Empirical Identifiability of Latent Class Models: Theoretical Analysis and Application to Multiple Heuristics' Modelling

F. Gonzalez-Valdes[1]; B.G. Heydecker[2] and J. de D. Ortúzar[3]

[1] Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile; e-mail: fagonzalezv@uc.cl

2 Centre for Transport Studies, University College London; e-mail: b.heydecker@ucl.ac.uk

[3] Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), Pontificia Universidad Católica de Chile, e-mail: jos@ing.puc.cl

Email for correspondence: jos@ing.puc.cl

Abstract

Latent class (LC) discrete choice models have been found to exhibit identifiability problems. Theoretical identifiability addresses this issue in general, but no empirical identifiability analysis has been performed previously for these models. Here, we analyse the identifiability of LC models and through this, establish that differences among classes are crucial in identification. We quantify the relationship between behavioural difference and empirical identifiability using maximum likelihood analysis, and proceed to show empirically that is informative in Bayesian estimation. Then, we simulate a common scenario of potential non-identifiability with multiple choice heuristics in a real transport mode choice context. Based on our simulation results, we show that Bayesian estimation procedures are more robust than likelihood maximisation whilst recovering our main results. We show a graphical diagnostic for identifiability and provide examples of model non-identifiability, weak identifiability and strong identifiability.

1. Introduction

Latent class (LC) models are finite mixture models representing several clusters of individuals (Kamakura and Russell, 1989). They have been widely applied either exclusively with exogenous variables (Rossetti et al., 2018) or in conjunction with latent variables in a MIMIC model (Hess and Stathopoulos, 2013); with diffuse choice sets (Ben-Akiva and Boccara, 1995); and either using exclusively utility maximization heuristics or allowing different choice heuristics within each latent class (Hess et al., 2012; Gonzalez-Valdes and Raveau, 2018).

For LC models to be useful, they must be identifiable. Walker and Ben-Akiva (2002) investigated theoretical and empirical identifiability. Here, we focus on the latter, where in theory the model can be identified, but due to the model structure and data, the Hessian matrix is singular or nearly so (Cherchi and Ortúzar, 2008). Huang and Bandeen-Roche (2004) explored theoretical identifiability in LC models specifying conditions for identifiability in each of the components of a latent class – latent variable choice model. Yet, conditions for empirical identifiability when no latent variables are used have not been addressed completely.

Among the applications of LC models, one of the greatest challenges in identifiability arises when multiple heuristic choice models are considered. Indeed, these LC models have resorted to latent variables (Hess and Stathopoulos, 2013) and normalizations (Leong and Hensher, 2012) for identifiability. Therefore, our empirical application will focus in this challenging context.

To understand empirical identifiability of LC models, first we develop a theoretical framework to analyse the interaction of the governing forces of identifiability. Then, we conduct Monte Carlo simulation experiments in a realistic transport context to assess the drivers of identifiability. The simulation of latent classes is performed in the context of modelling with multiple choice heuristics. We explore three meta-experimental variations: the type of choice heuristics, the proportions of each heuristic in the sample, and the correlation between the probability of using a given heuristic and the individual preferences. Our results provide a framework for practitioners to design experiments for LC models and explore the identifiability of multiple-choice heuristic models in practical applications.

2. Theoretical Analysis

2.1 Binary Case

2.1.1 The balance of latent classes

Suppose that individuals align their behaviour to either one of two latent classes, denoted as *a* and *b*, with probabilities π_a and $\pi_b = (1 - \pi_a)$ respectively. Let $P_{cqi}(\theta)$ be the probability that individual *q* chooses alternative *i* using parameters θ conditional on belonging to class *c*. Then $P_{qi}(\theta)$, the probability of choosing alternative *i* under this binary LC model, is given by (1):

$$P_{qi}(\theta, \pi_a) = \pi_a P_{aqi}(\theta) + (1 - \pi_a) P_{bqi}(\theta) \; . \tag{1}$$

The log-likelihood of this model with (θ, π_a) is given by (2), where $P_{cq*}(\theta)$ represents the probability that individual *q* would have chosen the selected alternative aligning their behaviour to latent class *c*:

$$l(\theta, \pi_a) = \sum_q \log\left(\pi_a \, P_{aq*}(\theta) + (1 - \pi_a) \, P_{bq*}(\theta)\right) \tag{2}$$

The maximum value of this function could arise either at a boundary or at an interior value of π_a. In the first case (i.e. $\pi_a \in \{0,1\}$), the optimal model consists of a single latent class. Whereas in the case of an interior solution (i.e. $\pi_a \in (0, 1)$), the two classes of individuals coexist in a mixture model. The solution depends upon the balance between the losses and gains in likelihood associated with including an additional class in the model and, therefore, reducing the proportion of the complementary one. A boundary solution will be obtained when it is optimal for the model to consider a single class of individuals, that is, when the improvements in likelihood from the inclusion of the other class do not compensate for the losses.

In the more interesting case of an interior solution, the likelihood is maximised when the likelihood function is stationary with respect to variations in the class membership probability π_a. This can be detected as an interior point at which the derivative of the log-likelihood function equals zero. Among the variables to examine, an interesting one is precisely π_a, because it indicates the proportion of each class and, therefore, connects them in the model. The first order condition regarding π_a is analysed next.

We start by considering the case where the class membership function π_a is constant across the population (i.e. the probability of class membership is the same for every individual). For the context of multiple-choice heuristics that we explore later, this is the most frequent formulation (Balbontin et al., 2017; Hess et al., 2012). Under this specification the following theorem describes the optimality for two class estimation or coexistence of the two latent classes:

THEOREM 1: Two latent classes coexist optimally in a discrete choice model with constant class membership probability if the vector θ of estimated parameters satisfies the following balance:

$$\sum_q \frac{P_{aq*}(\theta)}{P_{q*}(\theta)} = \sum_q \frac{P_{bq*}(\theta)}{P_{q*}(\theta)} \tag{3}$$

where $P_{q*}(\theta) = \pi_a P_{aq*}(\theta) + (1-\pi_a) P_{bq*}(\theta)$ represents the modelled probability that individual q chooses the alternative actually chosen.

PROOF: For an interior solution, the first order condition for the maximisation is given by (4):

$$\frac{\partial l(\theta,\pi_a)}{\partial \pi_a} = \sum_q \frac{P_{aq*}(\theta) - P_{bq*}(\theta)}{\pi_a\, P_{aq*}(\theta) + (1-\pi_a)\, P_{bq*}(\theta)} = 0\,. \tag{4}$$

Manipulation of (4) leads to (5):

$$\sum_q \frac{P_{aq*}(\theta)}{\pi_a\, P_{aq*}(\theta) + (1-\pi_a)\, P_{bq*}(\theta)} = \sum_q \frac{P_{bq*}(\theta)}{\pi_a\, P_{aq*}(\theta) + (1-\pi_a)\, P_{bq*}(\theta)}\,. \tag{5}$$

Using the definition of $P_{q*}(\theta)$, this is equivalent to (3).

Equations (3) and (5) indicate that when it is optimal for the model to include both latent classes, there is a balance between them. This balance is given by the sum of the ratio of the likelihoods of the class to the complete model, showing that the ratio of class-conditional to marginal probability of the observed choices is crucial. The magnitude of this sum is described by Theorem 2:

THEOREM 2: Two latent classes coexist optimally in a discrete choice model with constant class probabilities if the balance quantity in (3) with likelihood maximising parameters θ is equal to the sample size Q .

PROOF: Expanding the left-hand side of (3) leads to (6):

$$\sum_q \frac{P_{aq*}(\theta)}{P_{q*}(\theta,\pi_a)} = \sum_q \frac{\pi_a\, P_{aq*}(\theta)}{P_{q*}(\theta,\pi_a)} + \sum_q \frac{(1-\pi_a)P_{aq*}(\theta)}{P_{q*}(\theta,\pi_a)}$$

$$= \sum_q \frac{\pi_a\, P_{aq*}(\theta)}{P_{q*}(\theta,\pi_a)} + \sum_q \frac{(1-\pi_a)P_{aq*}(\theta)}{P_{q*}(\theta,\pi_a)} + \sum_q \frac{(1-\pi_a)P_{bq*}(\theta) - (1-\pi_a)P_{bq*}(\theta)}{P_{q*}(\theta,\pi_a)}$$

$$\Rightarrow \sum_q \frac{P_{aq*}(\theta)}{P_{q*}(\theta,\pi_a)} = \sum_q \frac{\pi_a\, P_{aq*}(\theta) + (1-\pi_a)P_{bq*}(\theta)}{P_{q*}(\theta,\pi_a)} + (1-\pi_a)\sum_q \frac{P_{aq*}(\theta) - P_{bq*}(\theta)}{P_{q*}(\theta,\pi_a)} \tag{6}$$

According to equation (1), every term in the first summation of the right-hand side of (6) is identically equal to one, therefore the summation adds to Q. The second summation is equal to zero because of stationarity (4) for the likelihood maximising parameters θ . Because of (3) and in light of the symmetry between the latent classes, the condition corresponding to class a applies equally to class b . Then, (7) describes the balance in a model with two latent classes and constant class membership function:

$$\textstyle\sum_q \frac{P_{aq*}(\theta)}{P_{q*}(\theta,\pi_a)} = \sum_q \frac{P_{bq*}(\theta)}{P_{q*}(\theta,\pi_a)} = Q\,. \tag{7}$$

The balance is broken (i.e. the optimal model contains only one latent class) when it is optimal to not include any amount of the other latent class, as discussed previously. A diagnostic condition for this is presented in (8) and (9) for the case including latent class a alone:

$$\left.\frac{\partial l(\theta,\pi_a)}{\partial \pi_a}\right|_{\pi_a=1} = \sum_q \frac{P_{aq*}(\theta)-P_{bq*}(\theta)}{\pi_a\, P_{aq*}(\theta)+(1-\pi_a)\, P_{bq*}(\theta)} > 0 \Rightarrow \pi_a^* = 1 \qquad (8)$$

$$\sum_q \frac{P_{aq*}(\theta)}{P_{bq*}(\theta)} < Q \Rightarrow \pi_a^* = 1 \qquad (9)$$

In this case of a single latent class a, $P_{a*}(\theta,\pi) \equiv P_{aq*}(\theta)$ so that $\sum_q \frac{P_{aq*}(\theta)}{P_{q*}(\theta,\pi)} = Q$. The result of Theorem 2 shows that this equality extends to each latent class in a binary model; this is generalised to multiple classes in Theorem 4.

The optimality of the single latent class a can be identified using (9) when $\pi_a^* = 1$ (or its counterpart for latent class b alone when $\pi_a^* = 0$). For this to occur, the prevalent latent class must perform well, even when individuals are better aligned to the other class. Conversely, for a balanced interior combination of classes to be optimal (10a) must hold for some observations and (10b) for others:

$$\frac{P_{aq*}(\theta)}{P_{bq*}(\theta)} > 1 \quad (10a) \qquad\qquad \frac{P_{bq*}(\theta)}{P_{aq*}(\theta)} > 1 \quad (10b)\,. \qquad (10)$$

Together, this means that in a balanced solution, each of the latent classes must perform best for some of the observations.

In the case that the class membership function π_a is not constant but is instead some function $\pi_a(\theta)$, the balance is stated in Theorem 3:

THEOREM 3: Two latent classes coexist optimally in a discrete choice model if the vector θ of estimated parameters satisfies the balance specified by (11):

$$\sum_q \frac{\frac{\partial \pi_a(\theta)}{\partial \theta} P_{aq*}(\theta) + \frac{\partial P_{aq*}(\theta)}{\partial \theta}\pi_a(\theta)}{P_{q*}(\theta)} = \sum_q \frac{\frac{\partial \pi_a(\theta)}{\partial \theta} P_{bq*}(\theta) - \frac{\partial P_{bq*}(\theta)}{\partial \theta}\left(1-\pi_a(\theta)\right)}{P_{q*}(\theta)} \qquad (11)$$

PROOF: Equation (12) states the stationarity condition required for optimality.

$$0 = \frac{\partial l(\theta)}{\partial \theta} = \sum_q \frac{\frac{\partial \pi_a(\theta)}{\partial \theta} P_{aq*}(\theta) + \pi_a(\theta)\frac{\partial P_{aq*}(\theta)}{\partial \theta} - \frac{\partial \pi_a(\theta)}{\partial \theta} P_{bq*}(\theta) + \left(1-\pi_a(\theta)\right)\frac{\partial P_{bq*}(\theta)}{\partial \theta}}{\pi_a(\theta) P_{aq*}(\theta) + \left(1-\pi_a(\theta)\right) P_{bq*}(\theta)} \qquad (12)$$

Equation (11) is a direct rearrangement of (12) that expresses stationarity in terms of the balance between the latent classes.

Suppose now that the set of parameters β of the class membership function is disjoint from the set θ affecting the choices themselves. Then Theorem 3 has the following corollary:

COROLLARY 3.1: If the class membership function, with parameters β, is independent from the choice heuristics, with parameters θ, the balance required of sensitivity of class membership is given by 13):

$$\sum_q \frac{\frac{\partial \pi_a(\beta)}{\partial \beta} P_{aq*}(\theta)}{P_{q*}(\theta,\beta)} = \sum_q \frac{\frac{\partial \pi_b(\beta)}{\partial \beta} P_{bq*}(\theta)}{P_{q*}(\theta,\beta)}\,. \qquad (13)$$

The analysis presented in this section identifies when it is optimal for the model to include more than one latent class. Nevertheless, the coexistence of latent classes does not guarantee that the model is identifiable. This empirical identifiability issue is addressed next.

2.1.2 Class behavioural diversity for empirical identifiability

To study the identifiability of a multiple latent classes, that is when the estimator can be identified uniquely without any parameter set being observationally equivalent, we assume that the model has an interior solution. If the model had a boundary solution (i.e. only one class was estimated), then the conclusion is that one heuristic consistently outperforms the other in explaining all population's behaviour.

For a parametric model to be identifiable, the information matrix (14) must be non-singular. Moreover, for greater precision in the parameter estimates, we desire that the covariance matrix $\boldsymbol{\Sigma}$ has values on the principal diagonal the square roots of which are small compared to the corresponding point estimates of parameters, which we describe as *strong identifiability*. The covariance matrix is related to the model via the Fisher information matrix F by (15):

$$F = -\mathbb{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta_x \partial\theta_y}\right) \tag{14}$$

$$\Sigma' = F^{-1} \tag{15}$$

The elements on the principal diagonal of Σ' provide the Cramér-Rao lower bound on the variance of estimation of the parameters θ in the corresponding elements of Σ. Thus, to obtain strong identifiability, the determinant of the information matrix should be large, hence requiring large values $-\mathbb{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta_x^2}\right)$ on its principal diagonal.

Similar to the analysis of the first order condition for the two-class case, we analyse the information matrix at the point determined by π_a. We analyse the case where the class membership function is constant. Thus, the diagonal element of the information matrix corresponding to π_a is given by the derivative of (4) with respect to π_a, and relates to the empirical identifiability of the class proportions:

$$\frac{\partial^2 l(\theta)}{\partial\pi_a^2} = -\sum_q \frac{\left(P_{aq*} - P_{bq*}\right)^2}{P_{q*}^2} \tag{16}$$

For F to have a large determinant, and thus for the standard errors of the estimators to be small, it is necessary for expression (16) to be large. Noting that the maximum likelihood estimates are obtained when the probability P_{q*}^2 is maximum, identifiability is determined by the numerator of (16). Thus, the expression $(P_{aq*} - P_{bq*})^2$ is an important element in the identification of latent classes. Large values of this expression are obtained when the classes exhibit disparate behaviour. We see from this that absence of substantial behavioural diversity between the two classes may cause identifiability problems. Therefore, this behavioural diversity requires not only that the functional forms are different but that it should also be reflected in the data used for estimation.

2.2 Multiple Latent Class Case

We now consider the general case in which several latent classes align with the population. We start by analysing the first order conditions to generalise the balance obtained in section 2.1. Then, the analysis of empirical identifiability is extended to this multiple class case.

Extending the notation of section 2.1, let π_c be the probability that individual behaviour aligns to class $c \in C$ so that $\sum_{c \in C} \pi_c = 1$ and $\pi_c \geq 0 \ \forall c \in C$. Then, the joint log-likelihood function $l(\pi, \theta)$ of the model is given by (19):

$$l(\pi, \theta) = \sum_q \log\left(\sum_{c \in C} \pi_c P_{cq*}(\theta)\right) \tag{19}$$

By extending Theorems 1 and 2, Theorem 4 indicates the necessary condition for the coexistence of several latent classes in a model:

THEOREM 4: Several latent classes $c \in C$ coexist optimally in a model when each of them contributes the same aggregated ratio $\sum_q \frac{P_{cq*}}{P_{q*}} = Q$.

PROOF: The constrained maximum likelihood subject to the sum constraint $\sum_{c \in C} \pi_c = 1$ (with Lagrange multiplier λ) and positivity constraints $\pi_c \geq 0 \ \forall c \in C$ (with Lagrange multipliers η_c $c \in C$) is obtained when the Lagrangian (20) is stationary with respect to $\pi_c \ \forall c \in C$:

$$\mathcal{L} = -l(\pi, \theta) - \lambda(1 - \textstyle\sum_{c \in C} \pi_c) - \sum_{c \in C} \eta_c \pi_c \, . \tag{20}$$

Differentiating the Lagrangian $\mathcal{L}$ with respect to π_c and equating to 0 for stationarity gives the necessary condition for the optimality with respect to the probability π_c:

$$\frac{\partial}{\partial \pi_c} \mathcal{L} = 0 \Leftrightarrow \sum_q \frac{P_{cq*}}{\sum_{a \in C} \pi_a P_{aq*}} = \lambda - \eta_c \ \Rightarrow \sum_q \frac{P_{cq*}}{P_{q*}} = \lambda - \eta_c \quad \forall c \in C \, .$$

The first-order Karush-Kuhn-Tucker (KKT) conditions for the positivity constraints on π_c with multiplier η_c are: $\pi_c \geq 0$, $\pi_c \eta_c = 0$, $\eta_c \geq 0$. According to the complementarity of π_c and η_c for each latent class c:

$$\pi_c > 0 \ \Rightarrow \ \eta_c = 0 \Rightarrow \ \sum_q \frac{P_{cq*}}{P_{q*}} = \lambda \, , \pi_c = 0 \ \Rightarrow \ \mu_c \geq 0 \Rightarrow \ \sum_q \frac{P_{cq*}}{P_{q*}} \leq \lambda \, . \tag{21}$$

Now applying the equation for P_{q*}, the stationarity condition for likelihood and the KKT conditions, we have

$$\begin{aligned} Q &= \sum_q \frac{\sum_{c \in C} \pi_c P_{cq*}}{\sum_{a \in C} \pi_a P_{aq*}} \\ &= \sum_{c \in C} \pi_c \sum_q \frac{P_{cq*}}{P_{q*}} \\ &= \lambda \sum_{c \in C} \pi_c - \sum_{c \in C} \pi_c \eta_c = \lambda \, . \end{aligned}$$

The sum constraint $\sum_{c \in C} \pi_c = 1$ yields the value λ for the first term in the last line, whilst the Karush-Kuhn-Tucker complementarity conditions $\pi_c \eta_c = 0 \ \forall c \in C$ yields the value 0 for the second term.

Using this in (21), $\pi_c > 0 \ \Rightarrow \ \sum_q \frac{P_{cq*}}{P_{q*}} = Q$. This proves Theorem 4 and extends the balance derived in section 2.1 to multiple latent classes: those present have identical aggregated ratio Q of P_{cq*}/P_{q*} for the alternatives chosen; others have values that are no greater than Q .

Theorem 4 presents the balance condition for the optimal point, but again does not guarantee the empirical identifiability of the latent classes. For the vector $\boldsymbol{\pi}$ to be identifiable, the information matrix should be non-singular and, therefore, the Hessian matrix of the Lagrangian should be positive definite. This requires that all principal submatrices of the Hessian that correspond to the second derivatives with respect to the proportions should have positive determinants. The mixed second partial derivatives of $\mathcal{L}$ are equal to those of the log-likelihood (because all the constraints are linear) and are stated in (22):

$$\frac{\partial^2}{\partial \pi_a \partial \pi_b}\mathcal{L} = \sum_q \frac{P_{aq*}P_{bq*}}{\left(\sum_{c\in C}\pi_c P_{cq*}\right)^2} = \sum_q \frac{P_{aq*}P_{bq*}}{P_{q*}^2} \tag{22}$$

Therefore, each 2×2 submatrix of this kind has the structure shown in (23):

$$\begin{bmatrix} \sum_q \frac{P_{aq*}^2}{P_{q*}^2} & \sum_q \frac{P_{aq*}P_{bq*}}{P_{q*}^2} \\ \sum_q \frac{P_{aq*}P_{bq*}}{P_{q*}^2} & \sum_q \frac{P_{bq*}^2}{P_{q*}^2} \end{bmatrix} \tag{23}$$

Because both elements on the principal diagonal are positive, the submatrix is positive definite if the determinant exceeds zero. Moreover, if the determinant *D* given by (24) is large, then the covariance matrix of the estimators is small:

$$D = \sum_{p\in Q} \frac{P_{ap*}^2}{P_{p*}^2} \sum_{q\in Q} \frac{P_{bq*}^2}{P_{q*}^2} - \left(\sum_{q\in Q} \frac{P_{aq*}P_{bq*}}{P_{q*}^2}\right)^2 \tag{24}$$

Before analysing (24) to determine when *D* will be positive, note that the analysis is useful in the case that latent classes are distinct. In that case, we cannot have $P_{aq*} = P_{bq*}\ \forall q$. Therefore, there will be a proportion of outcomes where class *a* outperforms the aggregate model and another proportion where its performance is worse. The quadratic structure of the expression $P_{cq*}^2 / P_{q*}^2 \ \ c \in C$ tends to amplify the difference when one model outperforms the other. Provided that each class outperforms every other class for some observations, then every determinant *D* of the form (24) will be positive and thus the model is theoretically identifiable. Empirical identifiability is addressed in Theorem 5.

THEOREM 5: If several latent classes coexist in an identifiable model, empirical identifiability increases as the value of the covariance of the latent classes decreases.

PROOF: To make the analysis more convenient, we introduce notation for the moments of the ratios of probabilities $\frac{P_{cq*}}{P_{q*}} \ \ c\epsilon C$. Let the first and second moments be respectively:

$$\mu_c = \mathbb{E}\left(\frac{P_{cq*}}{P_{q*}}\right), c \in C, \qquad \sigma_c^2 = Var\left(\frac{P_{cq*}}{P_{q*}}\right) \ \ c \in C \ \text{ and } \ \sigma_{ab} = Cov\left(\frac{P_{aq*}}{P_{q*}}, \frac{P_{bq*}}{P_{q*}}\right) \quad a, b \in C$$

With this notation, the expectation of elements involved in (24) can be written as:

$$\mathbb{E}\left(\sum_{q\in Q} \frac{P_{cq*}^2}{P_{q*}^2}\right) = Q(\mu_c^2 + \sigma_c^2) \ \text{ and } \ \mathbb{E}\left(\sum_{q\in Q} \frac{P_{aq*}P_{bq*}}{P_{q*}^2}\right) = Q(\mu_a\mu_b + \sigma_{ab})$$

Therefore, the expectation of (24) can be rearranged to express *D* as an unbiased sample estimate of the population quantity:

$$\frac{1}{Q^2}\mathbb{E}(D) = \mu_a^2\mu_b^2\left(\frac{\sigma_a^2}{\mu_a^2} - 2\frac{\sigma_{ab}}{\mu_a\mu_b} + \frac{\sigma_b^2}{\mu_b^2}\right) + \sigma_a^2\sigma_b^2\left(1 - \frac{\sigma_{ab}^2}{\sigma_a^2\sigma_b^2}\right). \tag{25}$$

Recall that from condition (21), for both classes a and b to be present in the model we need $\mu_a = \mu_b = 1$. If the choice probabilities are perfectly correlated, $\sigma_{ab}^2 = \sigma_a^2\sigma_b^2$: the right-hand side of (25) will be null so that the Hessian matrix would be singular in expectation. The expectation of the partial derivative of D with respect to the correlation σ_{ab} in (26) is negative, so that the expectation of the determinant increases as this correlation decreases. In particular,

$$\mathbb{E}\left(\frac{dD}{d\sigma_{ab}}\right) = -2Q^2(\mu_a\mu_b + \sigma_{ab}) = -2Q^2\mathbb{E}\left(\frac{P_{aq*}P_{bq*}}{P_{q*}^2}\right) \leq 0 \tag{26}$$

Consequently, estimation of the mixed model is better conditioned (as indicated by larger values of D) when correlation σ_{ab} is reduced and as sample size Q increases, proving Theorem 5.

The requirement for positive determinants of the principal submatrices of the Hessian, therefore, generalises the requirement for the binary classes case presented in section 2.1. To be identifiable, the behaviour of a class should outperform that of all other classes in at least one observation; the greater the behavioural difference, the bigger the determinant of (24) and hence the smaller the covariance matrix of the estimators.

3. Empirical Experiments for Identifiability

To test the theorems in an identifiable case that is not straightforward, we use the context of multiple-choice heuristics. Here, each choice heuristic is modelled under a different latent class. To build the experiment and to guarantee the presence of different choice heuristic and control the choice parameters, a synthetic population is generated. We investigate three dimensions affecting the choice process: the type of choice heuristic within each latent class, the proportion of each latent class (or choice heuristic) in the synthetic sample, and the correlation between the parameters of the probability of belonging to each class, and the parameters associated with their sensitivities for different attributes of the alternatives. Finally, for each of these three dimensions, ten experiments were performed.

The first dimension is the type of choice heuristic. The analysis of section 2 establishes that the difference between the latent classes is key to their identification. Three different choice heuristics were tested against RUM, which is used most widely, to investigate whether they could be identified in a practical context. These are: Elimination by Aspects –EBA– (Tversky, 1972), Stochastic Satisficing –SS– (González-Valdés and Ortúzar, 2018) and Random Regret Minimization –RRM– (Chorus et al., 2008).

The second dimension is the proportion of each latent class (or choice heuristic) in the sample. The results (5) and (7) show that the greater this proportion, the greater the number of observations for which it will outperform other heuristics, thus increasing its identifiability. Two proportions were tested: 70% of the sample chooses according to RUM and 30% according to the other heuristic, and vice versa, i.e. $\pi_c \in \{0.3, 0.7\}$.

Finally, the third dimension is the correlation between the choice and the class membership probabilities. The purpose of this dimension was to analyse how any such correlation would impact on identifiability. Correlation was introduced in a personal trait that affects both the probability of belonging to a class and the choice preferences.

We use a simulated dataset to investigate whether it was possible to capture a mixture of choice heuristics in a practical transport context. For estimation we required two components: a set of choice alternatives available to each individual and each individual's choice from their set. The choice sets for each individual were extracted from a real revealed preference dataset to represent a realistic scenario; the individuals' choices were simulated for the synthetic population under the various heuristics, to control the generating behaviours.

3.1 The Choice Sets

The choice sets were created based on a dataset from a transport survey in Santiago de Chile (Gaudry et al., 1989; Guevara, 2016), considering trips from home to work of 1,374 individuals, who chose among nine modes. Because we wanted to control the number of alternatives available in the experiment, the simulation choice sets were constrained to three alternatives. Moreover, as these were labelled, alternative-specific constants (ASC) could be estimated.

To create the simulated choice sets, two processes were performed separately: (i) fictitious choice sets of size 3 were created and (ii) each individual's choice was simulated from one of these sets. In the first step, real choice sets were sampled from the databank and then adjusted as follows. If the sampled choice set had fewer than three alternatives, it was discarded; if it had more than three alternatives, alternatives were deleted at random until the desired choice set size was obtained. Accounting for all our different choice sets, we had a total of 28,477 different choice sets to pool from. We repeated this procedure of uniform random sampling with replacement from the 1,374 individuals to generate a synthetic sample of 10,000 choices.

After the choice sets were built, the individual's choice was simulated under the specified heuristic. Each alternative in the choice sets was characterised by four attributes: monetary cost, in-vehicle time, walking time, and waiting time.

3.2 Synthetic Population and Choice Heuristics

To obtain a chosen alternative from the simulated choice sets, the following approach was used. For each individual in the sample, a binary variable was first generated to represent a socio-demographic attribute z (simply named *trait*) with probability p_z. Each simulated individual was also assigned independently to use one of two available choice heuristics: RUM and the contrasting one (i.e. EBA, RRM or SS). In each case, the probability π_R of using RUM was given by the inverse logit function (27).

$$\pi_R = \frac{\exp(\theta_0 + \theta_1 z)}{1 + \exp(\theta_0 + \theta_1 z)} \tag{27}$$

The choice heuristics used were: (i) RUM in its simplest form, the multinomial logit (MNL) model (McFadden, 1973), with linear and additive in parameters utility function. In some experiments the cost attribute was modified based on the individual's sociodemographic trait, to analyse the correlation between the class membership function and the parameters of the heuristics; (ii) *Random regret minimisation* (RRM, Chorus et al., 2008), selecting the μ-RRM formulation of RRM to increase the profundity of regret compared to the simplest version to increase the behavioural difference with RUM, and thus, the probability of identifying them jointly; (iii) *Satisficing* (Simon 1955), interpreted here as a heuristic according to which an individual chooses the first satisfactory (i.e. good enough) alternative they find, using the Stochastic Satisficing (SS) model of Gonzalez-Valdes and Ortúzar (2018); and (iv) *Elimination by Aspects* (EBA, Tversky, 1972), where every aspect is discrete, although the attributes may have continuous values, as in the present case. Acceptability thresholds are specified to achieve binary discrimination.

Because our focus was to work with models that exhibit identifiability issues, we preferred Bayesian estimation over maximum likelihood estimation, as the latter is prone to be captured in local optima, although Train (2009, p290) showed that this has no impact when the sample size Q is large (as in our case). Bayesian estimation by Markov Chain Monte Carlo, specifically Gibbs sampling, used the JAGS package (Plummer, 2016) for the R software system (R Core Team, 2016). For each parameter, ten thousand useful samples were obtained after burn-in. For the prior distributions, low precision zero-centred normal priors were used.

4. ANALYSIS OF RESULTS

Given the dimensions tested and the replications for each combination, a total of 120 experiments were undertaken. First, we analysed – across the various dimensions – the proportion of replications of each model that resulted in a balance between the choice heuristics (latent classes). Then, we verified that Theorem 2 held for the models that identified both latent classes notwithstanding being estimated using Bayesian methods. Then, for each of the three cases (one for each combination of latent classes), the average of the parameter estimates was tested against the corresponding target ones.

4.1 Analysis of Identifiability

A model is non-identifiable if the information matrix is singular. In our context of Bayesian estimation, no matrix inversion is required; nonetheless, model non-identifiability can be detected when the standard deviations of the parameters are extreme with associated instability of the Markov chain. Even though we have described identifiability, we detected different degrees of non-identifiability. Therefore, to distinguish degrees of identifiability that models may exhibit, we developed three further descriptions:

- *Strong identifiability*: all parameters of the model are estimated with acceptable standard deviations. Both latent classes are identified, thus there is a balance between them.
- *Weak identifiability*: most of the parameters of the model are estimated accurately but a small proportion of them are estimated with extreme standard deviations. Nevertheless, the model is able to identify the two classes clearly.
- *Non-identifiability*: either most parameters are estimated with extreme standard deviation, or no balance can be found between latent classes.

In section 2 we analysed how behavioural differences may impact identifiability of the latent classes. Figure 1 provides a graphical diagnostic that shows the distribution of behavioural differences between the RUM class and the other choice heuristic classes among the alternatives of the dataset. This is quantified by the absolute difference between the probabilities given by two choice heuristics. For example, if two heuristics a and b estimate probabilities P_{ai} (respectively) P_{bi} of choosing alternative i , then the difference is calculated as $|P_{ai} - P_{bi}|$.

Figure 1 shows that among the choice heuristics tested, the RRM latent class differs least from the RUM latent class. Thus, we expect the RRM latent class to be the one with the least chance of balance with RUM in this context. Conversely, each of the SS and EBA latent classes present a substantial behavioural difference from RUM. Note however, that because this analyses only one dimension of the information matrix, it is useful for generating hypotheses but does not guarantee support for them. We analysed each pair of latent classes separately and evaluated the results according to the three degrees of identifiability.

We also analysed separately the importance for this identifiability of the proportions of latent classes and each of the correlation cases. Due to lack of space, we cannot provide further results here, but these are available on request from the authors.

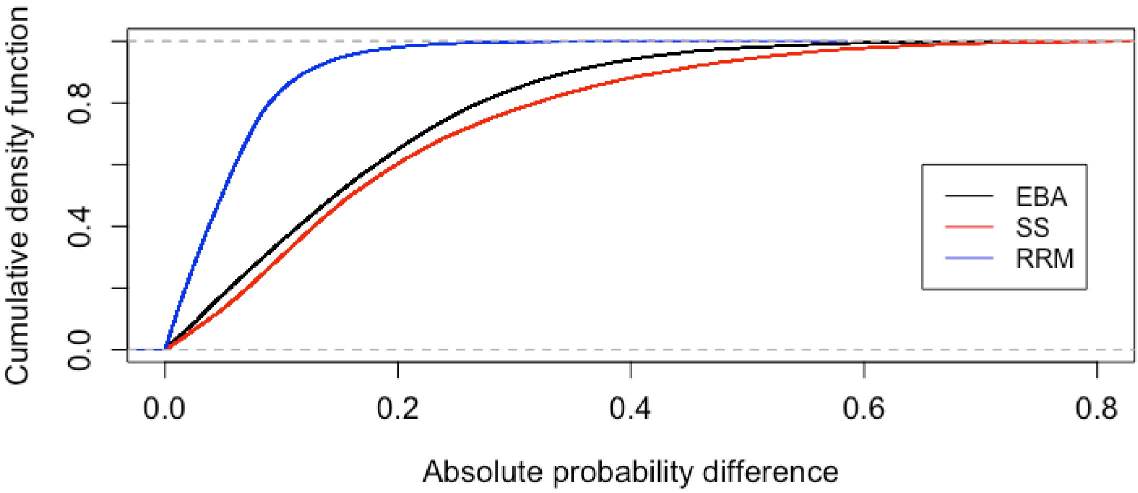

Figure 1: Behavioural difference between RUM and each of RRM, SS and EBA

5. CONCLUSIONS

The theoretical framework developed here of LC models provides a basis for analysis of their identifiability. Through this, we established two analytical conditions for this: first, a balance must exist between the latent classes and second, the behaviour of the classes must differ sufficiently that they can be identified with acceptable accuracy in their parameters. The balance that is required for joint estimation is quantified in terms of the size of sample used.

Our experiments show that estimation may fail to identify a pair of classes in a synthetic sample, even if the generating process contains a mixture of them. The existence of a balance depends on the performance of each class when interpreting the behaviour of the other. Indeed, the dominant class must perform poorly on some choices that were made following the other heuristic for the LC model to be able to estimate both of them.

In the practical experiments, we investigated different pairs of choice heuristics representing the latent classes. The link between the theoretical and empirical approaches was a graphical analysis that presents the difference between classes. In the present mode choice context, Random Regret Minimization was found not to differ much from Random Utility Maximisation. However, important behavioural differences from Random Utility Maximisation were exhibited by each of Stochastic Satisficing and Elimination by Aspects. Therefore, a worthwhile strategy could be to analyse the classes before estimating a combined model, which can be undertaken using straightforward diagnostic tests presented here. This way, with some testing parameters, modellers can examine whether the datasets are sufficiently rich in their choice behaviour to estimate the desired heuristics.

Our experiments show that, in principle, it is possible to estimate sophisticated class membership functions and latent classes simultaneously even when the difference in behaviour is given only by the choice heuristics. In fact, the model remains identifiable even if some variables affect both the class membership and choice levels of the model.

ACKNOWLEDGEMENTS

We are grateful to the Instituto Sistemas Complejos de Ingeniería (CONICYT PIA/BASAL AFB180003) and the BRT+ Centre of Excellence (www.brt.cl) for having funded this research.

REFERENCES

Balbontin, C., Hensher, D.A. & Collins, A.T. (2017) Integrating attribute non-attendance and value learning with risk attitudes and perceptual conditioning. *Transportation Research Part E: Logistics and Transportation Review* **97**, 172–91.

Ben-Akiva, M. & Boccara, B. (1995) Discrete choice models with latent choice sets. *International Journal of Research in Marketing* **12**, 9–24.

Cherchi, E. & Ortúzar, J. de D. (2008) Empirical identification in the mixed logit model: analysing the effect of data richness. *Networks and Spatial Economics* **8**, 109-24.

Chorus, C.G. (2010) A new model of random regret minimization. *European Journal of Transport and Infrastructure Research* **10**, 181–96.

Chorus, C.G., Arentze, T.A. & Timmermans, H.J.P. (2008) A random regret-minimization model of travel choice. *Transportation Research Part B: Methodological* **42**, 1–18.

Gaudry, M.J.I., Jara-Diaz, S.R. & Ortúzar, J. de D. (1989) Value of time sensitivity to model specification. *Transportation Research Part B: Methodological* **23**, 151–8.

González-Valdés, F. & Ortúzar, J. de D. (2018) The stochastic satisficing model: a bounded rationality discrete choice model. *Journal of Choice Modelling* **27**, 74–87.

Gonzalez-Valdes, F. & Raveau, S. (2018) Identifying the presence of heterogeneous discrete choice heuristics at an individual level. *Journal of Choice Modelling* **28**, 28–40.

Guevara, C.A. (2016) Mode-valued differences of in-vehicle travel time savings. *Transportation* **44**, 977–97.

Hess, S., Stathopoulos, A. & Daly, A. (2012) Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. *Transportation* **39**, 565–91.

Huang, G.H. & Bandeen-Roche, K. (2004) Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* **69**, 5–32.

Leong, W. & Hensher, D.A. (2012) Embedding multiple heuristics into choice models: an exploratory analysis. *Journal of Choice Modelling* **5**, 131–44.

McFadden, D. (1973) Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers of Econometrics*. Academic Press, New York.

Plummer, M. (2016) RJags: Bayesian graphical models using MCMC. R package 4-6.

R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rossetti, T., Guevara, C.A., Galilea, P. & Hurtubia, R. (2018) Modelling safety as a perceptual latent variable to assess cycling infrastructure. *Transportation Research Part A: Policy and Practice* **111**, 252–65.

Simon, H.A. (1955) A behavioural model of rational choice. *The Quarterly Journal of Economics* **69**, 99–118.

Train, K.E. (2009) *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

Tversky, A. (1972) Elimination by aspects: a theory of choice. *Psychol. Review* **79**, 281–99.

Walker, J. & Ben-Akiva, M. (2002) Generalized random utility model. *Mathematical Social Sciences* **43**, 303–43.