**Extending R package “idefix” for generating optimal designs for discrete choice experiments**

* **Introduction**

*Paraphrasing Train, the easiest and most widely used discrete choice model is the logit.*

The use of discrete choice experiments in applied environments has increased in the last 20 years.

This kind of experiments are characterized by a discrete choice response, where it can be nominal or ordinal.

To collect information of this kind, the respondent are asked to select an alternative, composed by different combinations of attribute levels, from a series of choice sets.

Given that the number of possible profiles or alternatives explodes as the number of attributes and levels increase, the use of optimal designs, also known as efficient designs, is widely use in several scenarios.

There are different optimality criteria to choose a final design, namely D, A, G and V optimal, each using different methodologies (see Kessels 2006). Since the focus of the majority of applications is to obtain precise estimates the D-optimal criterion is commonly used.

Even though this kind of designs have widely been used in research and applied sciences, there are not implementations to generate this kind of models in R. What people actually do, is generate an optimal design for a linear model and use it to collect the data and analyze it with a discrete choice model. This approach can be improved in the sense that instead of using the information matrix of a linear model (usually OLS) to maximize its determinant, the information matrix of non-linear discrete choices models can be used, such as the multinomial logit. information matrix.

The idefix package, already available in CRAN, generates D-optimal (or DB-optimal) designs for discrete choice experiments based on the multinomial logit model assuming homogeneity in the respondents and implements the individually adapted designs for the mixed multinomial model when the assumption of homogeneity does not hold.

The current state of this package generates these optimal designs using the Modified Fedorov Algorithm and sequential modified Fedorov algorithm for the multinomial logit and mixed multimonial logit model, respectively. This algorithm is computationally expensive and make its application in real life a little bit difficult.

In the following, an introduction to the multinomial logit model is given, continued by an explanation of the Individually Adaptive Sequential Bayesian (IASB) approach to generate individual designs for each respondent. Then, the modified Fedorov algorithm is presented, as well as a the Kullback-Leibler and coordinate exchange algorithm, which are part of the research proposal. Finally, a description of a simulation study is presented together with a detailed description of the research proposal.

* **Multinomial logit model**

This model is derived from Mcfaddens random utility model, where it is assumed that a respondent makes its final choice by measuring the utility of each of the alternatives presented. This utility can be expressed as:

\begin{equation}

U\_j = x^’\_j\beta + \epsiolon\_j

\end{equation}

Where j represents the different alternatives shown within a choice set, $x\_j$ is the corresponding row of the j-th alternative/profile in the design matrix x, \beta denotes the importance of the attribute levels and \epsilon\_j is an i.i.d extreme value error term. Now, the multinomial logit probability that a respondent chooses the j-th profile in the s-th choice set is

\begin{equation}

p\_{js}(\textbf{\beta}) = frac{\exp{x^{‘}\_{js}\beta}}{\sum^{J}\_{i=1}\exp{x^{‘}\_{is}\beta}}

\end{equation}

Because the errors are assumed to be independent, the choices of N respondents represent independent draws from a multinomial distribution. Therefore, the log-likelihood can be written as

\begin{equation}

LL(Y|X,\beta) = \sum{S}\_{s=1} \sum{J}\_{j=1} \sum{N}\_{n=1} y\_{jsn} ln(p\_{js}(\beta))

\end{equation }

Where Y is the matrix of choices from N respondents with elements $y\_{jsn}$.

Since the Idefix package chooses the best design based in the D-optimal criterion, i.e., the choice that minimize the determinant of the variance-covariance matrix of the parameter estimators or conversely maximize the determinant of the Information Matrix, it is of special interest the expression of the Fisher Information matrix:

\begin{equation}

M(X,\beta) = N \sum{S}\_{s=1} X^’\_s(P\_s-p\_s p\_s^’) X\_s

\end{equation}

Where $p\_s = [p\_{1s}, … ,p\_{js}]$ and {P\_s=diag[p\_{1s}, … ,p\_{js}]}.

It is important to notice that the information matrix depends on the unknown parameters through the probabilities, so initial parameter values are required before computing this matrix. Kessels mentioned that a Bayesian design approach works in this scenario, specifying a prior normal distribution for \beta, N(\beta|\beta\_0,\Sigma\_0).

I dunno if I should explain here that the d-optimal criterion changes.

* **Individually Adaptive Sequential Bayesian for Mixed Multinomial logit**