Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 47 (2020) 457–464

Transportation Research Procedia

www.elsevier.com/locate/procedia

22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18-20 September 2019, Barcelona, Spain

Discrete choice modeling using Kernel Logistic Regression

José Ángel Martín-Baos[a,*], Ricardo García-Ródenas[a], María Luz López-García[a], Luis Rodriguez-Benitez[b]

[a]*Department of Mathematics, Faculty of Computer Science. University of Castilla-La Mancha. 13071, Ciudad Real, Spain.*
[b]*Department of Information Systems and Technologies, Faculty of Computer Science. University of Castilla-La Mancha. 13071, Ciudad Real, Spain.*

Abstract

The Kernel Logistic Regression is a popular technique in machine learning. In this work this technique is applied to the field of discrete choice modeling. This approach is equivalent to specifying non-parametric utilities in random utility models. A Monte Carlo simulation experiment has been carried out to compare this approach with Multinomial Logit models, comparing the goodness of fit and the capability of obtaining the specified utilities.

Keywords: Kernel Logistic Regression; Random Utility Models; Non-parametric Utilities.

1. Introduction

Nowadays, artificial intelligence (AI) has achieved great popularity due to its success in applications such as autonomous vehicles, intelligent robots, image and voice recognition, automatic translation, etc. The construction of most of these intelligent machines is based on machine learning (ML) methods. These achievements have led to an increment in the usage of ML methods and there is an increasing interest in expanding the domain of applications, for instance in the field of transport planning.

Discrete choice methods based on Random Utility Theory (RUM) have been developed over the last four decades, and currently they have acquired a high degree of sophistication which allows them to obtain a set of measures such as willlingness to pay (WTP), value of the time (VOT), elasticities, market shares, etc. These measures quantify essential magnitudes of the studied problem and allow to evaluate the result of any intervention in the transport system. Discrete choice models describe how a rational decision-maker chooses an alternative between a set of choices depending on the characteristics of each one of them and the peculiarities of the decision-maker (Ben-Akiva and Bierlaire (1999a); McFadden (1978); Train (2009)). In the decision process there are some latent (unobservable)

* Corresponding author. Tel.: +34-926-29-53-00 Ext: 96683
E-mail address: JoseAngel.Martin@uclm.es

2352-1465
10.1016/j.trpro.2020.03.121

functions called utility functions which measure the interest of each alternative for a user. The decisor-maker is supposed to choose the alternative that maximizes his utility. The utility of each alternative consists of a deterministic part and a stochastic one. The probability distribution of this stochastic part determines the resultant model, being the multinomial logistic (MNL) model the most widespread example. These models are estimated using maximum likelihood estimation methodology which allows to determine an asymptotic distribution of the estimators and makes possible to test hypotheses in the parameters.

The combination of MNL with radial basis functions is known in the machine learning community as Kernel Logistic Regression (KLR) (Zhu and Hastie (2005)), which is derived without any probabilistic assumptions. Moreover, in KLR the parameter estimation problem is based on a penalized maximum likelihood estimation in which the goodness of fit criterion weighs the empirical risk and its complexity.

Espinosa-Aranda et al. (2018) propose a Nested Logit (NL) model with restrictions in which utilities are specified by radial basis functions. This paper extends the KLR to the situation of NL models with (exogenous or endogenous) constraints. The NL model with restrictions is defined implicitly through the resolution of an optimization problem. This characteristic leads to a bi-level structure of the estimation problem and the need to employ meta-heuristic methods instead of using a canonical method as Newton's method. Additionally, Espinosa-Aranda et al. (2015) uses this model in a passenger-centered train timetabling problem.

Discrete choice models require the modeler to specify a functional expression using the attribute set and a vector of parameters (parametric utilities), while in above ML models, such as KLR, this analytical expression is not necessary, being enough to choose a type of the so-called kernel function. We refer to this second approach as non-parametric utilities.

ML methods have been compared with the RUM models in the literature, being limited to analyzing which has the highest predictive capacity in the discrete choice problems. One approach that should be researched is to generalize the RUM theory in order to incorporate several achievements of the ML methods, thus allowing to take advantage of both methodologies. In this work the KLR methods are reviewed under a RUM perspective, showing that these approaches provide a way to specify utilities (non-parametric utilities) and associated estimation techniques. A controlled computational experiment have been conducted, using Monte Carlo simulation methods, to motivate the use of non-parametric utilities in non-linear phenomena.

2. Methodology

In this section, RUM and KLR methods are reviewed in order to introduce a common notation in a way they can be compared in the next section.

2.1. Random Utility Model

As described in Ben-Akiva and Bierlaire (1999b), utility theory assumes that the decision-maker's preference for an alternative can be captured by a value, which is called *utility*, and the decision-maker selects the alternative with the highest associated utility from his/her choice set. This approach presents some strong limitations in practical applications because the underlying assumptions of this concept are often violated. For this reason, some models such as Random Utility Models (RUMs) assume that the decision-maker has perfect discrimination capability. However, the analyst is considered to have incomplete information and, therefore, uncertainty must be taken into account.

In RUM the utility defined for a decision-maker n to select an alternative i from the choice set C_n is given by

$$U_{in} = V_{in} + \varepsilon_{in}, \tag{1}$$

where V_{in} is the *deterministic* (also called systematic) component of the utility, and ε_{in} is the unobserved component, which is a random term used to include the impact of all the unobserved variables that have an impact on the utility function. Hence, the probability that a decision-maker n chooses an alternative i from the choice set C_n is

$$P(i|C_n) = P\left(U_{in} \geq U_{jn} \quad \forall j \in C_n\right) = P\left(U_{in} = \max_{j \in C_n} U_{jn}\right). \tag{2}$$

Some assumptions are necessary to make the random utility model operational. The concept of utility is relative and not absolute, for this reason only the signs of the differences between utilities are relevant. This is shown in eq. (3) where location and scale parameters are introduced, represented as α and μ, respectively, and where $\mu > 0$.

$$\begin{aligned} P\left[U_{in} \geq U_{jn} \quad \forall j \in C_n\right] = \\ P\left[\mu U_{in} + \alpha \geq \mu U_{jn} + \alpha \quad \forall j \in C_n\right] = \\ P\left[U_{in} - U_{jn} \geq 0 \quad \forall j \in C_n\right]. \end{aligned} \tag{3}$$

The scale parameter is usually selected in order to obtain a convenient normalization of one of the variances of the unobserved component. Generally, the location parameter α is set to zero.

The hypotheses about the errors distribution ε_{in} determines the probability of choosing each alternative by the expression (3). The MNL models assume a Gumbel distribution and, therefore, in this case the probability of each alternative is given by the expression

$$P(i/V_n, C_n) = \frac{\exp(V_{in})}{\sum_{j \in C_n} \exp(V_{jn})}, \tag{4}$$

where V_n is the utility vector of the alternatives for the decision-maker n, i.e. $V_n = (V_{1n}, \cdots, V_{In})^T$, being I the total number of alternatives.

The deterministic term of the utility V_{in} of each alternative i, which is defined in Eq. (5), is a function that depends on the vector z_{in} of attributes of the alternative itself as perceived by the individual n and a vector S_n of characteristics of the decision-maker. It's written mathematically

$$V_{in} = V(z_{in}, S_n). \tag{5}$$

The previous equation can be simplified if an appropriate vector valued function h is used, which defines a vector of attributes x_{in} from z_{in} and S_n, that is

$$x_{in} = h(z_{in}, S_n). \tag{6}$$

Therefore, the deterministic utility V_{in} can be defined, for example, using a linear function

$$V_{in} = V_i(x_{in}, \boldsymbol{\beta}_i) = \boldsymbol{\beta}_i^\top x_{in} = \sum_{k=1}^{K} \beta_{ik} x_{ink}, \tag{7}$$

where $\boldsymbol{\beta}_i$ is the vector of parameters.

The utility functions depend on a vector of parameters $\boldsymbol{\beta}_i$ which needs to be estimated for each i. Let $\boldsymbol{\Theta}_\beta = (\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_I^\top)^\top$. The canonical method for estimating $\boldsymbol{\Theta}_\beta$ is the maximum likelihood estimation. It is assumed that the sample $\mathbf{X} = \{x_{in}\}$ of attributes for each decision-maker $n = 1, \cdots, N$ and for each alternative $i = 1, \cdots, I$ has been observed. The choice set for the decision-maker n is denoted by C_n and their decisions by $\mathbf{y}$ where $y_{in} = 1$ if the n decision-maker chooses the i alternative or $y_{in} = 0$, otherwise.

The likelihood of the sample is

$$p(\mathbf{y}|\mathbf{V}) = \prod_{n=1}^{N} \prod_{i \in C_n} P(i|V_n, C_n)^{y_{in}}. \tag{8}$$

The estimate of the $\boldsymbol{\beta}_i$ parameters of the utilities $V_i(x_{in}, \boldsymbol{\beta}_i)$ is obtained using Maximum Likelihood Estimation (MLE), by solving

$$\underset{\boldsymbol{\Theta}_\beta}{\text{Maximize}}\ LL(\boldsymbol{\Theta}_\beta) = \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}_\beta), \tag{9}$$

where

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}_\beta) = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \log P(i|\mathbf{x}_n, \boldsymbol{\Theta}_\beta, C_n). \tag{10}$$

and $\mathbf{x}_n = (x_{1n}^\top, x_{2n}^\top, \cdots, x_{in}^\top, \cdots, x_{In}^\top)^\top$.

Once the main concepts related to the RUM have been established, the following point introduces KLR.

2.2. Kernel Logistic Regression

Many machine learning methods approach the problem of classification from a non-statistical point of view. These procedures do not intend to explain the process of choosing for a specific user, but to develop procedures with the least classification error. For this reason they consider all available information about a user to explain the choice of the alternative i, i.e. $\mathbf{x}_n = (x_{1n}^\top, x_{2n}^\top, \cdots, x_{in}^\top, \cdots, x_{In}^\top)^\top$, and consider that all the users have the same choice set, i.e. $C_n = C$ for all decision-makers n. The goal is to predict the alternative chosen by the decision-maker n given the characteristic vector $\mathbf{x}_n$. Support Vector Machine (SVM) (Cortes and Vapnik (1995)) has been one of the most promising methods over the last few decades but with the emergence of deep networks this dominance has been weakened, Theodoridis (2015). KLR is considered a variant of SVM, which not only predicts the classification of an object (an individual's choice), but also estimates the probability of belonging to each category.

KLR builds some *latent functions*, $V_i(\mathbf{x})$ for all $i \in C$, which is equivalent to the systematic utility functions of RUM, but considering them as black box functions. Due to their equivalence they are denoted in the same way. The classification criteria is based on maximizing the expected utility

$$i^* = \arg \underset{i \in C}{\text{maximize}} \{V_i(\mathbf{x})\}. \tag{11}$$

If the function $\Psi(\mathbf{x})$ is added to all the functions $V_i(\mathbf{x})$ in KLR, the decision rule (11) does not change:

$$i^* = \arg \underset{i \in C}{\text{maximize}} \{V_i(\mathbf{x})\} = \arg \underset{i \in C}{\text{maximize}} \{V_i(\mathbf{x}) + \Psi(\mathbf{x})\}. \tag{12}$$

This illustrates that, just like the utility functions in RUM, the latent functions $V_i(\mathbf{x})$ are overspecified. Therefore, without prejudicing the explanatory capacity of the model, it can be assumed that $V_I(\mathbf{x}) = 0$.

KLR provides estimates of the class probabilities based on the functions $V_i(\mathbf{x})$ equivalent to the Eq. (4),

$$P(i|\mathbf{V}, C) = \frac{\exp(V_i(\mathbf{x}))}{1 + \sum_{i=1}^{I-1} \exp(V_i(\mathbf{x}))}. \tag{13}$$

The main problem is finding the latent functions $V_i : \mathcal{X} \mapsto \mathbf{R}$ for $i = 1, \cdots, I-1$. KLR search for functions $V_i(\mathbf{x})$ within function spaces named *Reproducing Kernel Hilbert Spaces* (RKHS). The RKHS space is a vector space which is univocally generated by the so-called *kernel function* $k(\mathbf{x}, \mathbf{x}')$, and its associated RKHS space is denoted by $\mathcal{H}_k$. The family of functions $\{k(\mathbf{x}, \mathbf{x}')\}_{x' \in \mathcal{X}}$ constitutes a basis of the vector space. Any element from $\mathcal{H}_k$ can be represented as a linear combination of basis elements, in particular from $V_i(\mathbf{x}) \in \mathcal{H}_k$, the expression of the non-parametric utilities is given by:

$$V_i(\mathbf{x}, \boldsymbol{\alpha}_i) = \sum_{n=1}^{N} \alpha_{in} k(\mathbf{x}_n, \mathbf{x}). \tag{14}$$

Two popular kernel functions are the dth polynomial kernel, $k(\mathbf{x}_n, \mathbf{x}) = (\mathbf{x}_n^\top \mathbf{x} + \rho)^d$, and the Gaussian kernel, $k(\mathbf{x}_n, \mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2\sigma^2}\right)$, where $\|\cdot\|$ is the Euclidean norm. Another is the squared exponential kernel with Automatic

Relevance Determination,

$$k(\mathbf{x}_n, \mathbf{x}) = \sigma^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_n)^\top \Lambda^{-1}(\mathbf{x} - \mathbf{x}_n)\right), \tag{15}$$

where Λ is a diagonal matrix of the squared lengthscale hyperparameters and σ^2 is a variance hyperparameter. This kernel function is particularly relevant when the attributes are measured on different scales.

Once a kernel function $k(\mathbf{x}, \mathbf{x}')$, has been chosen, the expression (14) shows that to obtain the functions V_i requires estimating the parameter vector, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \cdots, \alpha_{in}, \cdots, \alpha_{iN})^\top$. The parameter vector of the KLR model is denoted by $\boldsymbol{\Theta}_\alpha = (\boldsymbol{\alpha}_1^\top, \cdots, \boldsymbol{\alpha}_{I-1}^\top)^\top$.

The literature proposes a regularized function estimation in RKHS for the estimation of $\boldsymbol{\Theta}_\alpha$. This method suggests estimating the parameters by solving the following optimization problem

$$\underset{\boldsymbol{\Theta}_\alpha}{\text{Minimize}} \sum_{n=1}^{N}\sum_{i=1}^{I} L(y_{in}, V_i(\mathbf{x}_n, \boldsymbol{\alpha}_i)) + \frac{\lambda}{2}\sum_{i=1}^{I} \|V_i(\mathbf{x}, \boldsymbol{\alpha}_i)\|^2_{\mathcal{H}_k}, \tag{16}$$

where $L(\cdot)$ is a loss function that measures discrepancies between predicted and observed classifications, λ is a regularization parameter that controls the trade-off between goodness of fit and complexity of the model and the norm of the utility functions is computed in the space RKHS by $\|V_i(\mathbf{x}, \boldsymbol{\alpha}_i)\|^2_{\mathcal{H}_k} = \boldsymbol{\alpha}_i^T \mathbf{K} \boldsymbol{\alpha}_i$, being $\mathbf{K}$ the Gram matrix, defined by $\mathbf{K}_{n,n'} = k(\mathbf{x}_n, \mathbf{x}_{n'})$.

Note that Eq. (16) has the form of *loss + penalty*. The loss function $L(\cdot)$ allows many different ways of measuring the model adjustment, the KLR uses the negative value of the log-likelihood function as a loss function, i.e. $-\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}_\alpha)$. Therefore,

$$\sum_{n=1}^{N}\sum_{i=1}^{I} L(y_{in}, V_i(\mathbf{x}_n, \boldsymbol{\alpha}_i)) = -\sum_{n=1}^{N}\sum_{i=1}^{I} y_{in} \log P(i|\mathbf{x}_n, \boldsymbol{\Theta}_\alpha). \tag{17}$$

This procedure is called minimizing Regularised Negative Log-Likelihood (RNLL) or equivalently maximizing Penalised Maximum Likelihood Estimation (PMLE).

KLR operates with the utility functions as black boxes, without trying to explain the choice process. This procedure starts from a feature vector $\mathbf{x}_n$ for each decision-maker n and aims to predict the selected choice, represented by the dummy variable y_{in}, and to that end uses all the information in the classification process, without distinguishing between features of one alternative against the other. In other words, to estimate the utility V_i of the decision-maker n, it uses both the characteristics of the alternative i itself and those of the rest. Mathematically, RUM takes into account functions $V_i(x_{in})$, while KLR proposes the use of functions $V_i(\mathbf{x}_n)$ which depend on the whole vector $\mathbf{x}_n$.

The main difference between the approach followed by the KLR models, which is referred as *non-parametric*, and that followed by RUM, which is denominated as *parametric*, is how they specify the utility function. In the parametric approach it is necessary to define a functional expression in advance, establishing from the beginning the effect of each attribute against the others, while in the non-parametric approach the kernel functions $k(\cdot, \cdot)$ are chosen and this choice determines the RKHS $\mathcal{H}_k$ where the utilities $V_i(\mathbf{x}) \in \mathcal{H}_k$ are searched. The non-parametric approach presents the fundamental advantage that these utility specifications allow to approximate very diverse linear and non-linear phenomena, without the need to have prior knowledge of the phenomenon. Once RUM and KLR methods have been reviewed, in next section presents some numerical results obtained using both methodologies.

3. Numerical results

To compare KLR in relation with RUM, an experimental study has been designed. A Monte Carlo simulation has been designed in order to control the error term and the utility specification. The utility of the alternative i for the individual n is given by the expression:

$$U_{in} = V(x_{in1}, x_{in2}) + \varepsilon_{in}; \text{ with } i \in \{1, 2, 3\}, \tag{18}$$

where the error terms ε_{in} are independent and identically distributed (IID) random variables draw from a Gumbel distribution with scale parameter λ and location parameter 0.

For the simulation experiment three systematic utilities have been supposed:

$$V(x_{in1}, x_{in2}) = \beta_1 x_{in1} + \beta_2 x_{in2} \quad \text{Linear} \tag{19}$$

$$V(x_{in1}, x_{in2}) = {x_{in1}}^{\beta_1} {x_{in2}}^{\beta_2} \quad \text{Cobb-Douglas (CD)} \tag{20}$$

$$V(x_{in1}, x_{in2}) = \min\{\beta_1 x_{in1}, \beta_2 x_{in2}\} \quad \text{Minimum} \tag{21}$$

For each of the previous models three pairs of parameters have been considered, being the parameter $\beta_1 = 1$ and $\beta_2 \in \{0.5, 1, 2\}$. Moreover, each of these nine models has been defined using two levels of uncertainty, by means of the scale parameter $\lambda \in \{0.1, 0.01\}$. Therefore, 18 different models have been considered. For each model, a total of $N = 1000$ individuals was generated using a uniform distribution of x_{in} on the square $[0, 1] \times [0, 1]$. Finally, to obtain more accurate results, we have generated 100 samples of each model.

In this numerical experiment two multinomial logit models have been estimated using two different utility specifications

$$V_i(x_{in}, \boldsymbol{\beta}_i) = \beta_0^i + \beta_1 x_{in1} + \beta_2 x_{in2} \quad \text{(Linear utilities)} \tag{22}$$

$$V_i(x_{in}, \boldsymbol{\alpha}) = \sum_m \alpha_m \exp\left(\rho \|x_{in} - x_m\|^2\right) \quad \text{(Non-parametric utilities)} \tag{23}$$

Both kind of parameters have been estimated using MLE and all the analyses have been coded in the R programming environment (see The R Development Core Team (2008)). In order to estimate the linear multinomial logit model in R, the 'mlogit' package was used (see Croissant). The intercept of the first alternative is always fixed to 0, i.e $\beta_0^1 = 0$. Concerning KLR, the 'kernlab' package is used which provides kernel-based machine learning methods in R (see Karatzoglou et al. (2004)). Using these R packages, several functions have been constructed to estimate the non-parametric utilities using the MLE.

For each of the 18 different models, for each of the 100 generated samples and for each level of error λ, the Log-likelihood (LL) and the McFadden value $\rho^2 = 1 - \frac{LL(\widehat{\boldsymbol{\Theta}})}{LL(0)}$ are calculated, where $\widehat{\boldsymbol{\Theta}}$ is the estimate of the vector of parameters. Tables 1 to 3 show the mean and standard deviation value of the log-likelihood and the ρ^2 indices. These tables compare the goodness of fit between linear utilities given in Eq. (22) and the non-parametric utilities given in Eq. (23).

Table 1. Cobb-Douglas

Lambda	Parameters	Linear utilities		Non-parametric utilities	
		Mean value	Standard deviation	Mean value	Standard deviation
$\lambda = 0.01$	$\beta_1 = 1$	$\overline{LL} = -261.79$	$\sigma_{LL} = 18.83$	$\overline{LL} = -118.21$	$\sigma_{LL} = 14.81$
	$\beta_2 = 1$	$\overline{\rho^2} = 0.76$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.89$	$\sigma_{\rho^2} = 0.01$
	$\beta_1 = 1$	$\overline{LL} = -273.27$	$\sigma_{LL} = 18.05$	$\overline{LL} = -155.20$	$\sigma_{LL} = 19.47$
	$\beta_2 = 2$	$\overline{\rho^2} = 0.75$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.86$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -244.95$	$\sigma_{LL} = 19.77$	$\overline{LL} = -117.17$	$\sigma_{LL} = 13.80$
	$\beta_2 = 0.5$	$\overline{\rho^2} = 0.78$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.89$	$\sigma_{\rho^2} = 0.01$
$\lambda = 0.1$	$\beta_1 = 1$	$\overline{LL} = -601.99$	$\sigma_{LL} = 22.56$	$\overline{LL} = -537.13$	$\sigma_{LL} = 20.42$
	$\beta_2 = 1$	$\overline{\rho^2} = 0.45$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.51$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -701.07$	$\sigma_{LL} = 25.34$	$\overline{LL} = -616.16$	$\sigma_{LL} = 22.35$
	$\beta_2 = 2$	$\overline{\rho^2} = 0.36$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.44$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -549.99$	$\sigma_{LL} = 23.56$	$\overline{LL} = -503.17$	$\sigma_{LL} = 23.72$
	$\beta_2 = 0.5$	$\overline{\rho^2} = 0.50$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.54$	$\sigma_{\rho^2} = 0.02$

Table 2. Linear

Lambda	Parameters	Linear utilities		Non-parametric utilities	
		Mean value	Standard deviation	Mean value	Standard deviation
$\lambda = 0.01$	$\beta_1 = 1$	$\overline{LL} = -30.81$	$\sigma_{LL} = 7.21$	$\overline{LL} = -59.90$	$\sigma_{LL} = 6.73$
	$\beta_2 = 1$	$\overline{\rho^2} = 0.97$	$\sigma_{\rho^2} = 0.01$	$\overline{\rho^2} = 0.95$	$\sigma_{\rho^2} = 0.01$
	$\beta_1 = 1$	$\overline{LL} = -19.02$	$\sigma_{LL} = 5.40$	$\overline{LL} = -62.20$	$\sigma_{LL} = 6.24$
	$\beta_2 = 2$	$\overline{\rho^2} = 0.98$	$\sigma_{\rho^2} = 0.00$	$\overline{\rho^2} = 0.94$	$\sigma_{\rho^2} = 0.01$
	$\beta_1 = 1$	$\overline{LL} = -39.23$	$\sigma_{LL} = 7.84$	$\overline{LL} = -66.66$	$\sigma_{LL} = 7.04$
	$\beta_2 = 0.5$	$\overline{\rho^2} = 0.96$	$\sigma_{\rho^2} = 0.01$	$\overline{\rho^2} = 0.94$	$\sigma_{\rho^2} = 0.01$
$\lambda = 0.1$	$\beta_1 = 1$	$\overline{LL} = -322.90$	$\sigma_{LL} = 19.12$	$\overline{LL} = -324.48$	$\sigma_{LL} = 19.16$
	$\beta_2 = 1$	$\overline{\rho^2} = 0.71$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.70$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -207.11$	$\sigma_{LL} = 15.41$	$\overline{LL} = -210.75$	$\sigma_{LL} = 15.47$
	$\beta_2 = 2$	$\overline{\rho^2} = 0.81$	$\sigma_{\rho^2} = 0.01$	$\overline{\rho^2} = 0.81$	$\sigma_{\rho^2} = 0.01$
	$\beta_1 = 1$	$\overline{LL} = -395.32$	$\sigma_{LL} = 21.73$	$\overline{LL} = -396.72$	$\sigma_{LL} = 21.95$
	$\beta_2 = 0.5$	$\overline{\rho^2} = 0.64$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.64$	$\sigma_{\rho^2} = 0.02$

Table 3. Minimum

Lambda	Parameters	Linear utilities		Non-parametric utilities	
		Mean value	Standard deviation	Mean value	Standard deviation
$\lambda = 0.01$	$\beta_1 = 1$	$\overline{LL} = -486.01$	$\sigma_{LL} = 20.38$	$\overline{LL} = -186.77$	$\sigma_{LL} = 18.53$
	$\beta_2 = 1$	$\overline{\rho^2} = 0.56$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.83$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -467.20$	$\sigma_{LL} = 19.86$	$\overline{LL} = -215.99$	$\sigma_{LL} = 16.50$
	$\beta_2 = 2$	$\overline{\rho^2} = 0.57$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.80$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -471.46$	$\sigma_{LL} = 23.71$	$\overline{LL} = -226.80$	$\sigma_{LL} = 17.76$
	$\beta_2 = 0.5$	$\overline{\rho^2} = 0.57$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.79$	$\sigma_{\rho^2} = 0.02$
$\lambda = 0.1$	$\beta_1 = 1$	$\overline{LL} = -661.10$	$\sigma_{LL} = 25.78$	$\overline{LL} = -507.04$	$\sigma_{LL} = 24.27$
	$\beta_2 = 1$	$\overline{\rho^2} = 0.40$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.54$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -617.18$	$\sigma_{LL} = 20.42$	$\overline{LL} = -475.74$	$\sigma_{LL} = 20.10$
	$\beta_2 = 2$	$\overline{\rho^2} = 0.44$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.57$	$\sigma_{\rho^2} = 0.02$
	$\beta_1 = 1$	$\overline{LL} = -809.68$	$\sigma_{LL} = 19.97$	$\overline{LL} = -736.89$	$\sigma_{LL} = 22.70$
	$\beta_2 = 0.5$	$\overline{\rho^2} = 0.26$	$\sigma_{\rho^2} = 0.02$	$\overline{\rho^2} = 0.33$	$\sigma_{\rho^2} = 0.02$

Table 1 shows the result of the experiments using the models with $\lambda = 0.01$ and $\lambda = 0.1$ for the Cobb-Douglas systematic utility. The effect of the error term ε_{in} is lower in the model with $\lambda = 0.01$, therefore, the generated data contain more information about the phenomenon and the results of both linear and non-parametric utilities are better. Tables 2 and 3 shows the results for Linear and Minimum systematic utilities, respectively.

As it can be noticed, the non-parametric utilities overwhelm the linear utilities as they are able to generalize better and, consequently, it is possible to determine whether the model is non-linear and to adapt better to it. This can be checked with Cobb-Douglas and Minimum models, where non-parametric utilities adapt better to the model and give better results. That is not the case in linear models, where the non-parametric utilities slightly underperform linear utilities. Nonetheless, the Log-likelihood and ρ^2 values reported are very similar, so the differences do not seem significant.

4. Conclusions

This work shows, using a Monte Carlo study, that non-parametric utilities derived from KLR allow to capture non-linear effects of the decision process, which cannot be represented with linear RUMs. Moreover, KLR is also able

to approximate linear effects. Therefore, KLR frees the modeler to choose the functional expression of the utilities, because the non-parametric utilities can adapt to the phenomenon. Our future work will be focus on analyzing the theoretical aspects of these methods and the possibility of using KLR to retrieve post-analysis estimators such as willingness to pay, value of time, elasticities, market shares, etc.

Acknowledgements

The research of Martín-Baos, García-Ródenas and López-García was supported by Project TRA2016-76914-C3-2-P of the Spanish Ministry of Science and Innovation, co-funded by the European Regional Development Fund.

The research of Rodriguez-Benitez was supported by Project TIN2015-64776-C3-3-R of the Spanish Ministry of Science and Innovation, co-funded by the European Regional Development Fund.

References

Ben-Akiva, M., Bierlaire, M., 1999a. Discrete choice methods and their applications to short term travel decisions, in: Hall, R.W. (Ed.), Handbook of transportation science. Kluwer Academic Publishers, Massachusetts, pp. 5–33.

Ben-Akiva, M., Bierlaire, M., 1999b. Discrete Choice Methods and their Applications to Short Term Travel Decisions, in: Hall, R.W. (Ed.), Handbook of Transportation Science. Springer US, Boston, MA, pp. 5–33. URL: https://doi.org/10.1007/978-1-4615-5203-1_2, doi:10.1007/978-1-4615-5203-1{_}2.

Cortes, C., Vapnik, V., 1995. Supprot-Vector Networks. Machine Learning 20, 273–297. URL: http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/19549084, doi:10.1111/j.1747-0285.2009.00840.x.

Croissant, Y., . Estimation of multinomial logit models in R : The mlogit Packages. URL: https://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf.

Espinosa-Aranda, J., García-Ródenas, R., López-García, M., Angulo, E., 2018. Constrained nested logit model: formulation and estimation. Transportation 45, 1523–1557. doi:10.1007/s11116-017-9774-2.

Espinosa-Aranda, J., García-Ródenas, R., Ramírez-Flores, M., López-García, M., Angulo, E., 2015. High-speed railway scheduling based on user preferences. European Journal of Operational Research 246. doi:10.1016/j.ejor.2015.05.052.

Karatzoglou, A., Hornik, K., Smola, A., Zeileis, A., 2004. kernlab - An S4 package for kernel methods in R. Journal of Statistical Software 11, 1–20.

McFadden, D.L., 1978. Modelling the Choice of Residential Location, in: et al., A.K. (Ed.), Spatial Interaction Theory and Residential Location. North Holland, Amsterdam, The Netherlands, pp. 75–96.

The R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. URL: http://www.gnu.org/copyleft/gpl.html.

Theodoridis, S., 2015. Machine Learning. A Bayesian and Optimization Perspective.

Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press. URL: https://econpapers.repec.org/RePEc:cup:cbooks:9780521766555.

Zhu, J., Hastie, T., 2005. Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics 14, 185–205. URL: https://doi.org/10.1198/106186005X25619, doi:10.1198/106186005X25619.