Estimating mixed logit with non-parametric random variables.

Fabian Bastin*, Cinzia Cirillo and Philippe L. Toint****

*CERFACS
42, Avenue G. Coriolis - 31057 Toulouse Cedex 01, France
**University of Namur Transportation Research Group
8, Rempart de la Vierge - 5000 Namur Belgium

Abstract

The estimation of random parameters by means of mixed logit models is becoming current practice amongst discrete choice analysts, one of the most straightforward applications being the derivation of willingness to pay distribution over a heterogeneous population. In many practical cases, parametric distributions are *a priori* specified and the parameters for these distributions are estimated. This approach can however lead to many practical problems. Firstly, it is difficult to assess which is the more appropriate analytical distribution. Secondly, unbounded distributions often produce values ranges with difficult behavioral interpretation. Thirdly, little is known about the tails and their effects on the mean of the estimates. (Hess et al, 2005; Cirillo and Axhausen, 2006)

This paper extends the nonparametric methods in a classical context of mixed logit models. The random variables of the objective functions are assumed to be continuous, bounded, and independent, and we are interested by the inverse cumulating distribution functions. These functions are modeled by means of cubic B-splines with strictly increasing base coefficients, a sufficient condition to construct monotonic (increasing) functions. As a result, the number of parameters that have to be estimated increases; the information on the tails and on the shape of the random variables however should help the analyst to find the right parametric distribution for the random parameters (if this exists).

This technique is applied to simulated data and the ability to recover both parametric and non-parametric random vectors is tested. The non-parametric mixed logit model is also used on real data derived from a survey on electric car, whose prototype has been realized and tested in a number of cities in Europe. The data set, which is part of a European study called “Cybercar” is a Stated Preference experiment conducted in Brussels in 2002. The model presents multiple choices and is estimated on repeated observations.

1. Introduction

In mixed logit models estimation, investigators traditionally use parametric models involving specific functional forms and a finite number of unknown parameters. The early applications of mixed logit have used normal distributions for partworths. The use of unbounded distributions appears inappropriate in many cases: certain attributes are assumed to be positively (or negatively) valued by all individuals; moreover, a zero cost coefficient causes problems for the evaluation of the willingness to pay. In order to circumvent these difficulties, more recent and sophisticated models propose the adoption of bounded distributions, often obtained as simple transformations of normals. Train and Sonnier (2003) specify mixed logit models with lognormal, censored normal and Sb distributions bounded on both sides. They also suggest to adopt Bayesian procedure in order to avoid estimation problems encountered with log-normal distributions parameters.

Some investigators have questioned whether the underlying theory is capable of conveying sufficient information to enable a correct and successful specification of parametric models and have instead proposed the less restrictive nonparametric or semiparametric approaches to the problem. In that context, Dong and Koppelman (2003) assume that distributions are represented by a finite number of points and use the Bayesian method to recover their mass and the associated probabilities. They assert that Maximum Likelihood Mixed logit failed to recover the true mass points from simulated data, although no reasons are given to explain that failure. The empirical analysis reported by the authors show that Mass Point Mixed logit is superior to Parametric Mixed logit; however, those results are limited by the use of only two points along each of the parameter dimensions.

Hess et al. (2005) propose discrete mixture of GEV models over a finite set of distinctive support points. The major advantage of this approach is the lack of need for simulation processes. However, it should be noted that when the number of discrete points to be estimated increases numerical problems related to the nature of the log-likelihood function to be maximized can be encountered.

Hensher (2006) resolves the problem of behaviorally sign changes by imposing a global sign condition on the marginal disutility expression and gives an application on the valuation of travel-time savings for car commuters. He adopts a globally constrained Rayleigh distribution for total travel time parameterization, although his focus is not on the specific analytical distribution, but on the behavioral appeal of the imposition of a global sign condition.

Train and Weeks (2004) place distributional assumptions on the willingness to pay and derive the distribution of the coefficients. They major finding is that models using normal and lognormal distributions for coefficients (models in preference space) fit the data better than those in willingness to pay space but provide less reasonable distribution for the willingness to pay. They also conclude that it is not possible to identify the distribution to use in all situations and that the best distribution-fit is highly situation-dependent.

Fosgerau (2006) employs various non-parametric techniques to investigate the distribution of the travel-time savings from a stated choice experiment. The adopted methodology relies on a two-step estimation procedure; first a Klein & Spady (1993) estimator is used to estimate parameters in a linear index binary choice model with no assumptions on the error term distribution; then the distribution of the error term is estimated. The proposed method does not account for repeated observations and applies only to binomial choices.

This paper proposes B-spline curves for non-parametric mixed logit models. In the literature B-splines are known to provide a concise formulation for curves that are composed of many polynomial pieces, thereby automatically controlling the overall curve smoothness. (Farin,1991). Spline smoothing has been applied to a wide range of problems in many disciplines since Whittaker (1923) first introduces it. This technique is often used for nonparametric regression. Recent applications in such areas as meteorology, medicine and price modeling can be found in Singh et al. (1997), Jarvis and Stuart (2001), and Bao and Wan (2004). To date there have only been a handful of applications of this approach in econometrics (Engle et al., 1986; Koenker et al., 1994; Craig and Ng, 2001). To our best knowledge, this approach is new for mixed logit models estimation.

The paper is organized as follow. Section 2 briefly recalls the Mixed logit model formulation and the estimation techniques adopted to solve the related maximum log-likelihood problem. Non-parametric estimation of continuous variables is developed on Section 3, together with a short review of the constrained optimization techniques adopted by the authors. Section 4 is dedicated to results obtained on simulated data and in particular we show the ability to recover both parametric and non-parametric random vectors. Preliminary results on a real case study are given on Section 5. Conclusions and future perspective of research are then presented.

2. Mixed Multinomial Logit Model (MMNL) formulation

The mixed logit formulation is nowadays extensively used in transport modeling for its flexibility. In particular, MMNL models estimate taste variation, avoid the problem of restricted substitution pattern in standard logit model and account for state dependency across observations. Mixed logit probabilities are expressed by means of the integral of standard logit probabilities over a density of parameters:

$$P_{ij} = \int L_{ij}(\beta) f(\beta) d\beta, \qquad (1)$$

where:

$i(i=1,\ldots,I)$ is the individual index,

$j(j=1,\ldots,J)$ is the alternative index,

$L_{ij}(\beta)$is the logit probability and,

$f(\beta)$ is a density function.

The mixed logit derivation that we will use in our application is based on random coefficients, with a joint distribution $f(\beta)$ that is usually assumed to be continuous. The choice probability is in this case:

$$P_{ij}(\theta) = \int L_{ij}(\beta) \phi(\beta \mid \theta) d\beta, \qquad (2)$$

where $\phi(\beta \mid \theta)$ is the density with parameters vector θ.

In the case an individual *i* chooses among alternatives $j=1,\ldots,J,$ in choice situations $t=1,\ldots,T_i$ (panel data), the utility of alternative *j* at time *t* can be expressed as:

$$U_{ijt} = \beta_i x_{ijt} + \varepsilon_{ijt}, \qquad (3)$$

where ε_{ijt} is an error term, assumed to follow the extreme value distribution and to be independently and identically distributed between alternatives, individuals and time periods, $\beta_i = g(\beta \mid \theta)$ is the vector of parameters randomly distributed in the population and x_{ijt} is the vector independent variables. We observe for each individual the sequence of choices $y_i = (j_{i1},\ldots,j_{iT\,i})$. The probability to observe the individuals' choices is given by the product of logit probabilities L_{it} (Train, 2003):

$$P_{iy_i}(y_i \mid x_i, \beta) = \int \left(\prod_{t=1}^{T_i} L_{it}(y_{it} \mid \beta) \right) f(\beta) d\beta \, . \qquad (4)$$

2.1 MMNL Model estimation

The vector of unknown parameters is estimated by maximizing the log-likelihood function, i.e. by solving the problem:

$$\max_{\theta} LL(\theta) = \max_{\theta} \frac{1}{I} \sum_{i=1}^{I} \ln P_{iy_i}(\theta), \qquad (5)$$

where y_i is the vector of alternative choices made by the individual *i*. This involves the computation of $P_{iy_i}(\theta)$ for each individual $i(i=1,\ldots,I)$, which is impractical since it requires the evaluation of one multidimensional integral per individual. To approximate the integral of the value $P_{iy_i}(\theta)$, a frequently used approach is to choose for each individual a point set , $S_R=\{u_1,\ldots,u_R\}\subset(0,1)^s$, where s is the problem dimension, i.e. the number of random coefficients, convert the vectors u_{r_i} to the (multivariate) distribution of β, and then take the average value of the function over S_R. This leads to the simulated probability

$$SP_{iy_i}^R=\frac{1}{R}\sum_{r_i=1}^{R}\prod_{t=1}^{T_i}L_{ij_{it}}\left(\beta_{r_i},\theta\right), \qquad (6)$$

where R is the number of random draws β_{r_i}, taken from the distribution function of β . As a result, θ is now computed as one solution of the simulated log-likelihood problem

$$\max_{\theta} SLL^R(\theta)=\max_{\theta}\frac{1}{I}\sum_{i=1}^{I}\ln SP_{iy_i}^R(\theta). \quad (7)$$

We will denote by θ_R^* one solution of this last approximation (often called Sample Average Approximation, or SAA), while θ^* denotes the solution of the true problem (5).

3. Non-parametric estimation of continuous variables

Most of the studies devoted to estimation of parameters without assumptions on the underlying distributions are concerned with discrete distributions. Such a discrete treatment could lead to an arbitrary population segmentation, which can be avoided if we turn to continuous distributions.

If the parameters coefficients are random, a practical way to approximate the log-likelihood function is to construct the set S_R, by sampling the random vector distribution, using Monte Carlo or quasi-Monte Carlo techniques. Each component of the random vector is itself random, and if we assume independence between these components, we can consider each one separately. Then all what we have to do is to draw from univariate random variables. If *X* is an univariate random distribution, a well-known technique to generate draws from its distribution consist to sample an uniform on [0,1], hereafter denoted by *U*[0,1], and to apply the inverse cumulative distribution function F_X^{-1} to these draws:

$$S_X=\left\{F_X^{-1}(u),\, u\sim U[0,1]\right\},$$

where S_X represented the sample set drawn from the random variable *X*. It is usually assumed that F_X is available, the distribution *X* being known.

We will capitalize on this technique by assuming that the distribution of the random variable *X* is not known, but that F_X, or more precisely F_X^{-1}, can still be approximated in some way. If *X* is a random continuous variables, the only properties that F_X^{-1} has to satisfy are:

- $F_X^{-1}:[0,1]\rightarrow\Re,$

- F_X^{-1} is monotonically increasing,
- F_X^{-1} is continuous.

In other terms, we have to estimate an arbitrary continuous real function whose domain is [0,1], which is monotonically increasing.

Functions approximation is a large field of mathematics, and various technique are possible. We however seek an adequate balance between estimation capabilities and satisfaction of the conditions ensuring that we can interpret the function as an inverse cumulative distribution function. If we furthermore assume that the random variable *X* has a bounded support, an elegant way to achieve such a balance is the use of B-spline functions. The bounded support assumption is not too much restrictive, since extreme behaviours, corresponding to values of *X* tending to plus or minus infinity, are usually not welcome since that are difficult to interpret, and may produce failures inside the optimization process. We therefore consider the bounded support assumption as an advantage rather than a drawback of the proposition.

A B-spline function of degree *p* is a polynomial function of degree *p*, defined on the interval $[a,b]$, can be expressed as linear combination of *n*+1 basis functions $N_{i,p}(u)$, as follows:

$$C(u) = \sum_{i=0}^{n} N_{i,p}(u) P_i .$$

The coefficients $P_0, P_1, \ldots, P_n$ are called the control points, and u is the knot vector $(u_0 = a, u_1, \ldots, u_m = b)$. The basis functions can be constructed by recurrence on the degree *p*:

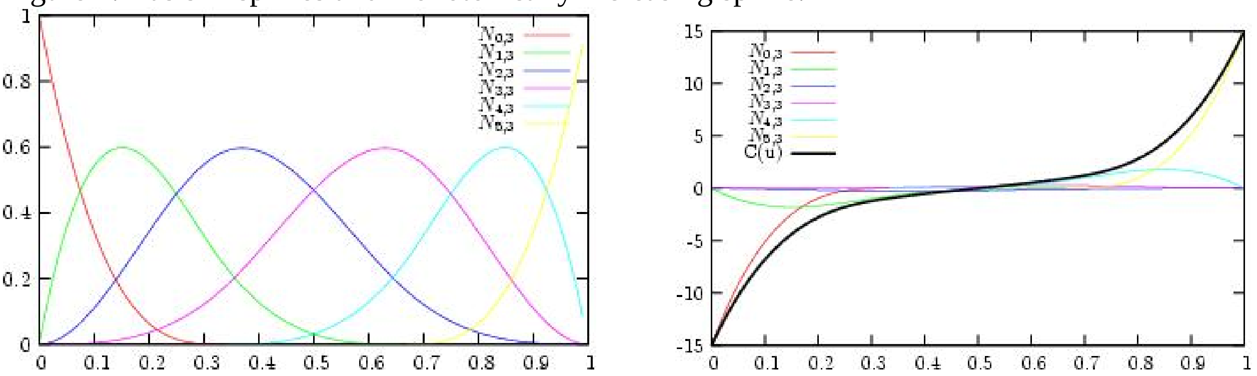
$$N_{i,0} = \begin{cases} 1 & \text{if } u \in [u_i, u_{i+1}), \\ 0 & \text{otherwise,} \end{cases}$$

and

$$N_{i,p} = \frac{u - u_i}{u_{i+p} - u_i} N_{i,p-1}(u) + \frac{u_{i+p+1} - u}{u_{i+p+1} - u_{i+1}} N_{i+1,p-1}(u)$$

The spline construction is illustrated in Figure 1.

Figure 1: Basis B-splines and monotonically increasing spline.



There are several types of knot vectors, but one especially convenient for our purposes is the *nonperiodic* (or *clamped* or *open*) knot vector, which has the form

$$U = \left\{ \underbrace{a,\ldots,a}_{p+1}, u_{p+1}, \ldots, u_{m-p-1}, \underbrace{b,\ldots,b}_{p+1} \right\},$$

that is the first and last knots have multiplicity p+1.

It is possible to show that the function $C(u)$ is p-1 continuously differentiable. In this paper, we will consider cubic B-splines, i.e. we will set p to 3. Another particularly nice property with respect to our needs is then that $C(u)$ is monotonically increasing if $P_0 \leq P_1 \leq \ldots \leq P_n$. As we will describe latter, this property that is quite easy to guarantee inside the estimation process.

For a more complete review of B-splines properties, we refer the reader to Piegl and Tiller (1996).

3.1 Constrained Optimization

When estimating the log-likelihood function, we have to solve a problem of the form

$$\min_{x \in C} f(x),$$

where $C \subset \Re^n$ is the feasible region. In our case, C will have the form

$$C = \left\{ x_{mi} \leq x_{(m+1)i} \leq \ldots \leq x_{(m+k)i} \right\},$$

That is C is the set of real numbers satisfying the monotonicity constraints. Such constraints can be easily dealt with projections. For simplicity, assume that we only have one non-parametric coefficient, so that C defines k ordered variables. C is then call the order-simplex; Figure 2 illustrates the order-simplex in three dimensions.

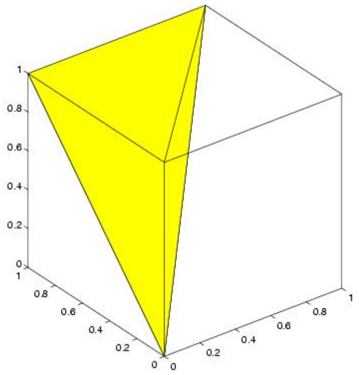


Figure 2: Order-simplex

The projection onto the order-simplex can be performed easily and efficiently, since several algorithms of complexity O(n) have been designed (Best and Chakravarti, 1990). Moreover, it is possible to adapt the trust-region approach to benefit from such projections. The main idea of a trust-region algorithm involves the calculation, at iteration k (with current estimate θ_k), of a trial point $\theta_k + s_k$ by approximately minimizing a model m_k of the objective function inside a trust region defined as

$$B_k = \left\{ \theta \text{ such that} \left\| \theta - \theta_k \right\| \leq \Delta_k \right\}, \qquad [13]$$

where Δ_k is called the trust-region radius. We will here use a quadratic model:

$$m_k(s) = SLL^R(\theta_k) + s^T \nabla_\theta SLL^R(\theta_k) + \frac{1}{2} s^T H_k s, \qquad [14]$$

where H_k is a symmetric approximation of the Hessian $\nabla^2_{\theta\theta} SLL^R(\theta_k)$. The predicted and actual decreases in the value of the objective function are then compared by computing the ratio

$$\rho_k = \frac{SLL^R(\theta_k + s_k) - SLL^R(\theta_k)}{m(\theta_k + s_k) - m(\theta_k)}, \qquad [15]$$

If this ratio is greater than a certain threshold, set to 0.01 in our tests, the trial point becomes the new iterate, and the trust-region radius is (possibly) enlarged. More precisely, if ρ_k is greater than 0.75, we set the trust-region to be the maximum between Δ_k and $2s_k$, otherwise we set $\Delta_k = 0.5\Delta_k$. If the ratio is below the bound, the trial point is rejected and the trust region is shrunk by a factor of 2, in order to improve the correspondence of the model with the true objective function. We have followed Conn et al. (10) in our choice of the parameters.

The only difference between the unconstrained case and the constraint case using projections lies in the computation of the step s in [14]. An efficient technique for the unconstrained case is the computation of the Steihaug-Toint point, that is an approximate truncated conjugated gradient minimizer of the model [14]. More precisely, we let the conjugated gradient method running until a sufficient decrease of the model has been achieved, or the boundary of the trust-region has been hit. The constrained case is managed by projecting the conjugated-gradient path onto the order-simplex, as illustrated in Figure 3.

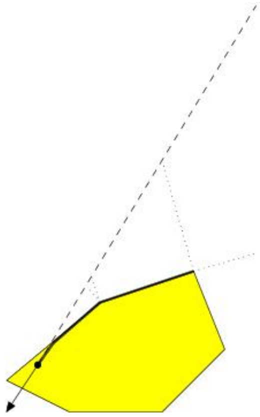


Figure 3: Projected Conjugated Gradient Path

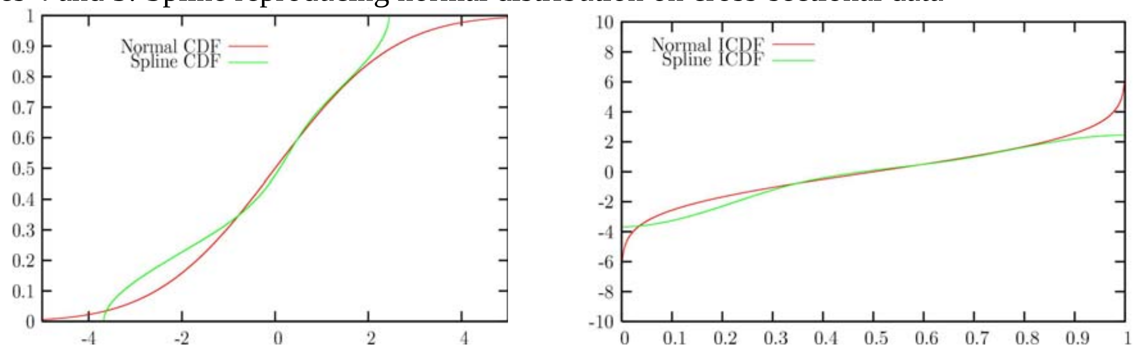
4. Simulations

In our simulated experiments we create two synthetic populations; the first data set is cross sectional and simulates 2000 individuals giving just one response, the second data set is a panel of 1000 individuals contributing with 2 observations each. The design contains 4 alternatives and 1 independent variable normally distributed with parameters N(0.5,1). We run two simulations on each of the data sets described; one supposing that the coefficient to be estimated was normally distributed with parameters N(0,4), the other assuming that the coefficient was lognormal distributed with parameters LN(1.133,0.604), where the first parameter is the mean of the lognormal distribution, and the second is its standard deviation. The main objective of our simulation study was to assess the ability of B-splines to recover the true (and known) distributions of our coefficient.

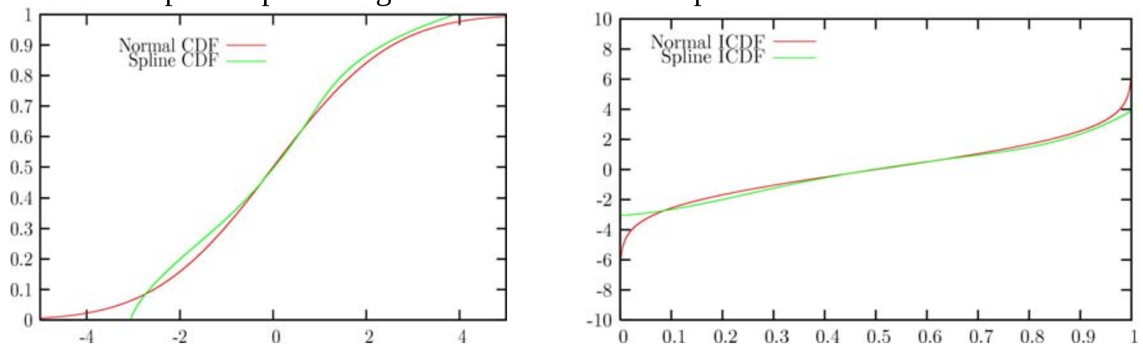
Results are represented on the following 8 Figures, we always report the cumulative distribution function (CDF) on the left and the inverse of the cumulative distribution function (ICDF) on the right.

As far as normal distribution is concerned, we note that B-splines approximate quite well the true distribution of the coefficient except on the tails. This should be expected since we approximate an unbounded distribution with a bounded distribution. The approximation is less accurate when trying to reproduce a coefficient with lognormal distribution, but the general behaviour is captured. In both cases results are better with panel data.

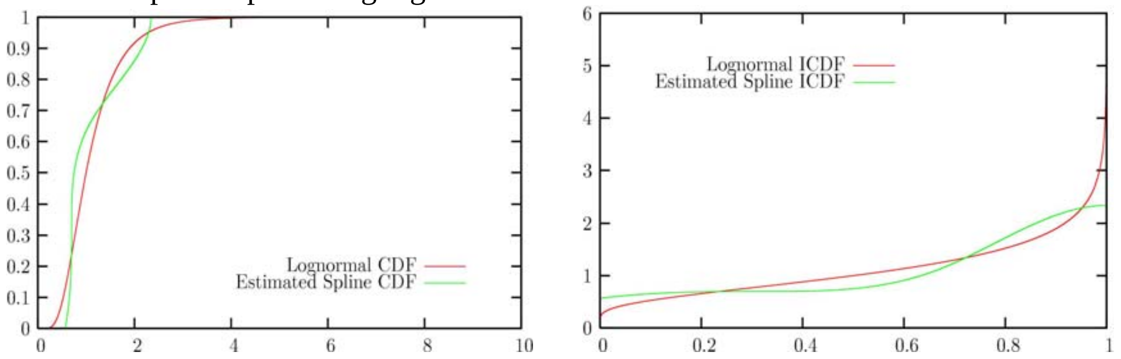
Figures 4 and 5: Spline reproducing normal distribution on cross-sectional data



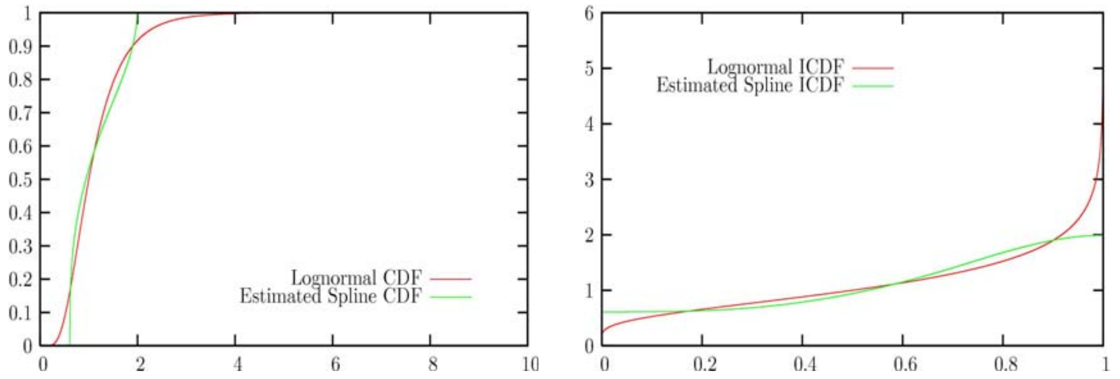
Figures 6 and 7: Spline reproducing normal distribution on panel data



Figures 8 and 9: Spline reproducing lognormal distribution on cross-sectional data



Figures 10 and 11: Spline reproducing lognormal distribution on panel data



5. Real case study: CyberCar survey.

Stated preference data are used to validate our methodology on a real case. The survey was conducted in Brussels (Belgium) in the first trimester 2003. The study was financed by the EU IST project called CYBERCAR. A consortium of 15 European countries has worked with French research institution INRIA to develop a self-steering CyberCar, which is currently being tested along the Riviera in southern France. The development and adoption of vehicles running autonomously without a driver on city streets at low speed (up to 30 km/h at the moment), while avoiding fixed and mobile obstacles, is the main goal of that project. More specifically, the survey aimed at estimating travel demand and willingness to pay for small and automated electric cars. It is an adaptive questionnaire administrated by professionals with the aid of the personal computers. The respondents, interviewed at their domicile, analyzed up to 18 scenarios based on the reference trip indicated and expressed their choices. The design contains 4 games, each of them proposes up to 4 alternatives on the screen. In total seven alternatives where available: car as driver, car as passenger, public transport, cyber-car, car sharing, bike and walk.

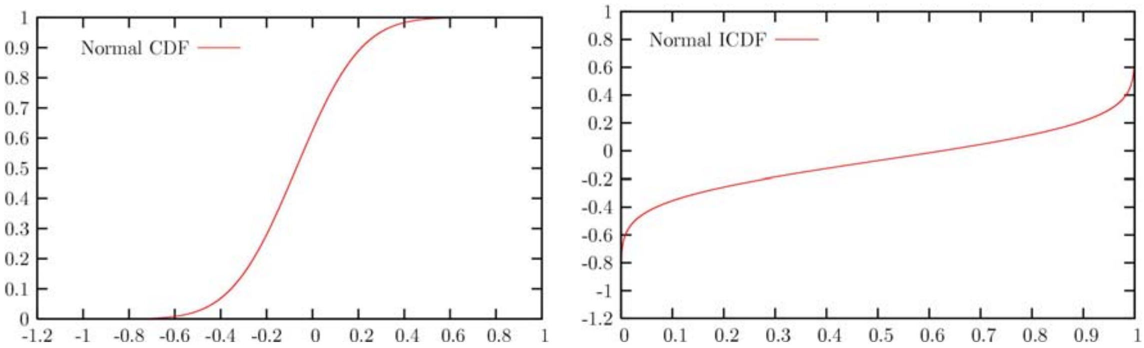
306 individuals completed the survey, giving a total number of 4824 observations for the estimation of the model. For the purpose of this paper we decided to estimate a model containing just time as independent variable. Although the model is partial, the extension to more complex specifications is quite straightforward. The levels adopted to derive travel time for each alternative in the experimental design are given in Table 1.

Table 1: Cybercar survey design

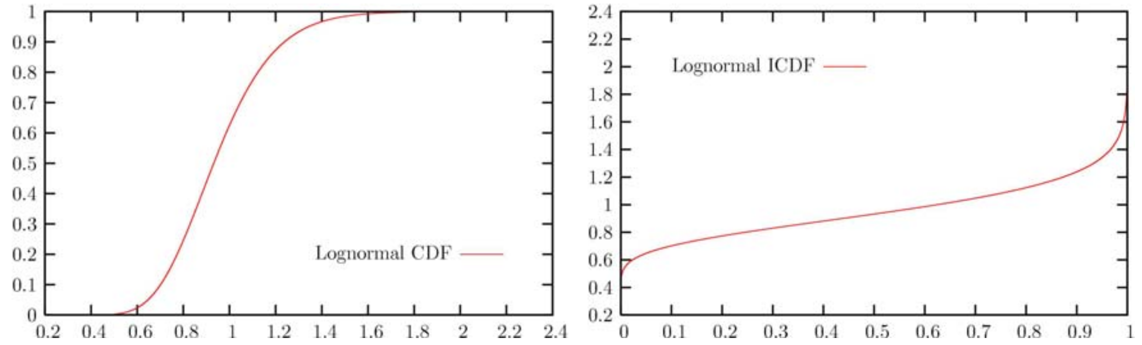
Variables	**Level of variation**		
Public Transport (PT)			
PT travel time	0	-15%	-30%
Car			
Car travel time	0	+25%	+50%
CyberCar			
CyberCar travel cost	= taxi fare	75% taxi fare	50% taxi fare
Carsharing			
Carsharing cost	0.30 EUR/km + 10 EUR	0.30 EUR/km + 15 EUR	0.30 EUR/km + 20 EUR

We report the estimation results in the following Figures; the three sets plot the time distribution estimated using respectively: normal distribution, lognormal distribution, B-spline. The final values of the log-likelihood are shown on Table 2. The goodness of fit indicator clearly indicates that the best model is the one using B-spline to estimate time distribution. It is also evident that the three results are very different from each other, and that what we assume to be the more realistic distribution (non parametric distribution) is very difficult to recover with classical parametric distributions. The sign of the observation has been set to negative values for the lognormal distribution, in order to cope with the positive sign of the random variable. However the final log-likelihood value suggests a poor adjustment of the distribution to the data. The value obtained with the normal distribution is much better, but an unexpected high number of respondents present a positive coefficient. The spline approximation manages to improve the log-likelihood value while giving a plausible behaviour of the coefficient sign, since nearly all individuals have a negative coefficient.

Figures 12 and 13: Real data – Time assumed to be normal distributed



Figures 14 and 15: Real data – Time assumed to be lognormal distributed



Figures 16 and 17: Real data – Time estimated with B-splines

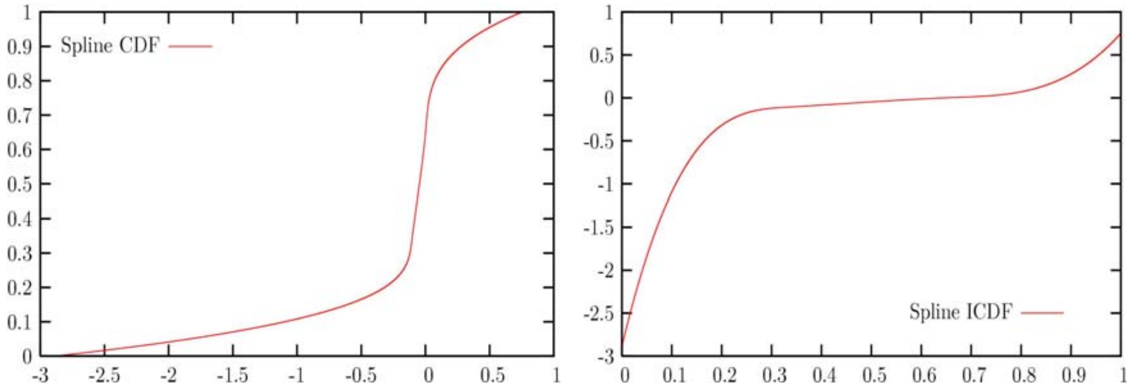


Table 2: Real data – Final log-likelihood values for the three estimated models

Distribution	Normal	Lognormal	B-Spline
Final Log-likelihood	-5161.33	-5673.43	-5120.72

6. Conclusions

Travel time variability has become one of the most debate full subjects in travel behavior. This problem is often approached with advanced demand models that allow the estimation of random coefficients with parametric distribution. This approach is not without drawbacks, especially since the distribution choice is not always clear. We have proposed to turn to nonparametric methods by adopting B-spline curves as polynomial approximations of arbitrary distributions, and we have implemented them into classical mixed logit formulation. Constrained optimization methods are used to deal with the monotonicity of the inverse of the cumulative distribution functions.

We have shown that parametric approach can fail to detect the real distribution and that non-parametric random variables could guide the analysts in search for the real shape of time distribution. Preliminary results on real data are extremely encouraging; not only the goodness of fit of the non-parametric model is highly better, but it also gives time distribution that would be very difficult to recover with classic parametric distributions.

Extension of that work to more complex models on real data is desirable. Comparison with bounded parametric distribution, such as censored normal and Sb-Johnson is also planned.

References:

Bao, H. X. H. and A.T.K. Wan (2004) On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data. *Real estate economics*, 32(3), 487-507.

Cirillo, C. and K.W. Axhausen (2006) Evidence on the distribution of values of travel-time savings from a six-week travel diary. *Transportation Research* **40A**, 444-457.

Conn, A., and, Gould, N.I.M., and Toint, Ph.L. (2000), Trust-Region Methods, SIAM, Philadelphia, USA.

Dong, X. and Koppelman, F. S. (2003), *Mass Point Mixed Logit Model: Development and Application*, paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, Switzerland.

Engle, R.F., C.W.J. Granger , J. Rice and A. Weiss (1986), “Semiparametric estimates of

Farin, G. (1991), NURBS for curve and surface design. SIAM Activity Group on Geometric Design.

Fosgerau, M. (2005), Investigating the distribution of the value of travel time savings, *Transportation Research* **41B**, 688-707.

Hensher, D.A. (2006) The Signs of the Times: Imposing a Globally Signed Condition on Willingness to Pay Distributions, Transportation, 33(3) 205-222.

Hess, S., M. Bierlaire and J.W. Polak (2005) Estimation of value of travel-time savings using mixed logit models. *Transportation Research* **39A**(3), 221-236.

Hess, S., M. Bierlaire and J.W. Polak (2005a) Discrete mixture of GEV models, paper presented at 5th Swisss Transport Research Conference, Monte Verità / Ascona, March 2005

Jarvis, C.H. and N. Stuart (2001), “A comparison among strategies for interpolating maximum and minimum daily air temperatures, part II: the interaction between the number of guiding variables and the type of interpolation method”, *Journal of Applied Meteorology* 40, 1075-1084.

Klein, R. and R. Spady (1993) An efficient semi-parametric estimator for binary response models. Econometrica 61 (2), 387-422.

Koenker, R., P. Ng and S. Portnoy (1994), “Quantile smoothing splines”, *Biometrika* 81, 673-680.

Piegl, L. A. and Tiller, W. (1996), The NURBS Book, second edition, Springer-Verlag, New York, NY, USA.

Singh, G.D., J.A. McNamara and S. Lozanoff (1997), “Morphometry of the cranial base in subjects with class III malocclusion”, *Journal of Dental Research* 76, 694-703.

the relation between weather and electricity”, *Journal of the American Statistical Association* 81, 310-320.

Train, K. and M. Weeks (2005) Discrete Choice Models in Preference Space and Willingness-to-Pay Space, Ch. 1, pp. 1-17, in Applications of Simulation Methods in Environmental Resource Economics, A. Alberini and R. Scarpa, (eds.), Springer-Verlag: Dordrecht, The Netherlands.

Train, K. and G. Sonnier (2005) Mixed Logit with Bounded Distributions of Correlated Partworths, Ch.7 pp.117in A. Alberini and R. Scarpa (eds.) *Applications of Simulations Methods in Environmental Resource Economics*, Kluwer Academics Publisher, Dordrecht, The Netherlands,

Whittaker, E.T. (1923), “On a new method of graduation”, *Proceedings of the Edinburgh Mathematical Society* 41, 63-75.