre-computed to speed up the search. The algorithm was implemented from scratch to be compatible with the rest of our code.

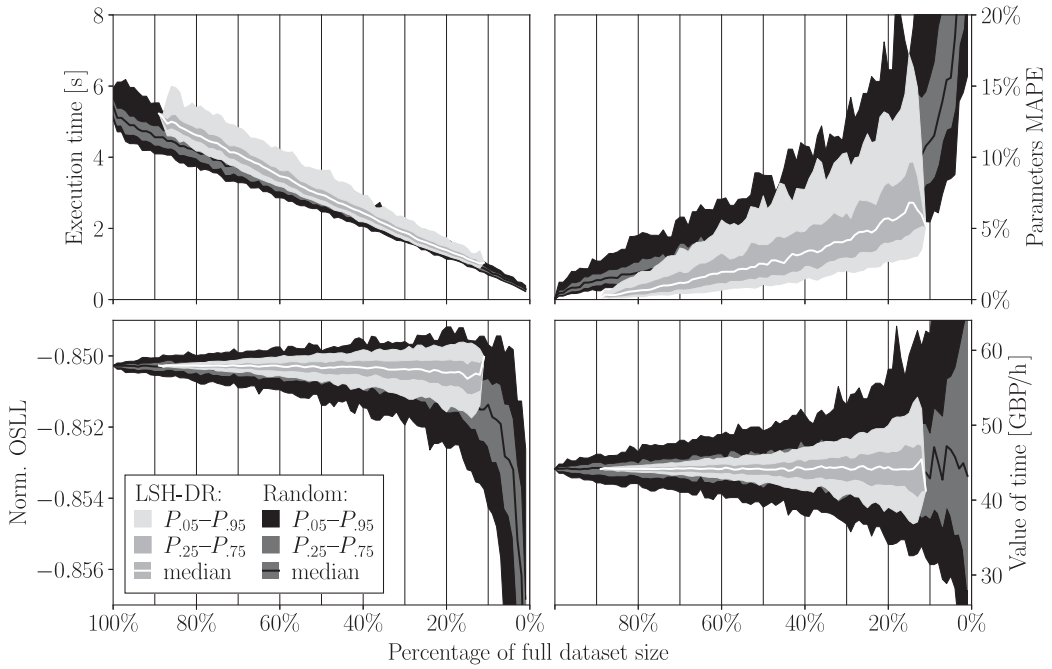


Fig. 6. Estimation of the NL-S model on samples generated by LSH-DR. The results obtained on random samples are also reported, for comparative purposes.

Closing remarks

The conducted experiments empirically demonstrate the validity and potential of our proposed method for the estimation of multinomial and nested logit models. In particular, they show that the LSH-DR algorithm outperforms random sampling for the three presented models, in exchange for a negligible increase in computational time. The algorithm is also shown to be several orders of magnitude faster than the alternative approaches proposed in the existing literature, while generating samples of comparable or superior quality. Finally, the reported results confirm that the samples generated by our algorithm yield accurate parameter estimates, which is of crucial importance in discrete choice modeling.

5. Conclusion

In this paper, we propose a simple and fast resampling technique designed to speed up the estimation of discrete choice models. The gain in computational time naturally comes at the cost of deteriorating the model estimation results; however, our method is specifically designed to mitigate this deterioration by preserving as much diversity as possible among the observations. As a result, the quality of the parameter estimates stays within reasonable ranges even for large reduction rates. The presented results additionally highlight the benefits of our method on the estimation of multinomial and nested logit models of small to medium sizes.

Intended future work includes the development and testing of more elaborate sampling strategies for sampling observations from buckets. For instance, those could be designed to increase the probability of choosing the most representative observations within each sub-bucket, or to rely on the content of the sub-buckets to generate synthetic prototypical observations. Additional investigation could also consist in developing LSH functions that can accommodate the analyst's knowledge of the dataset or the structure of the model of interest. For instance, one could consider a distinct projection for each alternative in the choice context and, in each projection, include only the variables that are relevant to the corresponding utility function; alternatively, one could associate probability distributions with greater means to specific variables so as to give them more importance in the hashing. Another natural progression of this work consists in extending our methodology to mixed logit models and maximum simulated likelihood estimation, which requires to carefully examine the interaction between locality-sensitive hashing and Monte Carlo integration: we suspect that blindly applying our method to halve the size of a dataset will not produce better estimation results than considering only half the number of draws in Monte Carlo integration. Rather, our intuition is that the two aspects necessarily need to be considered jointly. Finally, another promising direction of research consists in embedding the LSH-DR method within a stochastic optimization algorithm for model estimation, such as the one proposed in Lederrey et al. (2021). A recent study by the authors of this paper has shown that use of carefully selected and weighted batches of data — rather than random ones — could result in significantly better approximates of the gradient and Hessian matrix and, as a result, speed up convergence (Ortelli et al., 2023).

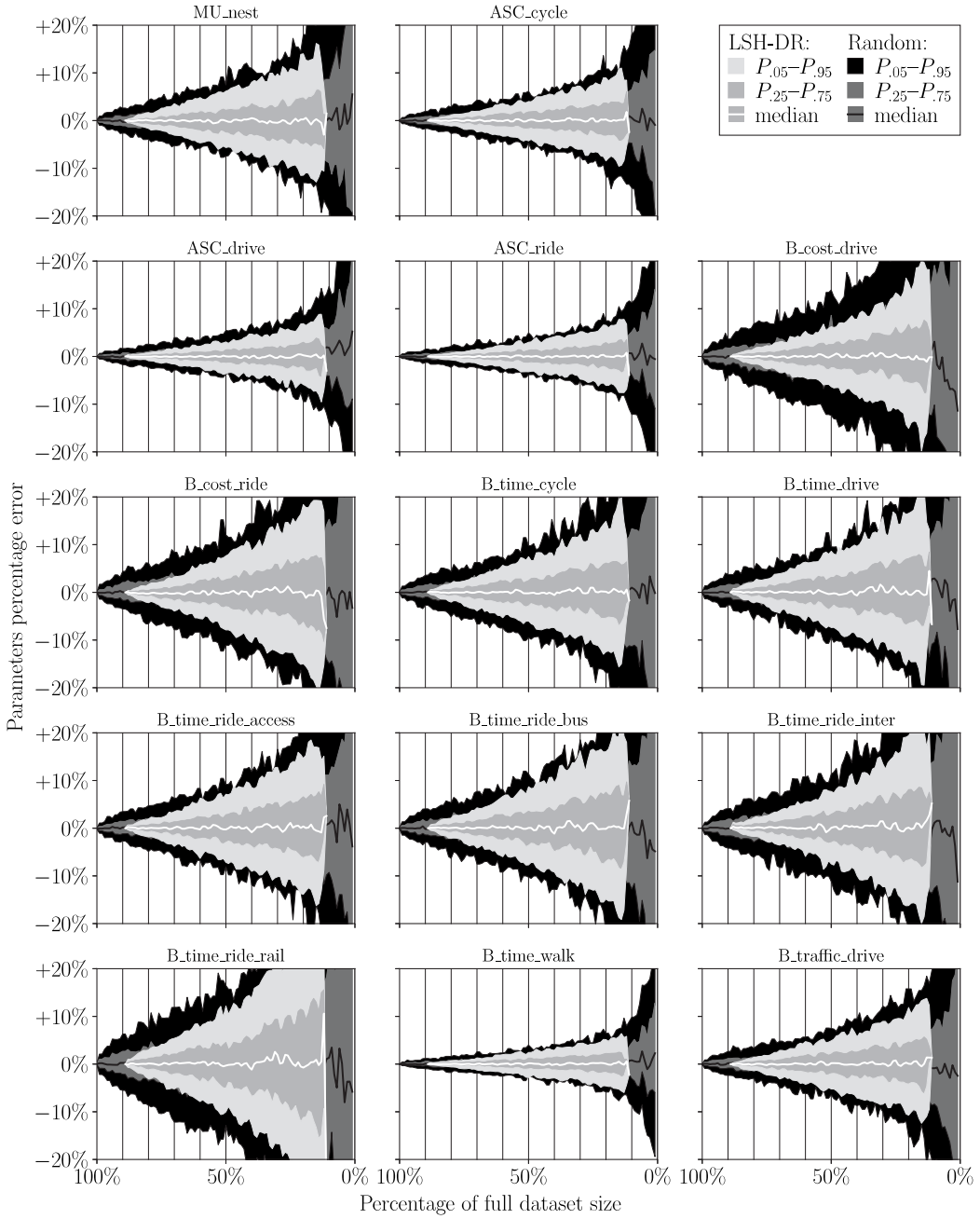


Fig. 7. Percentage error of the NL-S parameter estimates on samples generated by LSH-DR. The results obtained on random samples are also reported, for comparative purposes.

CRediT authorship contribution statement

Nicola Ortelli: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Matthieu de Lapparent:** Conceptualization, Supervision, Writing – review & editing. **Michel Bierlaire:** Conceptualization, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

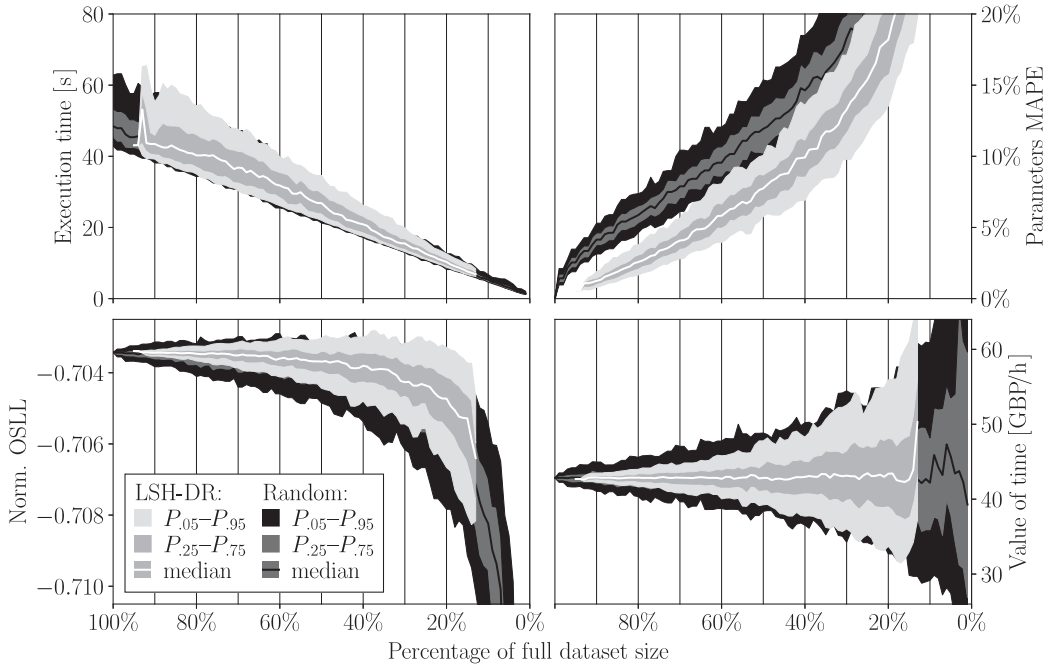


Fig. 8. Estimation of the MNL-L model on samples generated by LSH-DR. The results obtained on random samples are also reported, for comparative purposes.

Table 1
Specification of the MNL-S and MNL-L logit models (Hillel, 2019). All categorical variables are encoded using binary indicators. All explanatory variables are associated with alternative-specific parameters and all utility functions are linear in parameters. Constants included, the models consider 13 and 53 parameters, respectively.

Variable	Type	MNL-S				MNL-L			
		Walk	Cycle	Ride	Drive	Walk	Cycle	Ride	Drive
Travel time	cont.	×	×	×[a]	×	×	×	×[a]	×
Travel cost	cont.			×	×			×	×
Traffic level	cont.				×				×
Straight-line distance	cont.						×	×	×
Driver's license	bin.						×	×	×
Gender	bin.						×	×	×
Age:	cat.								
Child	*ind.*							×	×
Pensioner	*ind.*						×	×	×
Car ownership:	cat.								
One car in household	*ind.*						×[b]	×	×
More than one	*ind.*						×[b]	×	×
Trip purpose:	cat.								
Home-based work	*ind.*						×	×	×
Home-based education	*ind.*							×	×
Home-based other	*ind.*						×	×	
Employers' business	*ind.*						×	×	×
Time of departure:	cat.								
PM peak	*ind.*						×	×	×
Inter-peak	*ind.*						×		×
Day of week:	cat.								
Weekdays	*ind.*							×	×
Saturday	*ind.*						×	×	
Season:	cat.								
Winter	*ind.*						×		×

[a] The travel time of the "ride" alternative is split into four components, each associated with a distinct parameter: access and egress time, bus in-vehicle time, rail in-vehicle time and interchange time.

[b] The "cycle" alternative considers a single parameter associated to both car-ownership categories.

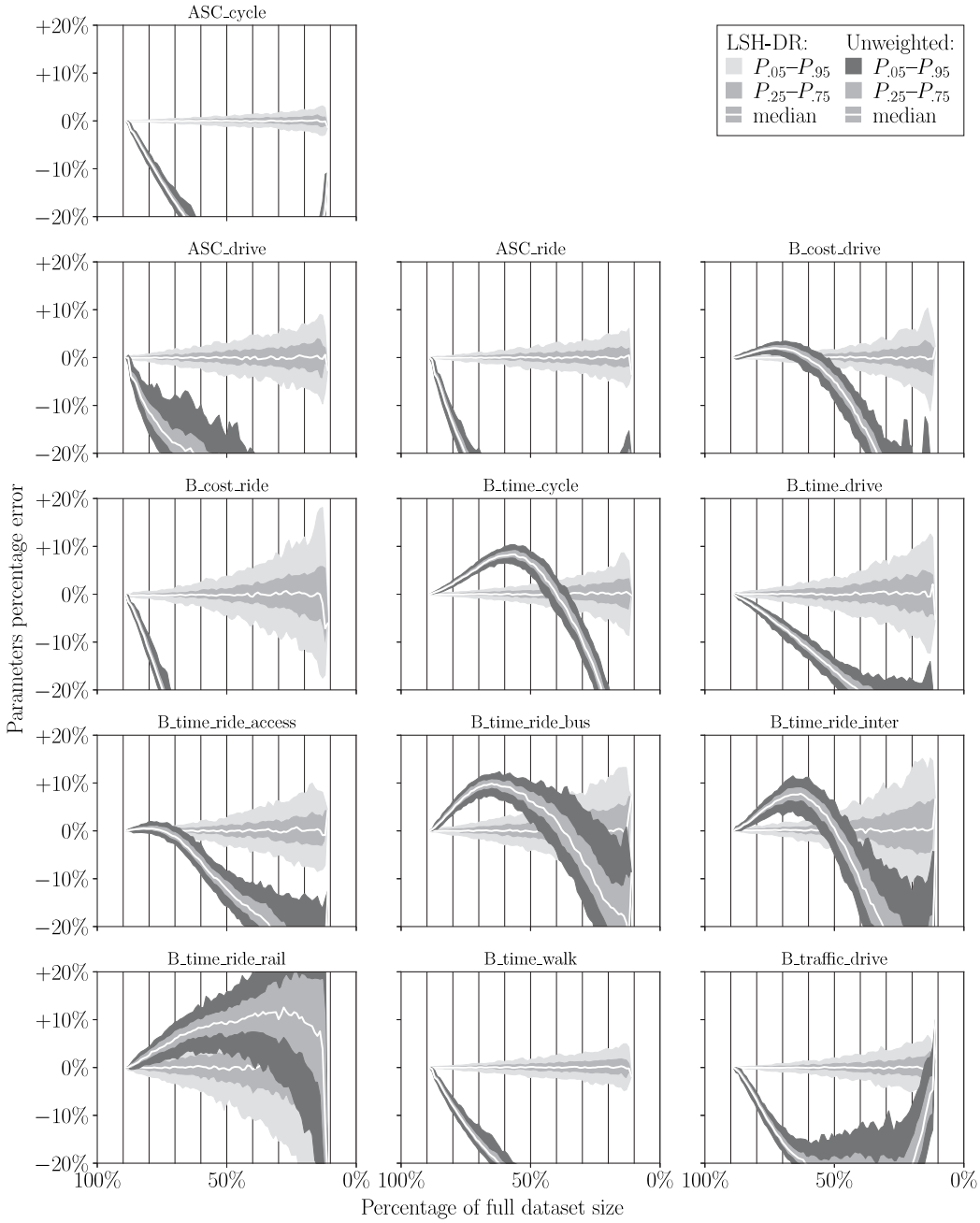


Fig. 9. Parameter estimates obtained in Experiment B, with and without weights.

Appendix

See Table 1 and Fig. 9.

References

Alexandropoulos, S.-A.N., Kotsiantis, S.B., Vrahatis, M.N., 2019. Data preprocessing in predictive data mining. Knowl. Eng. Rev. 34, e1.

Arnaiz-González, Á., Díez-Pastor, J.-F., Rodríguez, J.J., García-Osorio, C., 2016. Instance selection of linear complexity for big data. Knowl.-Based Syst. 107, 83–95.

Arteaga, C., Park, J., Beeramoole, P.B., Paz, A., 2022. xlogit: An open-source Python package for GPU-accelerated estimation of Mixed Logit models. J. Choice Model. 42, 100339.

Aslani, M., Seipel, S., 2020. A fast instance selection method for support vector machines in building extraction. Appl. Soft Comput. 97, 106716.
Bierlaire, M., 2023. A short introduction to Biogeme. Technical report, TRANSP-OR 230620, Transport and Mobility Laboratory, ENAC, EPFL.
Bierlaire, M., Krueger, R., 2020. Sampling and discrete choice. Technical report, TRANSP-OR 201109, Transport and Mobility Laboratory, ENAC, EPFL.
Castellanos, F.J., Valero-Mas, J.J., Calvo-Zaragoza, J., 2021. Prototype generation in the string space via approximate median for data reduction in nearest neighbor classification. Soft Comput. 25 (24), 15403–15415.
Chang, E., Shen, X., Yeh, H.-S., Demberg, V., 2021. On training instance selection for few-shot neural text generation. arXiv preprint arXiv:2107.03176.
Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S., 2004. Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth Annual Symposium on Computational Geometry. pp. 253–262.
Guevara, C.A., Ben-Akiva, M.E., 2013a. Sampling of alternatives in logit mixture models. Transp. Res. B 58, 185–198.
Guevara, C.A., Ben-Akiva, M.E., 2013b. Sampling of alternatives in multivariate extreme value (MEV) models. Transp. Res. B 48, 31–52.
Hillel, T., 2019. Understanding Travel Mode Choice: A New Approach for City Scale Simulation (Ph.D. thesis). University of Cambridge.
Hillel, T., Elshafie, M.Z., Jin, Y., 2018. Recreating passenger mode choice-sets for transport simulation: A case study of London, UK. Proc. Inst. Civ. Eng.-Smart Infrastruct. Constr. 171 (1), 29–42.
Lederrey, G., Lurkin, V., Hillel, T., Bierlaire, M., 2021. Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms. J. Choice Model. 38, 100226.
Leskovec, J., Rajaraman, A., Ullman, J.D., 2020. Mining of Massive Data Sets. Cambridge University Press.
Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 1977–1988.
Manski, C.F., McFadden, D., 1981. Structural Analysis of Discrete Data with Econometric Applications. MIT press, Cambridge, MA.
McFadden, D., 1978. Modeling the choice of residential location. In: Karlqvist, A., Lundqvist, L., Snickers, F., Weibull, J.W. (Eds.), Spatial Interaction Theory and Planning Models. North-Holland Publishing Company, pp. 75–96.
Molloy, J., Becker, F., Schmid, B., Axhausen, K.W., 2021. mixl: An open-source R package for estimating complex choice models on large datasets. J. Choice Model. 39, 100284.
Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Kittler, J., 2010. A review of instance selection methods. Artif. Intell. Rev. 34, 133–143.
Ortelli, N., de Lapparent, M., Bierlaire, M., 2023. Stochastic adaptive resampling for the estimation of discrete choice models. In: Proceedings of the 23rd Swiss Transportation Research Conference.
Ougiaroglou, S., Evangelidis, G., 2016. RHC: a non-parametric cluster-based data reduction for efficient k-NN classification. Pattern Anal. Appl. 19 (1), 93–109.
Ougiaroglou, S., Filippakis, P., Evangelidis, G., 2021. Prototype generation for multi-label nearest neighbours classification. In: Hybrid Artificial Intelligent Systems: 16th International Conference, HAIS 2021, Bilbao, Spain, September 22–24, 2021, Proceedings 16. Springer, pp. 172–183.
Park, Y., Qing, J., Shen, X., Mozafari, B., 2019. BlinkML: Efficient maximum likelihood estimation with probabilistic guarantees. In: Proceedings of the 2019 International Conference on Management of Data. pp. 1135–1152.
Paulevé, L., Jégou, H., Amsaleg, L., 2010. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. Pattern Recognit. Lett. 31 (11), 1348–1358.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Ren, H., Yang, B., 2019. Clustering-based prototype generation for imbalance classification. In: 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA). IEEE, pp. 422–426.
Rodrigues, F., 2022. Scaling Bayesian inference of mixed multinomial logit models to large datasets. Transp. Res. B 158, 1–17.
Saha, S., Sarker, P.S., Al Saud, A., Shatabda, S., Newton, M.H., 2022. Cluster-oriented instance selection for classification problems. Inform. Sci. 602, 143–158.
Schmid, B., Becker, F., Molloy, J., Axhausen, K.W., Lüdering, J., Hagen, J., Blome, A., 2022. Modeling train route decisions during track works. J. Rail Transp. Plan. Manag. 22, 100320.
Tsoleridis, P., Choudhury, C.F., Hess, S., 2022. Utilising activity space concepts to sampling of alternatives for mode and destination choice modelling of discretionary activities. J. Choice Model. 42, 100336.
van Cranenburgh, S., Bliemer, M.C., 2019. Information theoretic-based sampling of observations. J. Choice Model. 31, 181–197.
van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., Walker, J., 2021. Choice modelling in the age of machine learning-discussion paper. J. Choice Model. 100340.
Zhang, J., Liu, C., 2023. Fast instance selection method for SVM training based on fuzzy distance metric. Appl. Intell. 1–16.