A cross-validation approach for discrete choice models with sampled choice sets

Eric Larsen [*] Jean-Philippe Raymond [*] Emma Frejinger [*]
Angelo Guevara[†]

April 8, 2015

Extended abstract for hEART symposium 2015

Cross-validation can be used to assess and compare models' predictive performances. In brief, it consists in repeatedly partitioning a set of observations into two subsets: one used to estimate the models (training set) and one used to apply the models (validation set). A performance measure based on predicted choice probabilities is used to estimate the model's (out-of-sample) fit on the validation set. The issue using sampled choice sets is that the choice probabilities cannot be evaluated unless the observed alternative has been sampled. For estimation, chosen alternatives are added (with probability one) but that cannot be done in cross-validation since that would not correspond to a true prediction setting.

This paper makes two important contributions. First, we propose a loss-function as a performance measure for cross-validation that is well defined independently of the observed alternatives being sampled or not. The idea lies in using utilities as reference instead of probabilities since they can be evaluated independently of the choice set. Second, we show how the approach can be used to compare the predictive performance of path-based route choice models with sampled choice sets with a link-based recursive logit (RL) model (universal choice set).

There has been recent advances on choice set sampling for consistent estimation of multivariate extreme value, mixed logit and random regret minimization models applied in different contexts, e.g. location and route choice (Frejinger et al., 2009; Guevara and Ben-Akiva, 2013b; Guevara and Ben-Akiva, 2013a;

[*]Department of Computer Science and Operational Research, Université de Montréal and CIRRELT, Canada

[†]Facultad de Ingeniera y Ciencias Aplicadas, Universidad de los Andes, Chile

Guevara et al., 2014). These studies focus on deriving a sampling correction of utilities such that the pseudo maximum likelihood estimator is consistent. While numerous studies focus on the estimation problem, we are not aware of any literature focusing on validation using sampled choice sets. The validation step is important to check for over fitting and compare models' predictive performance. Our approach can be used for that purpose and it is relevant to any application that has universal choice sets that are too large to be practically feasible to deal with, or even too large to be enumerated (e.g. route choice application).

In the following we briefly describe the approach with a particular focus on the loss-function. Consider a given discrete choice model $P(i|\mathcal{U};\beta)$ where i is an alternative, $\mathcal{U}$ the universal choice set and β a vector of parameters. Moreover, consider one iteration of the cross-validation where a set of observations has been divided into a training $\mathcal{T}$ and validation $\mathcal{V}$ set. A choice set $D_n^{\mathcal{T}}$ is sampled for each observation $n \in \mathcal{T}$ and parameter estimates $\widehat{\beta}$ are obtained by maximizing the pseudo log-likelihood function over $P(i_n|D_n^{\mathcal{T}};\beta)\ \forall n \in \mathcal{T}$ where i_n is the observed alternative. This can be done using the state-of-the-art.

The next step consists in applying the estimated model to $n \in \mathcal{V}$. For this purpose we sample choice sets $D_n^{\mathcal{V}}\ \forall n \in \mathcal{V}$. Note that this set does not necessarily contain the observed alternative, and if it is not the case, we cannot compute $P(i_n|D_n^{\mathcal{V}};\beta)$. We define the average loss over this sample as the average probability the model assigns to alternatives $i \in D_n^{\mathcal{V}}$ having a higher deterministic utility (as defined by the model) than the observed alternative i_n. More precisely, the average loss is $\bar{\phi}^{\mathcal{V}}$

$$\bar{\phi}^{\mathcal{V}} = \frac{1}{|\mathcal{V}|}\sum_{n\in\mathcal{V}}\sum_{i\in D_n^{\mathcal{V}}}\delta_{in}P(i|D_n^{\mathcal{V}};\widehat{\beta})$$

where δ_{in} equals 1 if $V(i;\widehat{\beta}) > V(i_n;\widehat{\beta})$ and zero otherwise. This loss-function is well-defined independently of the observed alternative being sampled or not but its value depends on the sampled choice sets. As the number of iterations $k = 1,\ldots,K$ of the cross-validation increases (number of times the sample of observations is randomly divided into training and validation sets) and for choice sets $D_n^{\mathcal{V}}\ \forall n \in \mathcal{V}$ sufficiently large, the law of large numbers ensures that the average loss $\bar{\phi} = \frac{1}{K}\sum_{k=1}^{K}\bar{\phi}^{\mathcal{V}_k}$ converges.

We report numerical results using real data from a route choice application. There are two objectives with these results. First, to illustrate the approach and analyze the convergence of the average loss for different settings of the number of iterations, the sampling protocol and size of sampled choice sets. Second, to compare the predictive performance of path-based and link-based recursive logit (RL) route choice models (Fosgerau et al., 2013; Mai et al., 2015). The second objective implies an additional challenge since the RL model is based on the universal choice set whereas PSL is not. In order to have comparable loss functions

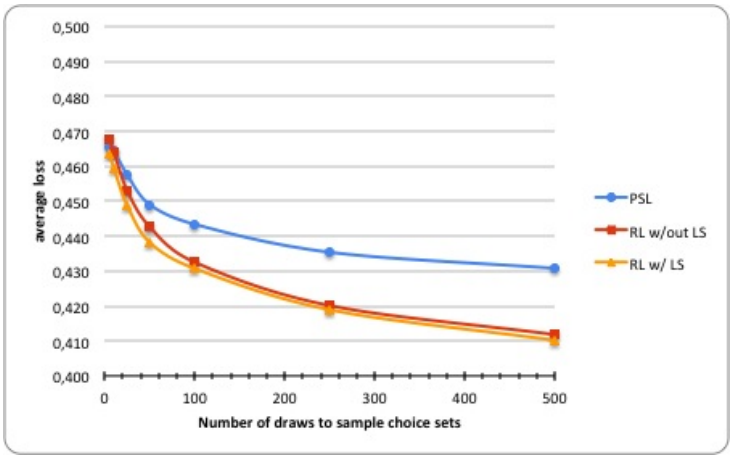

Figure 1: Average loss-function values as a function of number of draws used to sample choice sets

we normalize the RL choice probabilities such that they sum to one over the paths in the the sampled choice sets $D_n^{\mathcal{V}}$.

This is ongoing research and we report some preliminary results in this abstract for path size logit (PSL, Ben-Akiva and Bierlaire, 1999) and RL with and without link size attribute (Fosgerau et al., 2013). We use the Borlänge route choice data set that has been used in several studies, including Fosgerau et al. (2013). Figure 1 reports the average loss function values for one realization of $\mathcal{V}$ and different number of draws used to sample choice sets. These preliminary results suggest that the models have a similar prediction performance but the RL models are slightly better as the size of the choice sets increase. This is interesting since the RL models are based on the universal choice set. In the paper we will report a thorough numerical study comparing different route choice models and assessing the influence of (i) the number of iterations of the cross-validation (ii) the choice set sampling protocol, and (iii) the number of sampling draws on the average loss-function values.

References

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions, *in* R. Hall (ed.), *Handbook of Transportation Science*, Kluwer, pp. 5–34.

Fosgerau, M., Frejinger, E. and Karlström, A. (2013). A link based network route choice model with unrestricted choice set, *Transportation Research Part B*

56: 70–80.

Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling, *Transportation Research Part B* **43**(10): 984–994.

Guevara, C. A. and Ben-Akiva, M. E. (2013a). Sampling of alternatives in logit mixture models, *Transportation Research Part B* **58**(1): 185–198.

Guevara, C. A. and Ben-Akiva, M. E. (2013b). Sampling of alternatives in multivariate extreme value (MEV) models, *Transportation Research Part B* **48**(1): 31 – 52.

Guevara, C. A., Chorus, C. G. and Ben-Akiva, M. E. (2014). Sampling of alternatives in random regret minimization models, *Transportation Science* . forthcoming.
URL: *http://dx.doi.org/10.1287/trsc.2014.0573*

Mai, T., Fosgerau, M. and Frejinger, E. (2015). A nested recursive logit model for route choice analysis, *Transportation Research Part B* . Accepted for publication.