On the overspecification of multinomial and nested logit models due to alternative specific constants

M. Bierlaire[1], T. Lotan[2] and Ph. Toint[3]

Report 95/12 October 24, 1995

Abstract.
Alternative specific constants (ASCs), although playing an important role in random utility models, have received very little attention in theoretical studies. In this paper, we try to fill this gap by providing an analysis of the overspecification caused by ASCs to the log-likelihood function of multinomial and nested logit models. The theoretical results lead directly to a particular strategy of ASC specification, called here the *orthogonal strategy*. The analysis of the relationship between any two arbitrary strategies and the derivation of an interesting geometrical property of the orthogonal strategy provide a motivation to prefer the latter.

[1] Transportation Research Group — Department of Mathematics,
Facultés Universitaires ND de la Paix, Namur, Belgium
E-mail: mb@math.fundp.ac.be

[2] Service de Mathématique de la Gestion
Université Libre de Bruxelles, Brussels, Belgium
E-mail: tlotan@ulb.ac.be

[3] Transportation Research Group — Department of Mathematics,
Facultés Universitaires ND de la Paix, Namur, Belgium
E-mail: pht@math.fundp.ac.be

Current reports available by anonymous ftp from the directory
"pub/reports" on math.fundp.ac.be (internet 138.48.4.14)

1 Introduction

Alternative specific constants (ASCs) play an important role in multinomial and nested logit models: they correspond to the expectation of the random error term, and thus represent preferences, which are inherent and independent of specific attribute values, towards the alternatives. Furthermore, an incorrect choice of ASCs can cause model overspecification. Despite their importance, a detailed analysis of the role of ASCs is not available in the literature, at least to the authors' knowledge. Amongst the published references, Ben-Akiva and Lerman (1985) suggest avoiding overspecification of multinomial models by defining only J-1 ASCs for a choice set of size J. A common practice in this case is thus to assume that one (arbitrary) ASC is equal to zero. Simple generalizations of this principle to nested-logit models typically result in either equating one ASC to zero and estimating the rest (Daganzo and Kusnic, 1993), or in applying the same principle to all levels of the hierarchy (Ben Akiva M., private communication, March 1995).

In this paper, the role of ASCs is analyzed in detail from a complete description of the geometry of the overspecified log-likelihood function. A characterization of the overspecification due to the ASCs is fully developed, through the definition of the *invariant subspace*, which includes all the directions producing no changes to the log-likelihood. This leads to the identification of the subspace containing all the relevant geometrical information, that is the subspace orthogonal to the invariant subspace. After showing that an infinite number of different strategies can be used to overcome the overspecification problem, it is proved that any two strategies are equivalent up to a non-singular linear transformation of the variables of the optimization problem.

In addition to the presentation of commonly known strategies to remove overspecification, a new strategy, called here the *orthogonal strategy*, is described. It is deduced directly from the theoretical results presented in this paper and exhibits an interesting and desirable geometrical property.

The paper is organized as follows. Section 2 introduces the necessary notation and definitions. Section 3 develops the geometrical characteristics of the overspecified log-likelihood function, while Section 4 describes some specification strategies, analyzes their relationship and establishes the optimal geometric nature of the orthogonal strategy. Before concluding in Section 6, important algorithmic implications and practical comments are provided in Section 5.

2 Notation and definitions

The structure of a nested logit model is commonly represented by a directed tree $\mathcal{T}$, the vertices of which feature alternatives of the choice process (see for example, Ben-Akiva and Lerman, 1985, or Daly, 1987). The vertices of a tree are called *nodes*. The predecessor of a node k is called its *father* and is denoted by $\tilde{k}$. Sets of nodes are denoted by upper case calligraphic letters different from $\mathcal{L}$. For the sake of clarity, most of the definitions are using genealogical vocabulary, associating the choice tree with a family tree. The set

$$\mathcal{C}(k) = \{\ell \in \mathcal{T} \mid \tilde{\ell} = k\} \tag{1}$$

is called the set of the *children* of node k. The set of nodes in $\mathcal{T}$ is partitioned into three subsets: the root, the elemental nodes and the structural nodes. The root is the only node, noted r, that has no predecessor. The set of *elemental nodes* $\mathcal{E}$ is defined as the set of nodes without children. The set of *structural nodes* $\mathcal{S}$ is defined by

$$\mathcal{S} = \{\ell \in \mathcal{T}, \ell \neq r \mid \mathcal{C}(\ell) \neq \emptyset\}. \tag{2}$$

Typically, each structural node $k \in \mathcal{S}$ is associated with a non-zero *structural coefficient* $\theta_k \in I\!R$. The model is said to be *multinomial* if all nodes except the root are elementary.

In a logit model, the systematic part of the utility function associated by a given individual with an elemental node k has the following form:

$$U_k = c_k + \sum_i \beta_i x_i(k), \tag{3}$$

where $k \in \mathcal{E}$, $x_i(k)$ is the ith characteristic experienced by the individual for alternative k, β_i is the (unknown) coefficient representing the relative importance of characteristic i and c_k is the *alternative specific constant* (ASC) associated with this alternative. The ASC represents the expectation of the stochastic part of the utility function. Incorporating it in the deterministic part allows the assumption of zero mean to be always valid.

The utility function associated with a structural node k for a given individual has the form:

$$U_k = c_k + \sum_i \beta_i x_i(k) + \theta_k \sum_{\ell \in \mathcal{C}(k)} e^{U_\ell/\theta_k}, \tag{4}$$

where $k \in \mathcal{S}$, the first two terms of the right-hand side correspond to the part of the utility function specific to node k, and θ_k is the (unknown) structural coefficient associated with node k. We note that an ASC is associated with every node of the tree but the root. It is emphasized that, contrary to Daganzo and Kusnic (1993), specification of utility functions (and of ASCs) for structural nodes is enabled here. We also note that another form of the utility function (4) is sometimes used (see Daly, 1987, for instance), whose impact on the developments presented below is discussed in Section 5.

The choice model is defined by the specification of the probability $\mathrm{Prob}(k)$ that the given individual chooses alternative k. This probability can be specified recursively as follows.

- If $\tilde{k} = r$ (that is if k is a child of the root), then

$$\mathrm{Prob}(k) = e^{U_k} / \sum_{\ell \in \mathcal{C}(r)} e^{U_\ell}, \tag{5}$$

- otherwise,

$$\text{Prob}(k) = \text{Prob}(\tilde{k})\, e^{U_k} / \sum_{\ell \in \mathcal{C}(\tilde{k})} e^{U_\ell}. \tag{6}$$

Note that only (5) applies to multinomial models.

The log-likelihood function corresponding to an arbitrary sample can then be defined as

$$\mathcal{L}(\beta, \theta, c) = \sum_{j \in \{\text{individuals}\}} \log \text{Prob}(k_j), \tag{7}$$

where k_j is the alternative actually chosen by individual j in the sample, β is the vector of q coefficients β_i, θ the vector of $|\mathcal{S}|$ coefficients θ_ℓ, and c is the vector of $n = |\mathcal{T}| - 1$ ASCs. The coefficients are then estimated by the maximum likelihood method, that is by solving the problem

$$\begin{aligned} &\max_{\beta,\theta,c} \mathcal{L}(\beta, \theta, c), \\ &\beta \in I\!R^q, \;\; \theta \in I\!R^{|\mathcal{S}|}, \;\; c \in I\!R^n. \end{aligned} \tag{8}$$

3 Geometrical aspects of overspecification due to ASCs

It is well-known that a model including all n ASCs produces an overspecified log-likelihood function (see Ben-Akiva and Lerman, 1985). Such a situation must be avoided because it prevents many estimation procedures from finding an optimal solution to problem (8). While some of the more robust methods nevertheless succeed in converging (see, for example, Bierlaire, 1995), their speed of convergence can be very slow. This problem is due to the singularity of the second derivatives matrix (or Hessian matrix) of the log-likelihood function. Moreover, the usual estimation of the variance-covariance matrix, which involves the inverse of this Hessian, cannot be computed in this case. In this paper, we will concentrate on the overspecification arising from the ASCs and will not consider other possible sources depending on the data and model structure.

The overspecification in the multinomial case arises because adding the same quantity to the utilities of all alternatives does not affect the log-likelihood of the sample. This can be characterized by Theorem 1, where the vector space $I\!R^n$ is considered here with its orthonormal canonical basis $(e_1, \ldots, e_n)$, with $e_j \in I\!R^n$ being a vector composed only of zeros, except for the j-th component which is one.

Theorem 1 *Let $\mathcal{L}(\beta, c)$ be the log-likelihood function of an arbitrary sample for a multinomial model with n alternatives and q coefficients. Then*

$$\mathcal{L}(\beta, c) = \mathcal{L}(\beta, c + z) \quad \forall z \in \text{span}\{e\} \tag{9}$$

where $e = \sum_{j=1}^{n} e_j$ is the vector of all ones.

The proof is directly obtained from the form of the logit model in (5). We then say that $M = \text{span}\{e\}$ is an *invariant subspace* associated with $\mathcal{L}$.

Characterizing the overspecification in the nested case is slightly more complex. Indeed, the source of overspecification is at least twofold. First, the root and its children, considered as a multinomial submodel, may cause overspecification, as characterized by the next theorem.

Theorem 2 *Let $\mathcal{L}(\beta, \theta, c)$ be the log-likelihood function of an arbitrary sample for a nested model with n nodes, $|\mathcal{S}|$ structural nodes and q coefficients. Then*

$$\mathcal{L}(\beta, \theta, c) = \mathcal{L}(\beta, \theta, c + z) \quad \forall z \in span\{v_r\} \tag{10}$$

with

$$v_r = \sum_{k \in \mathcal{C}(r)} e_k \in \mathbb{R}^n. \tag{11}$$

The proof is directly obtained from (5). The vector v_r is composed of zeros, except for entries corresponding to the children of the root which are ones, that is

$$k \notin \mathcal{C}(r) \Longleftrightarrow \langle e_k, v_r \rangle = 0, \text{ and } k \in \mathcal{C}(r) \Longleftrightarrow \langle e_k, v_r \rangle = 1, \tag{12}$$

where $\langle \cdot, \cdot \rangle$ denotes the classical Euclidean inner product in $\mathbb{R}^n$.

The second source of overspecification is related to definition (4) of the utility of structural nodes. Indeed, if a constant is added to the utility of each child of a structural node k and substracted from the utility of k, then the global utility of k is unchanged, and so is the log-likelihood function.

Theorem 3 *Let $\mathcal{L}(\beta, \theta, c)$ be the log-likelihood function of an arbitrary sample for a nested model with n nodes, $|\mathcal{S}|$ structural nodes and q coefficients. Then, for each $k \in \mathcal{S}$, we have that*

$$\mathcal{L}(\beta, \theta, c) = \mathcal{L}(\beta, \theta, c + z) \quad \forall z \in span\{v_k\}, \tag{13}$$

where

$$v_k = \sum_{\ell \in \mathcal{C}(k)} e_\ell - e_k. \tag{14}$$

The proof is directly obtained from (4).

We next define the subspace M of $\mathbb{R}^n$, which is included in the invariant subspace.

Theorem 4 *The family of $s + 1$ vectors $\{v_r, \{v_k\}_{k \in \mathcal{S}}\}$ is a basis of the subspace*

$$M = span\{v_r\} \oplus \left[\bigoplus_{k \in \mathcal{S}} span\{v_k\} \right], \tag{15}$$

where the notation $\oplus$ denotes the direct sum of subspaces.

Proof. The desired result immediately follows from the definition of M and from (easy-to-prove) linear independancy of the family $\{v_r, \{v_k\}_{k \in \mathcal{S}}\}$. $\square$

We next prove that M completely characterizes the overall invariant subspace, that is it accounts for all the overspecification of the model due to the ASCs.

Theorem 5 *The subspace M is the overall invariant subspace for the likelihood function associated with the nested logit model for any given values of β and θ. Therefore, the dimension of this invariant subspace is $|\mathcal{S}| + 1$.*

Proof. We deduce from Daganzo and Kusnic (1992) that the log-likelihood function is strictly concave if β and θ are fixed. This result depends on the assumption made by these authors that no utility is associated with the $|\mathcal{S}|$ structural nodes and an arbitrarily selected elementary node, say the first. However, this causes no loss of generality compared with our formulation: it reduces to assuming, in this context, that the ASCs associated with structural nodes are zero and that the remaining part of their utility as well as the utility of the first node are transferred to elemental nodes other than the first. As a result, all degrees of freedom in specifying the ASCs are removed by fixing $|\mathcal{S}|+1$ of them, which in turn implies that the invariant subspace is of dimension at most $|\mathcal{S}|+1$. But this subspace is independent of the particular technique used for specificying the ASCs and the conclusion on its dimension is thus independent of the framework used by Daganzo and Kusnic. Combining this result with that of Theorem 4, we now deduce that the dimension of the invariant subspace is exactly $|\mathcal{S}|+1$, which finally implies that it is identical to M and yields the desired conclusion. □

Consequently, any nonzero vector z such that, for any β and θ,

$$\mathcal{L}(\beta,\theta,c)=\mathcal{L}(\beta,\theta,c+z), \tag{16}$$

belongs to M.

Consider for example a nested model with two structural nodes based on the tree structure shown in Figure 1. The associated invariant subspace M is then spanned by

$$\begin{array}{lcl} v_r & = & (\ 0\ ,\ 0\ ,\ 0\ ,\ 0\ ,\ 1\ ,\ 1\ )^T \\ v_3 & = & (\ 1\ ,\ 1\ ,\ -1\ ,\ 0\ ,\ 0\ ,\ 0\ )^T \\ v_5 & = & (\ 0\ ,\ 0\ ,\ 1\ ,\ 1\ ,\ -1\ ,\ 0\ )^T \end{array} \tag{17}$$

where the superscript ^{T} denotes the transpose.

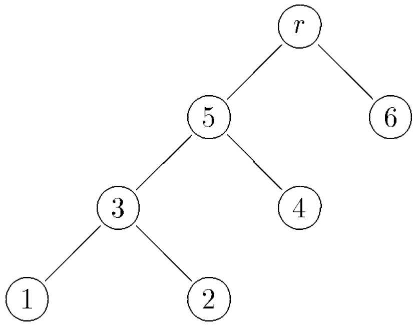

Figure 1: A tree structure

The previous results have shown that no information on changes in the log-likelihood function can be obtained in the invariant subspace M. However, since this function is typically not constant, there must be another part of $\mathbb{R}^n$ where the desired information, that is variations in function values and derivatives, can be found. The next theorem indicates that all this important information lies in the orthogonal complement of M in $\mathbb{R}^n$, defined by

$$C_\perp=\{w\in\mathbb{R}^n \mid \langle w,z\rangle=0 \text{ for all } z\in M\}. \tag{18}$$

Theorem 6 *Consider the gradient* $\nabla\mathcal{L}(\beta,\theta,c)$ *and the Hessian* $\nabla^2\mathcal{L}(\beta,\theta,c)$ *of* $\mathcal{L}$ *evaluated at any* $(\beta,\theta,c) \in I\!R^{q+|\mathcal{S}|+n}$. *Let also* $C_\perp$ *be the orthogonal complement of* M *in* $I\!R^n$, $P_\perp$ *be the orthogonal projector onto that subspace and*

$$\hat{P}_\perp = \begin{pmatrix} I & 0 \\ 0 & P_\perp \end{pmatrix} \tag{19}$$

be the orthogonal projector onto $I\!R^{q+|\mathcal{S}|} \times C_\perp$, *where* I *is the identity matrix of dimension* $q + |\mathcal{S}|$. *Then*

$$\nabla\mathcal{L}(\beta,\theta,c) = \hat{P}_\perp \nabla\mathcal{L}(\beta,\theta,c), \text{ and } \nabla^2\mathcal{L}(\beta,\theta,c) = \hat{P}_\perp \nabla^2\mathcal{L}(\beta,\theta,c)\hat{P}_\perp. \tag{20}$$

Proof. By definition of the invariant subspace and the orthogonality of $P_\perp$, we have that

$$\mathcal{L}(\beta,\theta,c) = \mathcal{L}(\beta,\theta,P_\perp c) = \mathcal{L}(\hat{P}_\perp(\beta,\theta,c)), \tag{21}$$

The desired result then immediately follows from chain-rule differentiation. □

Note that the proof of the theorem implies that the same result applies not only to the gradient and Hessian of the log-likelihood functions, but also to all higher order derivatives.

After examining the behaviour of the overspecified log-likelihood function and its derivatives, we are now in position to specify strategies whose purpose is to overcome this overspecification.

4 Strategies for resolving the overspecification

In order to avoid the underdeterminacy in c due to the existence of the invariant subspace M, it is necessary to constrain the log-likelihood's maximization to a subspace C complementary to M, that is for which

$$M \oplus C = I\!R^n. \tag{22}$$

We say that C is the *constraint subspace*, whose dimension is $\dim(C) = n - (|\mathcal{S}| + 1) = |\mathcal{E}| - 1$. We note here that there are infinitely many such subspaces, and that we only need to select one to avoid overspecification. The likelihood maximization (8) then reduces to

$$\hat{\Pi}(C) \stackrel{\text{def}}{=} \begin{cases} \max\limits_{\beta,\theta,c} \mathcal{L}(\beta,\theta,c) \\ \beta \in I\!R^q, \quad \theta \in I\!R^{|\mathcal{S}|}, \quad c \in C, \end{cases}$$

whose solution is a particular solution of (8). Defining R to be any full-rank matrix whose columns span C, the constrained problem $\hat{\Pi}(C)$ can be further reformulated as an unconstrained problem given by

$$\Pi(C,R) = \begin{cases} \max\limits_{\beta,\theta,c} \mathcal{L}(\beta,\theta,Rz) \\ \beta \in I\!R^q, \quad \theta \in I\!R^{|\mathcal{S}|}, \quad z \in I\!R^{|\mathcal{E}|-1}. \end{cases} \tag{23}$$

Since Theorem 6 indicates that all the relevant function and derivative information is contained in $C_\perp$, it is natural to choose the constraint subspace C to be identical to $C_\perp$. We call this particular ASC specification strategy the *orthogonal strategy*.

We now characterize $C_\perp$ from the definition of M. Using Hesse's normal form, the equation of a plane is directly deduced from the components of one of its normal vectors (see, for example, Kreyszig, 1972, p. 204): the hyperplane of all c orthogonal to an arbitrary vector v is defined by the equation

$$\sum_{j=1}^{n} \langle v, e_j \rangle c_j = 0. \tag{24}$$

We thus obtain that the hyperplane orthogonal to v_r defined by (11) is characterized by the equation

$$\sum_{j \in \mathcal{C}(r)} c_j = 0. \tag{25}$$

Similarly, the hyperplane orthogonal to v_k defined by (14) is characterized by

$$\sum_{j \in \mathcal{C}(k)} c_j - c_k = 0. \tag{26}$$

The constraint subspace $C_\perp$ is then the intersection of the hyperplanes defined by (25) and (26). Refering to the tree shown in Figure 1, constraints (25) and (26) can be written as

$$\left\{ \begin{array}{rl} c_5 + c_6 & = 0, \\ c_3 + c_4 - c_5 & = 0, \\ c_1 + c_2 - c_3 & = 0, \end{array} \right. \quad \text{or} \quad \left\{ \begin{array}{l} c_6 = -c_5, \\ c_4 = -c_3 + c_5, \\ c_2 = -c_1 + c_3, \end{array} \right. \quad \text{that is,} \quad R = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{pmatrix}. \tag{27}$$

As far as the authors are aware, the orthogonal strategy is not used in practice. We next describe several alternative techniques for choosing C that are applied in the field, both for the multinomial and the nested case. The technique most widely used for the multinomial case has been described namely by Ben-Akiva and Lerman (1985) and consists of selecting an arbitrary node b for which the ASC is constrained to be zero. The corresponding constraint subspace is then given by

$$C_m = \text{span}\{e_1, \ldots, e_{b-1}, e_{b+1}, \ldots, e_n\} = \bigoplus_{\ell \neq b} \text{span}\{e_\ell\}. \tag{28}$$

Note here that the n possible choices for b yield n different C_m. This technique is appealing because both the interpretation of the estimated ASCs and its practical implementation are straightforward.

Two methods seem to be used in practice for ASC specification in the case of nested models. The first is inherited from the technique we just described for the multinomial case and appears, for instance, in Daganzo and Kusnic (1992): an elemental node b is chosen arbitrarily and its ASC is constrained to be zero, as are the ASCs associated with structural nodes. It is called here the *elemental strategy*, because the hierarchical structure of the model is not taken into account.

Choosing an arbitrary $b \in \mathcal{E}$, the constraint subspace C_e is defined as

$$C_e = \bigoplus_{\substack{\ell \in \mathcal{E} \\ \ell \neq b}} \text{span}\{e_\ell\}. \tag{29}$$

This specification is very similar to (28). Applying this method to the example described by Figure 1, and arbitrarily choosing $b = 6$, we have

$$C_e = \text{range}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{30}$$

The second strategy also draws its inspiration from the multinomial practice. However, in contrast to the previous one, the structure of the tree is now explicitly taken into account. It is thefore called here the *structural strategy*. For the root r and each of the $|\mathcal{S}|$ structural nodes k, we select one and only one arbitrary child, denoted b_r and b_k, respectively, whose associated constant is constrained to be zero. We also define

$$\mathcal{B} \stackrel{\text{def}}{=} \{b_r\} \cup \{b_k \mid k \in \mathcal{S}\} \tag{31}$$

to be the set of all such nodes. The constraint subspace C_h is then given by

$$C_h = \bigoplus_{k \in \mathcal{T}\setminus\mathcal{B}} \text{span}\{e_k\}. \tag{32}$$

Considering the tree shown in Figure 1 and selecting $b_3 = 2$, $b_5 = 3$ and $b_r = 6$, we have that

$$C_h = \text{range}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{33}$$

After specifying different strategies for ASC specification, we now devote the rest of this section to compare them. Recalling that the log-likelihood function remains invariant along directions lying in M, we denote by

$$\overline{M}(c) = \{c + z \mid z \in M\} \tag{34}$$

the translated invariant subspace of the log-likelihood function at any point of the form (β, θ, c). We then define the (generally non-orthogonal) projection onto a constraint subspace C along M by

$$F_C c = \overline{M}(c) \cap C, \tag{35}$$

for any $c \in I\!R^n$ and any choice of the subspace C, where the intersection of the right-hand side is unique because of (22). This projection has the important property that it does not alter the value of the log-likelihood function, in the sense that

$$\mathcal{L}(\beta, \theta, c) = \mathcal{L}(\beta, \theta, F_C c) \tag{36}$$

for all choices of C, β, θ and all $c \in \mathbb{R}^n$. Because $\overline{M}(c) = \overline{M}(F_C c)$ holds by definition for all c and for each constraint subspace C, we deduce the important property that, for any constraint subspaces C_1 and C_2, the restriction of F_{C_2} to C_1 is bijective from C_1 to C_2 and

$$F_{C_1} F_{C_2} c_1 = c_1, \tag{37}$$

for each $c_1 \in C_1$. Note also that $F_{C_\perp} = P_\perp$, the orthogonal projection onto $C_\perp$ as defined in Theorem 6.

The next theorem emphasizes that moving from one particular unconstrained problem to another is equivalent to a linear transformation of the variables.

Theorem 7 *Let C_1 and C_2 be two constraint subspaces verifying (22), and spanned by the columns of the full rank matrices R_1 and R_2, respectively. We then have that*

$$\mathcal{L}(\beta, \theta, R_2 z_2) = \mathcal{L}(\beta, \theta, R_2 T z_1)$$

for any $z_1, z_2 \in \mathbb{R}^{|\mathcal{E}|-1}$, where the linear nonsingular operator T is defined as

$$T = R_2^+ F_{C_2} R_1, \tag{38}$$

where the superscript $^+$ denotes the Moore-Penrose generalized inverse.

Proof. The result is directly obtained from definition (35). The nonsingularity of the linear operator results from (37) and the definition of the generalized inverse of a full rank matrix $R^+ = (R^T R)^{-1} R^T$, yielding $T^{-1} = R_1^+ F_{C_1} R_2$. □

This theorem raises the question whether local maxima of $\Pi(C_1)$ correspond, through the mapping T, to local maxima of $\Pi(C_2)$. This point is settled by the following easy corollary.

Corollary 8 *Let C_1 and C_2 be two constraint subspaces verifying (22), spanned by the columns of the full rank matrices R_1 and R_2, respectively. Also let (β, θ, z_1) be an (isolated) local maximum of problem $\Pi(C_1, R_1)$. Then $(\beta, \theta, T z_1)$ is an (isolated) local maximum of problem $\hat{\Pi}(C_2, R_2)$.*

Proof. The result immediately follows from the observation that problem $\Pi(C_2, R_2)$ can be deduced from $\Pi(C_1, R_1)$ by the nonsingular linear transformation T. □

This confirms the well-known result that the optimal values of β and θ do not depend on the particular specification of the ASCs, that is on the particular choice of the constraint subspace C.

The next theorem investigates the geometric properties of the log-likelihood function for different ASC specification strategies, that is for different choices of C. This is obtained from the comparison of the curvature of $\mathcal{L}(\cdot)$ in directions belonging to C and $C_\perp$, where $\mathcal{L}$ is considered, for given β and θ, as a function of c only. This curvature is given by the Rayleigh quotient

$$\rho(c, s) \stackrel{\text{def}}{=} \frac{\langle s, \nabla^2 \mathcal{L}(c) s \rangle}{\|s\|^2}, \tag{39}$$

where $\|\cdot\|$ is the Euclidean norm.

Theorem 9 *Let C be an arbitrary constraint subspace verifying (22). Also denote by $P_\perp$ the orthogonal projector onto $C_\perp$. Then we have that, for all $c \in I\!R^n$ and $s \in C$,*

$$\rho(c, s) = \rho(c, P_\perp s) \cos^2 \alpha, \tag{40}$$

where α is the angle between s and $P_\perp s$.

Proof. The proof is obtained immediatly from Definition (39) of the Rayleigh quotient, from the definition of $\cos \alpha$,

$$\cos \alpha = \frac{\langle s, P_\perp s \rangle}{\|s\| \, \|P_\perp s\|},$$

and from the invariance of the Hessian under orthogonal projection, as shown by Theorem 6, that is

$$\langle s, \nabla^2 \mathcal{L}(c) s \rangle = \langle P_\perp s, \nabla^2 \mathcal{L}(c) P_\perp s \rangle.$$

□

As an immediate corollary, we observe that the maximum curvature of the objective function is obtained with directions lying in the orthogonal subspace $C_\perp$, as illustrated in the following example.

Consider a binomial model, where the systematic component of the utility functions is composed only of ASCs: $V_1 = c_1$ and $V_2 = c_2$. A sample of size ℓ is available, where $\frac{\ell}{2}$ individuals have chosen alternative 1, while the other $\frac{\ell}{2}$ have chosen alternative 2. Obviously, the maximum likelihood is obtained with any value of c_1 and c_2 such that $c_1 = c_2$.

Figure 2 represents the invariant subspace M and two possible constraint subspaces: $C_\perp$ (the orthogonal subspace) and C_m (the subspace used with the classical strategy, defined by (28)).

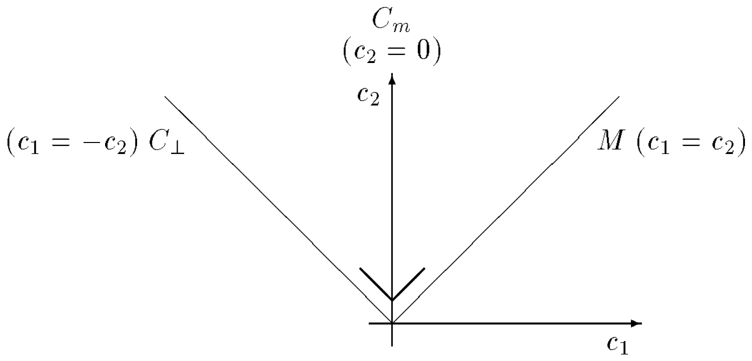

Figure 2: Subspaces for a simple binomial example

Constraining the log-likelihood function $\mathcal{L}(c_1, c_2)$ to subspaces M, $C_\perp$ and C_m, respectively, yields the corresponding one-variable functions which are plotted on Figure 3. Using Theorem 9, and noting that the angle $\alpha = 45°$ (and so $\cos^2 \alpha = \frac{1}{2}$), the curvature of the constrained log-likelihood function is doubled if the orthogonal strategy is preferred over the classical one.

5 Algorithmic implications and practical comments

We now analyze the impact of choosing one particular strategy on the behaviour of algorithms which maximize the log-likelihood function. Changing the strategy is equivalent to a linear

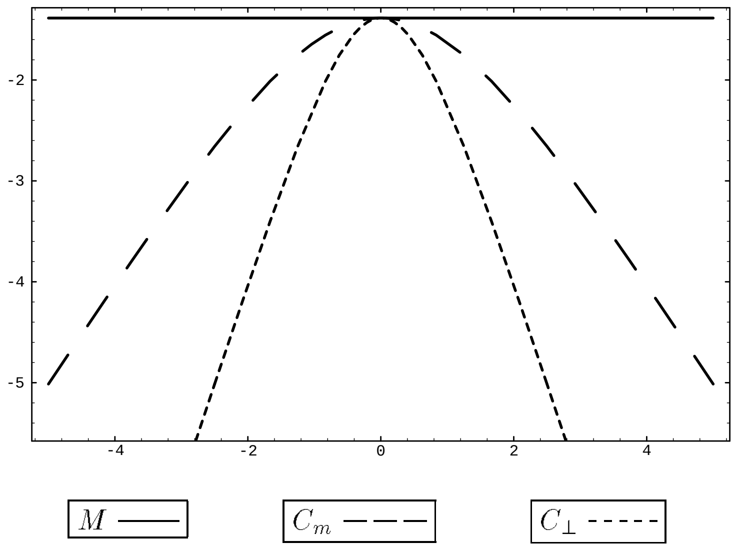

Figure 3: Variation of the curvature

transformation of the variables, as proved by Theorem 7. Some algorithms are independent of such transformation: Newton's method and quasi-Newton's methods of the Broyden family, combined with line searches, are some of the most popular techniques having this property. A detailed analysis of the invariance of algorithms is provided by Fletcher, 1987, pages 57–62. However, most of the methods designed to explicitly handle non-concavity in the quadratic model, like trust region methods (used namely in the HieLoW package by Bierlaire, 1995) and modified Newton methods (the version described by Berndt, Hall, Hall and Hausman, 1974 is used namely in ALOGIT (Daly, 1987) and in *LIMDEP, Version 7.0*, 1995), are not independant from a linear transformation of the variables.

In theory, trust region and modified Newton's methods can both be viewed as resolving the indefiniteness or singularity of the Hessian by substracting from it a positive multiple of the identity matrix, so as to obtain a safely negative definite matrix $\nabla^2\mathcal{L} - \lambda I$. The geometrical property of the orthogonal stimpliesrategy, characterized by Theorem 9, insures λ to be minimum in this case. The choice of the orthogonal strategy therefore implies an optimal behaviour of such algorithms, in that the true Hessian is perturbed as little as possible.

Finally, we observe that, if the estimation algorithm does not feature a scaling invariant stopping criterion, then the maximization can be terminated prematurely, if the orthogonal strategy is not used. Indeed, from Theorem 6 and the strict triangular inequality, the norm of the gradient is smaller than the norm of its projection onto the orthogonal subspace. This effect increases when the chosen constraint subspace approaches the invariant subspace.

Before concluding the paper, we mention some practical aspects of the ASC specification in general, and of the orthogonal strategy in particular.

1. The geometrical property of the orthogonal strategy described by Theorem 9 is based on an orthogonal projection onto $C_\perp$. Therefore, the matrix $R_\perp$ in problem $\Pi(C_\perp, R_\perp)$ has to be orthogonal. It is not a severe restriction, because the matrices typically arising in this context can be orthogonalized through a Gram-Schmidt or a Householder process (see, for

instance, Golub and Van Loan, 1989).

2. We emphasize that the choice of the alternatives whose ASC are constrained to be zero in elemental and structural strategies is not arbitrary. Indeed, a different choice corresponds to a different constraint subspace. The behaviour of an algorithm which is not invariant with regard to a linear transformation of the variables, can therefore differ from one choice to another.

3. The results of estimating ASCs using the orthogonal strategy may not be interpreted in a straightforward way. However, an equivalent and clearer solution can be easily obtained from the characterization (15) of the invariant subspace. Considering the example corresponding to Figure 1, suppose that $(c_1^*, c_2^*, c_3^*, c_4^*, c_5^*, c_6^*)$ are the values of the ASCs provided by the orthogonal strategy. It can be easily verified that the corresponding values for the elemental strategy, where ASC associated to node 1 has been constrained to be zero, are $(0, c_2^* - c_1^*, 0, c_4^* - c_3^* - c_1^*, 0, c_6^* - c_3^* - c_1^* - c_5^*)$. This result is directly obtained from the expression (17) of the invariant subspace basis, and is easy to implement.

4. The formulation (4) of the utility function of structural nodes is used by several authors, like Daganzo and Kusnic (1992), Brownstone and Small (1989), Ben-Akiva and Lerman (1985) or Forinash and Koppelman (1993). However, an alternate formulation is mentioned in some other references (see, for instance, Daly, 1987, Ortúzar and Willumsen, 1995), namely

$$U_k = c_k + \sum_i \beta_i x_i(k) + \theta_k \sum_{\ell \in \mathcal{C}(k)} e^{U_\ell}. \tag{41}$$

   We stress that the characterization of the invariant subspace described in this paper is no longer valid with hierarchical models using (41). In this case, it can be shown that the invariant subspace depends on θ, the basis vectors v_k, defined by (14), being then

$$v_k = \sum_{\ell \in \mathcal{C}(k)} e_\ell - \theta_k e_k. \tag{42}$$

   If a classical strategy is used, the angle between the invariant subspace and the constraint subspace is a function of θ. There is therefore no guarantee that the constraint subspace verifies (22) for any value of θ. Moreover, if the orthogonal strategy is preferred, it results in non-linear utility functions, which is usually not supported by existing estimation packages.

5. Automatic specification of all the strategies described in this paper has been included in HieLoW (Bierlaire, 1995).

6 Conclusion

The overspecification of the log-likelihood function of multinomial and nested logit models, due to the ASCs has been analyzed in detail and a geometrical description has been derived. The proposed formulation for the *invariant subspace* has been proved to fully characterize the overspecification due to the ASCs. Moreover, the subspace containing all the relevant geometrical information has been identified as the subspace orthogonal to the invariant subspace. This result has provided a natural technique, called the *orthogonal strategy*, to overcome the overspecification problem, by imposing feasible solutions to lie in this particular subspace. Since the orthogonal strategy does not seem to be used in practice, other classical strategies have also been presented. The relationship between any two arbitrary strategies has been characterized by a non-singular linear transformation of the maximization problem variables. We recommend to use the orthogonal strategy in practical applications, not only because of its maximum curvature property, which has desirable algorithmic implications, but mainly because all the geometrical information of the likelihood function lies in the orthogonal subspace. The orthogonal strategy is therefore the most natural way of removing the overspecification.

7 Acknowledgements

Useful discussions with M. Ben-Akiva, C. Bhat, E. Cascetta, A. Daly, B. Dellaert, S. Madanat and R. Mishalani were very much appreciated. The financial supports granted by the AGIR program of the Région Wallonne to the first author, and by a grant from Human Capital and Mobility of the Commission of the European Communities to the second author are also gratefully acknowledged.

References

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, USA.

Berndt, E. U., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement* **3/4**: 653–665.

Bierlaire, M. (1995). A robust algorithm for the simultaneous estimation of hierarchical logit models, *GRT Report 95/3*, Department of Mathematics, FUNDP.

Brownstone, D. and Small, K. A. (1989). Efficient estimation of nested logit models, *Journal of Business & Economics Statistics* **5**(1): 67–74.

Daganzo, C. F. and Kusnic, M. (1992). Another look at the nested logit model, *Technical Report UCB-ITS-RR-92-2*, Institute of Transportation Studies, University of California at Berkeley.

Daganzo, C. F. and Kusnic, M. (1993). Two properties of the nested logit model, *Transportation Science* **27**(4): 395–400.

Daly, A. (1987). Estimating "tree" logit models, *Transportation Research B* **21**(4): 251–268.

Fletcher, R. (1987). *Practical Methods of Optimization*, second edn, J. Wiley and Sons, Chichester.

Forinash, C. V. and Koppelman, F. S. (1993). Application and interpretation of nested logit models of intercity mode choice, *Transportation Research Record Issue* **1413**: 98–106.

Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations*, second edn, Johns Hopkins University Press, Baltimore.

Kreyszig, E. (1972). *Advanced engineering mathematics*, third edn, J. Wiley and Sons.

LIMDEP, Version 7.0 (1995). Econometric Software, Inc., Bellport, NY, USA.

Ortúzar, J. D. and Willumsen, L. (1995). *Modelling Transport*, 2nd edn, J. Wiley and Sons, Chichester (England).