

SIMULATED ANNEALING TECHNIQUE APPLIED TO MIXED LOGIT ESTIMATION

Julio C. Brégains^{(1),(2)}, J. Benito Pérez López⁽¹⁾, Alfonso Orro Arcay⁽¹⁾

jbregains@udc.es; jbperez@udc.es; aorro@udc.es

(1) Railways and Transportations Group, School of Civil Engineering, Campus de Elviña s/n – CP 15071 – La Coruña – Spain, Tlf: (+34) 981.167.000–1450, Fax: (+34) 981.167.170

(2) Electronic Technology and Communications Group, Department of Electronic and Systems, Faculty of Informatics, Campus de Elviña s/n – CP 15071 – La Coruña – Spain, Tlf: (+34) 981.167.000–1350, Fax: (+34) 981.167.160

ABSTRACT

In this work, and as an alternative to the traditional maximization techniques based upon gradients, the Simulated Annealing technique is applied to estimate a mixed logit model for transport mode choice. The method is applied to a specific case with real survey data, in order to analyze and discuss its advantages and drawbacks.

Keywords: Mixed Logit; Travel Behaviour; Simulated Annealing

INTRODUCTION

Discrete choice models based upon the random utility maximization theory are used in transport to study the behavior of travelers. Particularly, they are used to predict the choice of different modes, and, at a less extent, the choice of travels destination. In current practice, the most used models correspond to multinomial logit and nested logit. In the last decade, the popularity of the mixed logit has increased between professionals and researchers due to its flexibility (Mc-Fadden, 2000) and easy estimation (Hensher, 2003). By applying this model it is possible to specify the random heterogeneity in the users' tastes in a more flexible and direct way than with classical probit schemes (Daganzo et al., 1977), which represent the available alternative in the current technical literature.

Once a functional form for both the utility and the error is assumed, the estimation of the logit model is focused on obtaining the estimators of the parameters that specify such a utility. This can be done by a classical approach or with the help of bayesian techniques (Allenby, 1994). In the classical approach, the mixed logit parameters are usually estimated by maximizing the simulated log-likelihood on a sample of choices.

In the same way the log-likelihood function of the linear (on the parameters) multinomial logit model is globally concave –thus a single maximum being secured (Mc-Fadden, 1973)—, this is not held by other discrete choice models. In the case of mixed logit, even with a linear specification on its parameters, there is no guarantee of the existence of a unique maximum. For this reason, the use of numerical maximization techniques that maintain robustness even when dealing with non-concave optimizing functions becomes necessary (Walker, 2001:58). Likewise, it becomes necessary to verify that a global maximum has been obtained by performing the optimization with different starting points.

The classical maximization processes make use of known numerical techniques, which are constituted, mainly, by variations of the gradients method. Such techniques are relatively fast. Nevertheless, since they have to perform numerical first derivatives (i.e., gradients), and, in some cases, inverse Hessian matrices (which imply first and second numerical partial derivatives), in some cases they carry convergence problems, more often than not related with Hessian matrices determinants equal, or near, to zero. Another of the difficulties of the gradient methods lies on their easiness of stagnating into local extremes, something that becomes particularly restrictive in cases where the maximizing function possesses a considerable variability.

One of the alternatives proposed to solve the problems of deterministic techniques mentioned above is the use of stochastic methods, such as, for example, Monte Carlo (Metropolis, 1949), Simulated Annealing (Kirkpatrick, 1983), or Particle Swarm Optimization (Kennedy, 2001). All of these methods share the advantage of dealing adequately with the difficulties described above, sometimes at the price of being slower than their deterministic counterparts. In the field of the discrete choice models, this kind of techniques has been applied to a probit model (Liu, 2000), after adding a non-linear programming post-process to refine the search for the optimum solution.

Other maximization algorithms, with restrictions and with trust regions, applied specifically to mixed logit, such as BTR (Basic trust-region algorithm) and BTRDA (Basic trust-region algorithm with dynamic accuracy) (Bastin, 2004), have obtained significant reductions in the time computing. The BIOGEME software (see Bierlaire, 2003 and Bierlaire, Bolduc and Godbout, 2003) propose other estimation algorithms, including the possibility of setting additional restrictions. In order to solve unconstrained nonlinear problems whose second derivatives matrices are singular at a local solution, Bierlaire and Thémans (2009) propose an iterative procedure which allows identifying a singularity in the objective function and then artificially appending curvature to it during the course of the optimization algorithm.

In this work, the use of the simulated annealing as an alternative to the traditional methods of maximization of the log-likelihood function applied to modal split is proposed. Even this proposal does not intend to establish such a method as a general solution to the optimization problem, it could be considered to be perfectly adequate in certain cases, as an alternative to the other available techniques.

In order to study the abovementioned proposal, the work is divided into three parts.

The justifications are presented in the first part (Method), through a brief description of both the mixed logit applied to the transport planning –in particular, to modal split–, and the estimation of such a model (log-likelihood maximization) by means of the use of several deterministic techniques. Next, the simulated annealing method is described in more detail.

In the second part (Examples), the techniques are applied to a particular case, which corresponds to a set of data obtained by a real poll made in Santiago de Chile city (Ortúzar, 1983). This set is particularly adequate for testing and comparing the optimization methods described in the first part, since it has been found that, with it, some of the classical techniques experience convergence problems. The simulated annealing and the other techniques are then applied to this set, being the obtained results conveniently compared among themselves. Some of the models specifications (such as the number of utility random coefficients –which are considered to be linear in the attributes) are also further modified, in order to observe the different results, not only with the various optimization techniques, but with the same technique after varying these coefficients.

Finally, the Results are analyzed and the Conclusions are presented.

METHOD

Mixed Logit

Within the framework of the theory of maximizing the random utility of the discrete choice model, the individual is assumed to know the importance of the available alternatives and to always choose the alternative of greatest utility. For the analyst, this utility will have a representative component (V) and a random component (ε).

In the random coefficients approach (see Train, 2003), this decomposition is treated in a slightly different, although completely equivalent, way. One assumes that the utility of alternative j for individual n can be specified as

$$U_{nj} = V_{nj}(\beta) + \varepsilon_{nj} = \sum_{k=1}^K \beta_{kn} x_{knj} + \varepsilon_{nj} = \beta^T x_{nj} + \varepsilon_{nj} \quad (1)$$

where the coefficients β represent the individual's tastes, and vary within the population according to a distribution established by the analyst, with density function $f(\beta|\theta)$ dependent on certain underlying parameters θ (e.g., the distribution's mean b and standard deviation σ); the x 's are attributes of alternative j : characteristics of the trip or socioeconomic characteristics of the individual.

In the case applied to modal split in transports, the probability P_{nj} that an individual n chooses a transport j among J available options can be represented as:

$$P_{nj} = \int \frac{e^{V_{nj}(\beta)}}{\sum_{i=1}^J e^{V_{ni}(\beta)}} f(\beta | \theta) d\beta = \int \frac{e^{\beta^T x_{nj}}}{\sum_{i=1}^J e^{\beta^T x_{ni}}} f(\beta | \theta) d\beta \quad (2)$$

The mixing distribution $f(\beta | \theta)$ captures both the variance and the correlation of the non-observed features. In most applications of mixed logit, for f are used a variety of continuous distribution functions such as normal, log-normal, uniform, etcetera, even though discrete functions are also considered.

In general, the integral given in equation (2) does not lead to a closed form solution, and therefore its value is commonly estimated by numerical simulation. The symbolic solution is then replaced by the simulated probability SP_{nj} which is obtained by an adequate discretization of (1). If the V_{nj} function is specified by means of random coefficients, such discretization is performed by generating R values of β in $f(\beta | \theta)$, called β_r , obtained whether by using pseudo-random values or by generating them quasi-randomly, as, for example, by drawing Halton sequences (Halton, 1960). The simulated probability, which finally depends on the θ parameters of f , is the average value obtained by the equation:

$$SP_{nj}(\theta) = \frac{1}{R} \sum_{r=1}^R \frac{e^{\beta_r^T x_{nj}}}{\sum_{i=1}^J e^{\beta_r^T x_{ni}}} \quad (3)$$

Log-likelihood Maximization Techniques: Classical Approaches and Simulated Annealing.

The estimation of the θ parameters of the utility functions is performed by maximizing the simulated log-likelihood SLL (Train, 2003), in which the P_{nj} values are replaced by their corresponding SP_{nj} given in (3), considering all of the J alternatives and the N sample individuals:

$$SLL(\theta) = \sum_{n=1}^N \sum_{j=1}^J \ln SP_{nj}(\theta) \quad (4)$$

The usual maximization techniques are focused on procedures based upon the computation of gradients, generally of second order, making use of first and second numerical derivatives (gradients and Hessians) of the SLL function.

The Newton-Raphson (NR) procedure can be used, but it has two drawbacks; calculation of Hessian in each step and lack of guarantee of improvement in every step if the function is not globally concave. Variable metric algorithms can be used to achieve function increments. Methods of this type are Davidon-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS), that approximate the Hessian using several points of the likelihood function. However, Bernd, Hall, Hall and Hausmann method (BHHH) is used frequently in this field. This method utilizes the fact that the SLL function is the sum of observations log-likelihoods. The BHHH method uses the average outer product of the scores¹ as an approximation of the Hessian. All these methods are properly described in Train (2003).

The Simulated Annealing represents one of the alternatives to the classical optimization techniques. The name and operation of such an algorithm is based upon an analogy with the annealing of metals or the gradual cooling of crystalline structures applied to a material, in order to reduce its defects and to increase the size of its crystals (Kirkpatrick, 1983). The initial heat produces the disarrangement and the random change of the initial positions of the atoms (which corresponds to a given global internal energy of the material), by means of the generation of states of higher energies (excitation levels). The slow cooling the material is exposed to allows obtaining a configuration whose global energy is lower than the initial one.

To explain the operation of such a technique, it is useful to sketch a flux diagram, considering the special case in which a Boltzmann distribution is applied, and specified for the case of perturbing the $\beta_j \theta$ coefficients, which would lead to the maximization of the SLL, see Fig. 1.

Let us consider that the measured utilities V_{nj} consist of the sum of K terms, composed of Q fixed coefficients b_q , and D random coefficients of mean value b_d and standard deviation σ_d . Under these assumptions, the following conditions apply:

$$\begin{aligned} q \in \Omega_Q; d \in \Omega_D \\ \Omega_Q \cup \Omega_D = \{1, 2, \dots, K\}; \Omega_Q \cap \Omega_D = \emptyset \end{aligned} \tag{5}$$

where Ω_Q and Ω_D correspond to the indexes of the variables with fixed and random coefficients, respectively, whereas \emptyset represents the empty set.

In this way, the variables θ established in equations (1) to (4) are specified by a vector that contains the set of values b_q , b_d and σ_d .

Let us call θ_C the value of the vector θ that corresponds to some iteration of the SA algorithm, and θ_P the value of the same vector on a previous iteration.

¹ The score of an observation is the derivative of that observation's log-likelihood with respect to the parameters β evaluated at step's β .

The first step of the algorithm, see Fig. 1, consists of initializing conveniently the values of the θ_P vector, whether by assigning to it some values randomly chosen, restricted within the range $\theta_L \leq \theta_P \leq \theta_U$, or making it equal to any other vector the researcher considers to be more appropriate².

The value of the maximum log-likelihood LL_B the algorithm is searching for (i.e. best LL) has an initial value which is determined by $LL(\theta_P)$. Then, an initial temperature T_i is assigned to T , being the latter a variable, called Temperature, that will decrease in a factor $0 < \alpha < 1$ as the number of iterations increases, up to achieving a minimum (final) value given by T_f . The temperature reduction loop (see the dashed gray line on Fig. 1), establishes the log-likelihood value of the previous iteration $LL_P = LL(\theta_P)$. From the next step on, the vector θ_P is perturbed by a quantity $\delta\theta$, being this perturbation taken on a random basis³. The log-likelihood is then updated with this new value $LL_C = LL(\theta_C)$, which is compared with that of the previous iteration. If the current LL is better than the previous one ($LL_C \geq LL_P$), the solution is always accepted, and if, besides, LL_C is better than LL_B , then this configuration becomes the optimum solution at that iteration. The algorithm, nevertheless, allows accepting some cases in which the current solution is worse than the previous one (i.e. $LL_C < LL_P$), by applying the Boltzmann criterion, in which the value $e^{(LL_C - LL_P)/T}$ is compared with a value generated randomly (a value between 0 and 1 of a uniform standard distribution). This condition allows the algorithm to alter its searching regions, reducing, in this way, the possibility of stagnation in local extremes. As T lowers, nevertheless, this condition becomes more and more difficult to meet, because $e^{(LL_C - LL_P)/T}$ becomes smaller and smaller⁴. As a consequence of this, as the algorithm progresses, the θ jumps toward regions of lower LL become less frequent, with a tendency to allow only solutions that increase its value. For each temperature decreasing loop, ($T \leftarrow \alpha T$, i.e., the value of T is updated by setting it equal to αT), M perturbations of θ_P are performed. The algorithm stops when the final temperature T_f is achieved.

An example carried out with the described SA technique is presented below. The performance of such a technique is compared with the classical maximization routines given in the modal split estimation literature.

² The expression $\theta_L \leq \theta_P \leq \theta_U$ means that a bounded value between two extremes is assigned to each component of vector θ_P ; for example $-u \leq b_q \leq u$, $-v \leq b_d \leq v$ and $0 \leq \sigma_d \leq w$, with $u, v, w \in \mathcal{R}_{>0}$, $d \in \Omega_D$ y $q \in \Omega_Q$ ($\mathcal{R}_{>0}$ being the set of the real positive numbers).

³ The components of the vector $\delta\theta$ are quantities randomly generated under uniform standard distribution, conveniently biased and resized by the researcher.

⁴ The problems for meeting such a condition depend on how close T is to T_i , the latter being arbitrarily small. This condition is limited, naturally, by the computing capacities of the numerical tool used to run the algorithm. Such a tool is described upon section EXAMPLES, see below.

EXAMPLES

In order to show the scope and limitations of the proposed method, some examples are presented below.

Data description.

A set of transversal section data are considered, which were obtained by means of a real poll corresponding to the Las Condes-Center of Santiago de Chile corridor (Ortúzar, 1983). The set is composed of a sample of 622 interviewees, each one having at his/her disposal $J=9$ transport alternatives. The logit model applied is specified by considering the V_{nj} composed of a linear combination of $K=12$ descriptive variables (attributes), 4 of them being available for each individual. The 12 descriptive variables are distributed along 8 alternative specific constants and 4 additional values corresponding to travel time in vehicle (whether as driver, as bus/taxi/subway passenger, or as a combination of any of them), travel costs (costs per minute whether as driver, as bus/taxi/subway passenger, or as a combination of any of them), walking time (whether as driver or as passenger) and waiting time (as driver or as passenger). A more detailed description of the data can be found in the corresponding reference (Ortúzar, 1983). This database has been selected mainly because previous classic optimization procedures applied to it had demonstrated to have some problems to estimate the mixed logit parameters. The main aim of this investigation is to analyze the utility of simulated annealing as an alternative method for estimating these problematic cases.

For studying the fitting of mixed logit to the poll variables, it is possible to specify several configurations with random or fixed coefficients. This is described on next section.

Techniques and presented cases.

A broad set of classical optimization techniques based upon gradient methods (or its derivations) were exploited. All of them have been obtained from a set of routines developed previously by Keneth Train, David Revelt and Paul Ruud from the University of Berkeley, California (Mixed Logit Estimation Routine for Cross-Sectional Data, Train, Revelt and Ruud 1996, 1999), using Gauss, a well known matrix programming language (Aptech Systems, 2001, 2002). This code (TRR) is available for free on the internet, at the professor Train's webpage (<http://elsa.berkeley.edu/~train/software.html>). At the beginning of the main code of such tool it is indicated that the user can freely develop modified versions of it, but claiming these modified versions not being shared, if distributed, as alternative versions, but as new programs themselves. In such a case, the authors ask for the programmers to indicate in the code that it corresponds to a modified version of the original, the modifications to be itemized and the authors of such modifications to be specified as well. For carrying out the current research, the original TRR program has been modified by the authors (us) in order to incorporate the corresponding simulated annealing code.

With the purpose of performing an exhaustive comparison, all the routines available on TRR has been used. The nomenclatures used for naming all the routines are listed on Table 1.

The data has been estimated by considering four different coefficient distributions, in all cases the random coefficients sharing normal distribution, namely:

β_{12} : all of the coefficients are fixed, id est.: $\Omega_Q = \{1,2,3,4,5,6,7,8,9,10,11,12\}$. This configuration corresponds to a multinomial logit, whose estimation is very easy, but that was included to perform a first evaluation of the capability of the SA against the other routines. The results of this first estimation were then used as initial values of the remaining methods.

β_{11+1} : $\Omega_Q = \{1,2,3,4,5,6,7,8,9,10,11\}$, $\Omega_D = \{12\}$; this refers to a configuration with 11 fixed coefficients and a random one, which corresponds to the waiting time attribute.

β_{10+2} : $\Omega_Q = \{1,2,3,4,5,6,7,8,9,10\}$, $\Omega_D = \{11,12\}$; corresponds to random coefficients for waiting time and walking time.

β_{9+3} : $\Omega_Q = \{1,2,3,4,5,6,7,8,9\}$, $\Omega_D = \{10,11,12\}$; corresponds to random coefficients for travel cost, waiting time and walking time.

β_{9+1+2} : $\Omega_Q = \{1,2,3,4,5,6,7,8,9,11,12\}$, $\Omega_D = \{10\}$; corresponds to one random coefficient for travel cost.

β_{8+1+3} : $\Omega_Q = \{1,2,3,4,5,6,7,8,10,11,12\}$, $\Omega_D = \{9\}$; corresponds to one random coefficient for travel time in vehicle.

Results and discussion.

Table 2 shows the SLL values obtained for every configuration (Ω_Q, Ω_D), and specified for each of the different maximization techniques. The reported values refer to program executions with 400 repetitions ($R = 400$), and using Halton sequences; the limit in the number of iterations of any of the classical maximizations is taken to be 100. Further analyses have been performed with different values of R , diverse starting points, and up to a maximum of 500 iterations. In Table 2, below the *SLL* are written, within brackets, the computing times, stated in seconds⁵. The third row indicates the cause of sudden haltering of the classical algorithms. Case A indicates normal convergence. Case B denotes the algorithm haltered with no returned results as a consequence of reaching a singular numerical Hessian. Symbol C means the algorithm reached the (fixed) maximum number of iterations without converging; in this particular case it was verified that the algorithm was whether not able to achieve any improvements (the last value of the function is specified in

⁵ The computations have been performed with Gauss software tool, running in a laptop having two processors working at 2.4GHz, with an available RAM of 2GB under a 32-bit operating system.

the table) or not capable of returning reasonable solutions after discarding (numerically) too many observations (the last value of the function is not specified in this case).

The DFP (ML) technique has turned out to be inadequate due to numerical errors, and, for that reason, it was not included in the table. The SD(ML) and PRCG(ML) techniques have not reached convergence points (even after having been tested with 500 iterations as a limit).

From analyzing the data, it can be stated that the SA has performed function values close to the convergence point that the other algorithms reached almost homogeneously. It has been verified that the obtained parameter values were, likewise, very similar. In the β_{9+3} case, with which other algorithms were not able to converge, it has been found that the SA reached a function value that is above that of the halting point of the other methods.

From the behavior analysis point of view, it was found, by using customary tests, that there is no significant randomness in cases β_{11+1} (waiting time) and β_{8+1+3} (travel cost). For β_{10+2} , it is verified in the same manner that the standard deviation for the walking time is significant, but not for the waiting time, which is in accordance with the above results. It is equally significant the improvement and the deviation coefficient value in case β_{9+1+2} (in-vehicle time).

The behavior of the SA routine appears to be the most robust of them all, with incremental *SLL* values which seem to be related directly with the number and significance of the random coefficients, becoming a very feasible alternative for replacing, or complementing, the classical algorithms applied to log-likelihood maximization.

In contrast, nevertheless, it is necessary to comment that the algorithm reveals disadvantages that, even though are not severe, allows arguing that it deserves some further investigation in order to find some improvements to be presented in future works.

One of the disadvantages of the algorithm is its slowness (when compared to the deterministic techniques presented here) for finding a solution. It can be seen that the computing time ranges, for the given examples, from 2244.12 to 13601.76 seconds (37.4 minutes to 4.96 hours, respectively). This, however, does not constitute a real limitation: the algorithm largely compensates this slowness with a definite robustness, allowing the researcher to sacrifice time for the sake of guaranteed good results.

Another disadvantage points towards the necessity of adjusting of some initial parameters, i.e.: T_i , T_F , θ_{pi} and $\delta\theta$ range (see Fig. 1). Such adjustment influences not only on the algorithm results, but on its computing time. Unfortunately, there is not any systematic method for assigning such values, being the user compelled to a trial and error procedure until acquiring some experience that allows him estimating, with more or less success, the required values to reach the calibration that leads to optimum results. In most cases it is necessary to re-run the algorithm several times, due to the fact that the final result depends on the initial value of the variables θ_p . In the cases presented on Table 2, the final values were obtained after carrying out the routine twice or three times, taking the final θ_p of previous executions as initial values, and modifying, at the same time, the remaining initial

parameters (T_i , T_F and $\delta\theta$ range). The computing times given in Table 2 take into account those re-executions, being the total time obtained as the sum of the partial times.

A last disadvantage is the incapability for the final value to be really established as a maximum value, since the SA halts just after reaching a minimum temperature T_F , giving no possibility of knowing if the returned final value of the algorithm corresponds to a stationary point. Unfortunately, verifying this is not possible by calculating the numerical gradient, since this is the main fallibility reason of the other methods. In such cases it is possible to solve partially the problem by re-running the algorithm until no improvements are obtained. Further investigations could follow also this line for performing a final post-processing in order to guarantee the presence of a maximum and also obtaining error values for the estimators. As said before, it is worth noting that the additional efforts derived from these problems are compensated by the stability of the algorithm, something that can be observed from the results presented in this work, at least for the considered examples.

CONCLUSIONS

This work has presented an exploration about the utility of the simulated annealing as a tool prepared for estimating mixed logit discrete choice models applied to transport. For this purpose, the technique has been adequately tuned, and its effectiveness and feasibility for maximizing the simulated log-likelihood function has been shown by presenting some examples. At present, a weak point of this technique lies on its comparatively larger computing times needed for obtaining a final solution. In an analysis with real data, the algorithm has been proven to be robust enough for reaching a function maximum value, returning a solution in cases where the classical algorithms fail. In this sense, the proposed technique could be considered to be an alternative tool to be exploited in such cases, bearing in mind that a slightly deeper participation of the researcher is required.

ACKNOWLEDGEMENTS

This work has been supported by the Spanish Ministry Responsible for Public Works, under research project “SIMETRIA: Simulation Models for the Evaluation of Multimodal Transportation Scenarios Global and Regional” (P063/08).

REFERENCES.

- ALLENBY, G. and LENK, P. (1994), Modeling household purchase behavior with logistic normal regression, *Journal of the American Statistical Association* 89, 1218–1231.
- APTECH SYSTEMS (2001). *Maximum Likelihood Estimation for GAUSS (Version 5.0)*. Software Documentation, version June, 6, 2001.
- APTECH SYSTEMS (2002). *GAUSS User Guide*. Software Documentation, version October, 31, 2002.

- BASTIN, F. (2004). *Trust-region algorithms for nonlinear stochastic programming and mixed logit model*. Ph. D. Thesis in Mathematics. Facultés universitaires Notre-Dame de la Paix. Faculté des Sciences. Namur, Belgium, 12-03-2004.
- BIERLAIRE, M. (2003). *BIOGEME: a free package for the estimation of discrete choice models*. STRC 2003 (3rd Swiss Transport Research Conference) Monte Verita, Ascona, March 2003.
- BIERLAIRE, M., BOLDUC D. and GODBOUT M. H. (2003). *An introduction to BIOGEME (Version 0.7)*. December 17, 2003.
- BIERLAIRE, M. and THÉMANS, M. (2009). Dealing with singularities in nonlinear unconstrained optimization. *European Journal of Operational Research* 196, pp. 33–42
- DAGANZO, C. F., BOUTHELIER, F., and SHEFFI, Y. (1977). Multinomial Probit and Qualitative Choice: A Computationally Efficient Algorithm. *Transportation Science* 11: 338–358.
- HALTON, J. (1960). On the efficiency of evaluating certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik* 2, 84-90.
- HENSHER, D. A. and GREENE, W. H. (2003). The Mixed Logit model: The state of practice. *Transportation* 30, pp. 133–176.
- KENNEDY, J. and EBERHART, R. C. (2001). *Swarm Intelligence*. Morgan Kaufmann.
- KIRKPATRICK, S., GELATT, C. D., VECCHI, M. P., 1983, *Optimization by Simulated Annealing*, Science, New Series 220 (4598): 671–680.
- LIU, Y. and MAHMASSANI, H.S. (2000). Global maximum likelihood estimation procedure for multinomial probit (MNP) model parameters. *Transportation Research B* 34, pp. 419–449.
- MC FADDEN, D. (1973). Conditional logit analysis of qualitative choice behavior. En P. Zarembka (ed.), *Frontiers in Econometrics*, 105-142, Academic Press, New York, 1973.
- MC FADDEN, D. and TRAIN, K., 2000, *Mixed MNL Models For Discrete Response*, Journal of Applied Econometrics, 15 (5): 447–470.
- METROPOLIS, N. and ULAM, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association* 44 (247), pp. 335–341.
- ORTÚZAR, J.de D. and DONOSO, P. (1983). Survey design, implementation, data coding and evaluation for the estimation of disaggregate choice models in Santiago de Chile. *Second International Conference on New Survey Methods in Transport*. Sydney, 1983.
- TRAIN, K. 2003. *Discrete Choice Methods with Simulation*, Cambridge University Press.
- WALKER, J. (2001). Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables. Ph.D. Dissertation, Massachusetts Institute of Technology, February 2001.

FIGURES AND TABLES

Figure 1. Simulated Annealing flux diagram prepared for maximizing the log-likelihood (SLL) of the cases given in the examples (see text for variables description).

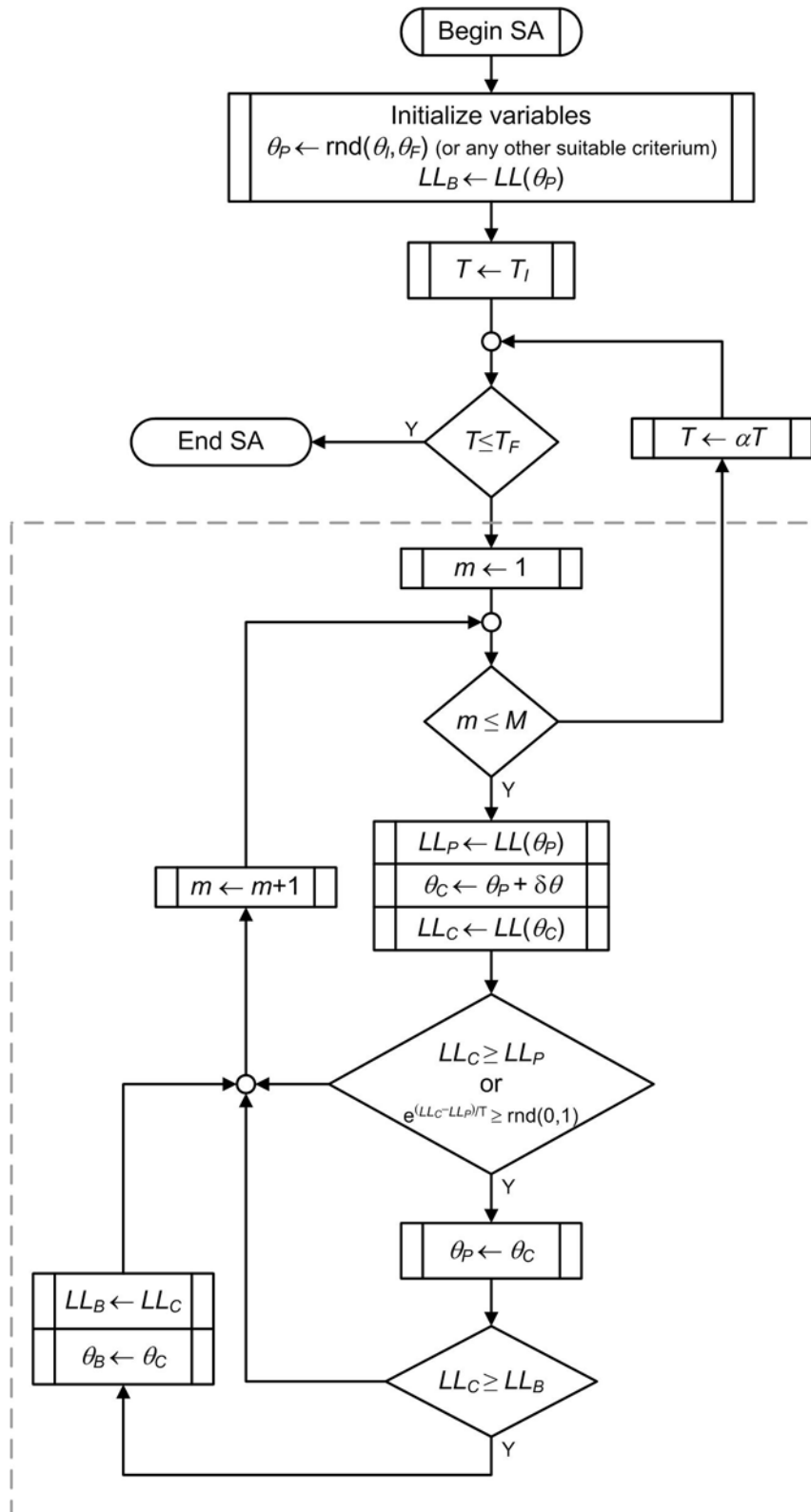


Table 1. Nomenclatures for each of the $SLL(\theta)$ maximization techniques used in the examples.

NOMENCLATURE	TECHNIQUE	DESCRIPTION
NR(PR)	Newton–Raphson	Routines developed by Paul Ruud and used in TRR code.
BHHH(PR)	Berndt, Hall, Hall and Hausman	
SD(ML)	Steepest–Descent	Routines included in the modulus for maximum likelihood estimation run under Gauss program (Aptech, 2001) and used in TRR code.
BFGS(ML)	Broyden, Fletcher, Goldfarb and Shanno	
DFP(ML)	Davidson, Fletcher and Powell	
NR(ML)	Newton–Raphson	
BHHH(ML)	Berndt, Hall, Hall and Hausman	
PRCG(ML)	Polak–Ribiere–type Conjugate Gradient	
SA	Simulated Annealing	Technique presented in this work, included in a separate version of the TRR code.

Table 2. Results of the simulated log-likelihood given for each coefficient distribution and obtained with the techniques described in this work. An XX indicates a case where the routine did not return any *SLL* value (unexpected halt). Below the *SLL* values, the algorithm computing times are given in seconds (in parenthesis). For each technique, third row gives the cause of the algorithm to halt (see text).

Maximization Technique	Coefficients Specification					
	β_{12}	β_{11+1}	β_{10+2}	β_{9+3}	β_{9+1+2}	β_{8+1+3}
NR(PR)	-869.984	XX	XX	XX	-864.951	XX
	(0.15)	--	--	--	(19.64)	--
	A	B	B	B	A	B
BHHH(PR)	-869.984	-869.983	XX	XX	XX	XX
	(0.19)	(33.14)	--	--	--	--
	A	A	B	B	B	B
SD(ML)	-885.573	-869.985	-865.239	-861.825	-866.141	-869.985
	(12.08)	(154.75)	(208.66)	(187.29)	(156.11)	(228.45)
	C	C	C	C	C	C
BFGS(ML)	-869.985	-869.985	XX	XX	-864.953	-869.985
	(1.27)	(22.15)	(852.95)	(1126.59)	(73.93)	(60.47)
	A	A	C	C	A	A
NR(ML)	-869.985	-869.985	-864.499	XX	XX	XX
	(0.33)	(43.73)	(86.14)	(2999.07)	(3866.08)	(9365.68)
	A	A	A	C	C	C
BHHH(ML)	-869.929	XX	XX	XX	-864.953	XX
	(0.41)	(1093.09)	(1320.81)	(1488.32)	(39.16)	(1011.81)
	A	C	C	C	A	C
PRCG(ML)	-885.573	-869.985	-865.239	-861.825	-866.141	-869.985
	(12.55)	(184.59)	(193.87)	(188.50)	(178.90)	(242.22)
	C	C	C	C	C	C
SA	-869.992	-869.987	-864.565	-860.099	-864.876	-870.007
	(2244.12)	(6251.36)	(6349.79)	(17854.96)	(16499.88)	(13489.36)