On the Information Matrix in Mixed Logit Models Estimation

Fabian Bastin and Cinzia Cirillo

Advanced discrete choice models—in particular, mixed logit models—are used extensively in transportation. Although much progress in estimation techniques has made them numerically appealing, their properties have not been fully explored. This lack of exploration sometimes leads to confusing quality measurements and misinterpretation of the estimates. In this paper, the regularity conditions for which the information equality holds are reviewed, and some underlying technical difficulties in the context of mixed logit modeling are discussed. This paper specifically addresses the questions of correlations between estimated parameters and the validity of the asymptotic normality assumption in complex models, as nonparametric formulations. In the latter case, the population is resampled with the use of bootstrap principles to construct confidence intervals on the estimated parameters. Numerical tests on simulated data are presented to assess the relevance of the problem and the validity of the methods proposed.

Discrete choice models rely heavily on statistical theory of model estimation. Demand modelers in the transportation field use a sample of observations from the process being modeled, randomly drawn from the whole population, to estimate unknown parameters of some utility function (often assumed to be linear). In most travel demand analysis, the population can be assumed to be infinite, although any sample consists of a finite number of observations [see, for instance, Ben-Akiva and Lerman (*1*), chap. 2]. Several methods are used to find estimators with known statistical properties: maximum likelihood, least squares, or squared moment conditions. More recently, simulated likelihood functions have been introduced to estimate more complex discrete choice models, with no closed mathematical formulation for the choice probability (*2*).

For simple models and correct specification assumptions, which is in practice difficult to guarantee, the estimates asymptotically converge to a normal distribution centered on zero and with a variance–covariance matrix equal to the inverse of the information matrix. In mixed logit models, especially those specified as random coefficients, the coefficients are assumed to vary over the population with an underlying density function; this density can be parametric or nonparametric [see, for instance, Fosgerau and Bierlaire (*3*) and Bastin et al. (*4*)]. The number of parameters to be estimated for each attribute depends on the distributional form. Normal and log-normal distributions require the estimation of mean and standard deviation (*5*); truncated normal also requires the estimation of the mass(es); Johnson SB requires the lower bound and the upper bound of the distribution to be fixed or estimated (*6*); and nonparametric models require a number of parameters depending on the supporting points chosen by the analysts. As shown in this paper, the fact that several parameters are affected to the same factor creates correlations when the standard information matrix is used. The correct calculation of the variance–covariance matrix is relevant also for the construction of optimal experimental design in stated choice data collection (*7, 8*). Most methods proposed in the literature consider the minimization of the standard errors obtained from data collected and hence the maximization of the asymptotic *t*-statistics [for a study in the context of mixed logit estimation, see Bliemer and Rose (*9*)]. When the regularity conditions needed to estimate the variance–covariance matrix of the estimates are violated as in nonparametric mixed logit, different inference methods should be considered. Bootstrap techniques in particular can be used to estimate confidence intervals, bias, and variance of an estimator or to perform statistical tests. The rest of this paper is organized as follows. A brief introduction to mixed logit models is presented next, followed by a discussion of the mathematical background relevant to the use of the information matrix. Results from three simulated experiments, including D-optimal design and nonparametric random coefficient logit models, are reported next. Conclusions and findings from the proposed analysis are given in the final section of the paper.

MIXED LOGIT MODEL

The mixed logit model is a quite general formulation for individual choices between discrete options. Consider a set of N individuals, each one having to choose one alternative within a finite set Aj. A utility U_{nj} is associated with each alternative A_j in $A(n)$, as perceived by individual n. With reliance on the econometric theory, it is also assumed that individuals aim to maximize their utility, but not all components are observed. Instead, the utility U_{nj} is decomposed as the sum of a deterministic part $V_{nj}(\beta)$, where β is a vector to estimate, and a random, unobserved part ϵ_{nj}. The probability choice is then

$$L_{nj}(\beta) = P\left\lfloor V_{nj}(\beta) + \epsilon_{nj} \geq V_{na}(\beta) + \epsilon_{na},\ \forall A_a \in A(n)\right\rfloor$$

where

- L_{nj} = probability that individual n chooses alternative j;
- P = probability operator;
- V_{nj} = deterministic part of utility of alternative j, as perceived by individual n; and
- ϵ_{na} = random part of utility of alternative a, for individual n.

F. Bastin, Department of Computing Science and Operational Research, University of Montreal, CIRRELT, CP 6128, Succursale Centre-Ville, Montreal, Quebec H3C 3J7, Canada. C. Cirillo, Department of Civil and Environmental Engineering, University of Maryland, 1173 Glenn L. Martin Hall, Building 088, College Park, MD 20742. Corresponding author: F. Bastin, bastin@iro.umontreal.ca.

Transportation Research Record: Journal of the Transportation Research Board, No. 2254, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 11–18.
DOI: 10.3141/2254-02

In the rest of this paper only linear utilities will be considered (as is standard practice in the transportation field). The probability expression is of course dependent on the distribution choice for ϵ_{nj} when the ϵ_{nj}'s are assumed to be independent and identically distributed. When the ϵ_{nj}'s are assumed to be independent and identically distributed Gumbel distributions among the individuals and alternatives, the traditional logit probability is obtained. In the mixed logit framework, the assumption that β is a constant vector is relaxed and it is allowed to be random with cumulative distribution function (CDF) $F_B(\beta)$ so that the probability choice L_{nj} is now conditional on the realization β, and the unconditional probability is

$$P_{nj} = E_\beta\left[L_{nj}(\beta)\right] = \int L_{nj}(\beta)\, dP_\beta(\beta) \tag{1}$$

Therefore β cannot be directly estimated, so it will be assumed that β can be described as $\beta = (\Gamma, \theta)$, where Γ is a random vector (in practice, an m-dimensional uniform, $U[0, 1]^m$), and θ is some parameter vector to be estimated. In other words, a distribution family is assumed for β, parameterized by θ. If, moreover, the vector β is continuous, Equation 1 can be rewritten as

$$P_{nj}(\theta) = \int L_{nj}(\gamma, \theta)\phi(\gamma, \theta) d\gamma$$

where $\phi(\gamma, \theta)$ is the density of β, with parameters vector θ, and γ is a realization of the random vector Γ. In the case in which the same individual can express several choices, for each individual the sequence of choices $y_n = (j_{n1}, \mathrm{K} j_{nT_n})$ is observed, which can be assumed to be correlated, and the data will be considered as panel data. A simple way to accommodate this situation is to assume that the heterogeneity is present only on the population level, not on the individual level. The probability of observing the individuals' choices is then given by the product of logit probabilities $L_{nj_{nt}}$, as expressed by Train (*2*):

$$P_{ny_t}(\theta) = \int \left(\prod_{t=1}^{T} L_{nj_{nt}}(\gamma, \theta) \right) \phi(\gamma, \theta) d$$

The parameters θ are estimated by maximizing the log likelihood function

$$\mathrm{LL}(\theta) = \ln \sqrt[N]{\prod_{n=1}^{N} \ln P_{nj_n}(\theta)} = \frac{1}{N} \sum_{n=1}^{N} \ln P_{nj_n}(\theta) \tag{2}$$

NONPARAMETRIC MIXED LOGIT

Mixed logit models based on nonparametric random coefficients allow the estimation of taste heterogeneity without imposing strong assumptions on the underlying distributions. They are gradually replacing discrete treatments of the parameters that could lead to an arbitrary population segmentation. In this study it is assumed that the components of the random vector β are themselves random and can be considered separately if they are independent. A well-known technique to generate a set S_X of draws from a univariate random variable $\boldsymbol{X}$ is the inversion technique. It consists of sampling a uniform $U[0, 1]$ and applying the inverse cumulative distribution function F_X^{-1} to these draws:

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

It is usually assumed that F_X^{-1} is available (or at least some numerically good approximation of it), the distribution $\boldsymbol{X}$ being known. This method is well-known in the random numbers generation literature (*10, 11*) and is also popular in the context of variance reduction methods [see, for instance, L'Ecuyer (*12*)]. One can capitalize on this approach by expressing the inverse cumulative distribution function as some element in a functional space:

$$F_X^{-1}(\mathrm{o}) = \sum_{k=1}^{\infty} q_k h_k(\mathrm{o})$$

where h is a basis function, so that the set of h defines the basis of the considered functional space, and $\{h_k, k = 0, \mathrm{K}, \infty\}$ constitutes a basis of this space, and the q_k's are the coordinates, to estimate, of the (cumulative) distribution function F_X (if the basis cardinal is finite, and equal to n; h_k and q_k are just set to 0, for $k > n$), and k is a dimensional index. If it is furthermore assumed that the random variable $\boldsymbol{X}$ 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 restrictive because extreme behaviors, corresponding to values of $\boldsymbol{X}$ tending to plus or minus infinity, are usually not welcome. In this paper, cubic B-splines will be considered, and

$$q_1 \le q_2 \le \mathrm{K}$$

will be required to ensure that F_X^{-1} is monotonically increasing [for more details, see (*4*)].

MATHEMATICAL FOUNDATIONS OF INFORMATION MATRIX

Although practitioners often make use of the information matrix properties to make inferences about the estimated parameters, the underlying theory is barely explained. It is nevertheless thought that in the presence of complex models, it is important to understand the properties of the estimators for a valid application to real case studies and a correct interpretation of the results. In this section, those properties are reviewed, and more technical details are provided by Newey and McFadden (*13*).

The statistical estimator of the parameters vector θ_0 is considered, and θ_0 is defined as a function of the form

$$h(\mathrm{o}) : \chi \to \Theta \subseteq \Re^K$$

where χ is the sample space (here, the population and choice situations) and Θ is the parameter space (i.e., the set of values that the estimator is allowed to take). The estimator will be written as

$$\hat{\theta} = h(X_1, X_2, \mathrm{K}, X_N) := h(X)$$

where K is the dimension of the parameter space θ.

A particular value of this estimator, based on a particular sample realization, $x := (x_1, x_2, \mathrm{K}, x_N)$, is called an estimate of θ_0, denoted by

$$\hat{\theta} = h(x_1, x_2, \mathrm{K}, x_N) := h(x)$$

The sampling distribution of $\hat{\theta}$ is defined as the distribution of $h(X)$, and its density function is denoted by

$$f(\hat{\theta}; x_1, x_2, \mathrm{K}, x_N) := f(\hat{\theta}; x)$$

The distribution of the sample is denoted by $f(x; \theta_0)$, the joint distribution of the random vector $X := (X_1, X_2, \mathsf{K}, X_N)$ under the parameter θ_0. Unfortunately, the form of f is rarely known, so an approximation of it is used, denoted hereafter by $\hat{f}$. A special case, illustrated in Equation 2, is the maximum likelihood estimator, defined as

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \hat{f}\left(\tilde{\theta}; x\right)$$

For an independent and identically distributed sample, one has

$$\hat{f}\left(\tilde{\theta}; x\right) = \frac{1}{N}\prod_{i=1}^{N} \hat{f}\left(\tilde{\theta}; x_i\right)$$

and it is usually preferable to solve the mathematically equivalent but numerically more tractable program

$$\max_{\theta \in \Theta} \frac{1}{N}\sum_{i=1}^{N} \ln \hat{f}\left(\tilde{\theta}; x_i\right)$$

Also $\hat{\theta}_{ML}$ is rewritten as $\hat{\theta}_N$ to stress the relationship of the estimator to the sample size N.

A probability model is said to be regular if the distribution of the sample $\hat{f}(x; \theta) = \hat{f}(x_1, x_2, \mathsf{K}, x_N; \theta)$ satisfies the regularity conditions stated below [for more details, see Spanos (*14*) and Lehmann and Casella (*15,* chap. 2)]. The conditions are

R1. The parameter space Θ is an open subset of $\Re^K$, $K < N$.

R2. The support of the distribution $X_0 := \{x \mid \hat{f}(x; \theta) > 0\}$ is the same for all $\theta \in \Theta$.

R3. The score function (or the score), defined as the gradient of $\ln \hat{f}(x; \theta)$ at θ, $s(x; \theta) = \nabla_\theta \ln \hat{f}(x; \theta)$, exists and is finite for all $\theta \in \Theta$, $x \in \chi_0$.

R4. Considering the estimator $h(x)$, one can interchange differentiation with respect to θ and integration with respect to x, that is, $\nabla_\theta \int h(x)\, \hat{f}(x; \theta) dx = \int h(x)\, \nabla_\theta\, \hat{f}(x; \theta) dx < \infty$.

For such regular probability models, the Fisher information for the sample $(x_1, x_2, \mathsf{K}, x_N)$ is defined as the matrix of second moments of the score evaluated at the true parameters vector:

$$I(\theta) = E\left[s(X; \theta) s(X; \theta)^T\right] \tag{3}$$

which can be estimated as

$$I_N(\theta_N) = \frac{1}{N}\sum_{i=1}^{N} s(x_i; \theta_N) s(x_i; \theta_N)^T \tag{4}$$

It is often stated that the information matrix is defined as the opposite of $H(\theta_0)$, the Hessian of the log likelihood [see, for instance, Train (*2,* sect 8.6)]. One can, however, identify the information matrix to $H(\theta_0)$ only at the price of additional assumptions that are developed below, leading to the famous information matrix equality property. It is therefore more correct and safer to keep Equation 3 as the definition of the information matrix, especially because many popular models do not meet the requirements needed for this equality.

Asymptotic Normality of Maximum Likelihood Estimators

Let $L(\theta; x) = \ln \hat{f}(\theta; x)$. Under R1, the gradient of the log likelihood function vanishes at $\hat{\theta}_N$:

$$0 = \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \ln f\left(\theta_N; x_i\right)$$

In addition to R1 through R4, the following assumptions are also required:

R5. $L(o, x)$: $\Theta \to [0, \infty)$ is continuous at all points $\theta \in \Theta$.

R6. For all $\theta_1 \neq \theta_2$, where $\theta_1, \theta_2 \in \Theta$, $f(\theta_1; x) \neq f(\theta_2; x)$.

R7. $E[\ln f(\theta; X)]$ exists, and $\theta_0 = \arg\max_{\tilde{\theta}} E\,[\ln f(\tilde{\theta}; X)]$.

R8. $1/N\, L(\theta; X)$ converges almost surely to $E\,[\ln f(\theta; X)]$ for all $\theta \in \Theta$.

R9. $\ln L(\theta; x)$ is twice differentiable in open interval around θ.

R10. θ_0 and $\hat{\theta}_N = \arg\max_{\tilde{\theta}} 1/N\, L(\tilde{\theta}; X)$ belong to some compact subset of Θ, almost surely, for N large enough.

These conditions ensure that the estimator $\hat{\theta}_N$ is consistent, that is, $\hat{\theta}_N \to \theta_0$, even if the distribution $\hat{f}$ is misspecified ($\hat{f} \neq f$). Then, from the Slutzky theorem [see Newey and McFadden (*13*), sect. 3],

$$\sqrt{N}\left(\hat{\theta}_{\mathrm{ML}} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}(\theta_0) I(\theta_0) H^{-1}(\theta_0)\right) \tag{5}$$

Assume, furthermore, the following:

R11. The operations of integration of differentiation can be swapped for the second derivatives of $f(x; \theta)$ when considering $h(x)$, that is,

$$\nabla^2_{\theta\theta} \int h(x) f(x; \theta) dx = \int h(x) \nabla^2_{\theta\theta} f(x; \theta) dx$$

Under the regularity conditions R1 through R11, it can be shown that

$$I(\theta_0) = -E\left[\nabla^2_{\theta\theta} \ln f(X; \theta_0)\right] \tag{6}$$

that is, the information matrix is equal to the opposite of the log likelihood Hessian, evaluated at θ_0 [see, for instance, Lehmann and Casella (*15,* Lemma 5.3)]. This equality, known as the information matrix equality, is often used in practice, as the variance–covariance matrix can then be reduced to $H^{-1}(\theta_0)$ in Equation 5.

Exceptions and Violations to Normality Assumptions

Assumption R11 does not hold in the presence of misspecification; in that case the information matrix equality is not valid. For consistent specifications, $H^{-1}(\theta_0) I(\theta_0) H^{-1}(\theta_0)$ still gives the asymptotic variance–covariance matrix, which can be estimated as $H_N^{-1}(\hat{\theta}_N) I_N(\hat{\theta}_N) H_N^{-1}(\hat{\theta}_N)$. An important example of consistent misspecification is when the observations X_1, K, X_N are dependent, so that one has

$$f(\theta, X_1, \mathsf{K}, X_N) \neq \prod_{i=1}^{N} f(\theta; X_i)$$

That result happens, for instance, when a multinomial logit is estimated by using panel data, because the observations from a given individual are typically correlated. Many studies have been devoted to proving the consistency of the estimator, under the false assumption of independence (but with an otherwise correctly specified model). This situation shows that the estimator $H_N^{-1}(\hat{\theta}_N) I_N(\hat{\theta}_N) H_N^{-1}(\hat{\theta}_N)$ is still valid, but not $H_N^{-1}(\hat{\theta}_N)$, which cannot be viewed any longer as an approximation of the information matrix. This point is briefly illustrated in Case Study 2. This outcome is of course not a novel result (*16, 17*), but that is often ignored in transportation applications.

Assumption R7 is violated when considering mixed logit models, as often there is more than one solution to the maximum likelihood problem. For instance, a normally distributed parameter is characterized by two parameters, the mean and the standard deviation. However, the optimization program can deliver a negative standard deviation if no constraint is imposed, and this standard deviation can be interpreted in absolute value. However, the first part of R7 can be enforced by reducing the feasible set, for instance in the normal case, imposing the estimated standard deviation to be positive. But it also means that for a fixed number of draws, the estimator is not consistent because it is biased, so that one should be careful with the derived variance–covariance matrix. In practice, there are various ways to correct the bias, either by using "intelligent draws" (*18*) or by directly adding a correction to the log likelihood (*19*). This correction nevertheless does not affect the optimization bias, also described by Bastin and Cirillo (*19*), that is present especially when the log likelihood simulation estimator has a large variance, which is usually the case if there is a lot of heterogeneity in the population. That large variance makes the consistency assumption, and consequently the information matrix equality, more doubtful. It is then safer to apply the robust estimator while one cannot guarantee even the asymptotically normality, especially for large population sizes, because the log likelihood then tends to have more bias (*20*). In this paper, to limit these effects, the lattice rules proposed by Munger et al. will be used, designed specifically for mixed-logit models that perform significantly better than other constructions as Halton sequences (*18*).

Under the assumption that the estimator is normally distributed, the statistic

$$t_{\theta_i} = \frac{\hat{\theta}_i - \theta_{0,i}}{\hat{\sigma}_i}$$

where $\hat{\theta}_i$ designs the ith component of $\hat{\theta}$ and $\hat{\sigma}_i$ is its estimated standard deviation, follows a Student distribution, with mean 0 and variance 1, which can be approximated by an $\mathcal{N}(0, 1)$. In practice, the number of observations is usually large enough to make this approximation reliable. A confidence interval of level α can be constructed as

$$\left(\theta - Z_{\alpha/2}\hat{\sigma}_i, \hat{\theta}_i + Z_{\alpha/2}\hat{\sigma}_i\right)$$

where $Z_{\alpha/2}$ is the $1 - \alpha/2$ quartile of an $\mathcal{N}(0, 1)$. Considering the statistical test

$$H_0 : \theta_i = 0, H_1 : \theta_i \neq 0$$

for some i in $\{1, \mathsf{K}, K\}$, H_0 is rejected if 0 does not belong to this interval.

Bliemer and Rose (*9*) developed the expression of the Hessian in the context of mixed logit; the information matrix estimation (Equation 4) is used as an approximation of the Hessian in the context of Newton-based methods, leading to the popular Berndt, Hall, Hall, and Hausman (BHHH) approach (*21*). However, the equality (Equation 6) holds only at the true parameters. In the context of mixed logit modeling, this can sometimes lead to poor behavior of the BHHH method (*22*), as it cannot be ensured in practice that the probability family has been correctly identified. Here the log likelihood maximizations are performed by using a modified version of AMLET, taking advantage of more robust trust-region methods and SR1 or Broyden, Fletcher, Goldfaro, and Shanno (BFGS) approximations (*23*). In this context, R1 could seem unusual to some readers, as one often requires the feasible set to be compact, for instance to guarantee convergence of the optimization procedure or to establish consistency of the estimator [see Newey and McFadden (*13*), sect. 1]. However, one has to ensure that the maximum estimator does not lie on the feasible set boundary to establish its asymptotic distribution, as the log likelihood gradient would be required to cancel out at that point. Therefore, one cannot make an inference on the basis of the information matrix when R1 is violated, as will be illustrated in Case Study 3.

NUMERICAL EXPERIMENTS

Having reviewed the theory behind the information matrix, one can numerically explore, in three case studies, its behavior when estimating mixed logit models. The first case study is aimed at exhibiting creation of correlations between estimated parameters and analyzing the applicability of the information equality; the effect of the coverage of the integration space is also investigated. The second case study is about the relevance of the information matrix for efficient stated choice designs. Finally, the third case study proposes an alternative method to calculate *t*-statistics when the hypotheses of applicability of the information matrix are violated.

Case Study 1. Parametric Random Coefficient Experiment

The first case study involves a synthetic panel of 1,000 individuals, each contributing 10 observations. The model contains five alternatives, each specified with three attributes normally distributed with mean zero and variance one. One of the three coefficients is constant ($\beta = -0.4$); the remaining two are independent and normally distributed ($\beta_2 = \mathcal{N}(0.2, 1.0)$, $\beta_3 = \mathcal{N}(0.8, 0.25)$). Results relative to this experiment are given in Table 1. Different types of draws have been used for model estimation (Monte Carlo and randomized lattices), each time using 4,093 draws per individual. The deriving *t*-statistics, given in parentheses and computed with the robust variance–covariance

TABLE 1 Case Study 1. Model Estimation

Variable	MNL (*t*-stat.)	MMNL Monte Carlo	MMNL Lattice (*t*-stat.)
β_1	–0.326 (26.44)	–0.400 (33.30)	–0.400 (33.20)
β_2	0.171 (6.86)	0.255 (4.69)	0.254 (4.64)
σ_2	—	0.978 (16.73)	0.980 (16.41)
β_3	0.810 (36.98)	0.811 (24.58)	0.811 (24.91)
σ_3	—	0.486 (15.49)	0.486 (15.71)

NOTE: MNL = multinomial logit; MMNL = mixed multinomial logit; — = not applicable; *t*-stat. = *t*-statistic.

matrices, are not substantially different from each other, confirming that there is no bias due to poor coverage of the integration space. The multinomial logit recovers the ratio between the mean of the distribution (but not their individual values), and the mixed logit formulation matches the data used for population generation. The *t*-statistics also tend to be smaller in absolute value when the robust information matrix for the multinomial logit is used, but a clear pattern for the mixed logit formulation is not observed. Most important, computed correlations, as reported in Table 2, are very low for the multinomial logit, which is in line with the fact that all parameters were generated independently. However, significantly higher values are observed when random coefficients are estimated. A correlation as high as 0.91 is reported between the mean and the standard deviation of β_3. The model exhibits a great deal of heterogeneity, by construction, diverging largely from the multinomial logit situation. For a given factor, several parameters compete to capture information, leading, it is assumed, to the observed correlations. The reported results in Table 2 are based on lattice draws, but similar results have been obtained with Monte Carlo draws.

Because one is faced with panel data, it is also interesting to compare the opposite of the Hessian and the information matrix, evaluated at the solution. For the multinomial logit, as expected, because of the misspecification, quite different matrices were obtained:

$$-H\left(\hat{\theta}\right) = \begin{pmatrix} 6.956 & 0.135 & 0.578 \\ 0.135 & 7.191 & -0.255 \\ 0.578 & -0.255 & 0.615 \end{pmatrix}$$

$$I\left(\hat{\theta}\right) = \begin{pmatrix} 7.297 & 0.207 & 0.956 \\ 0.207 & 32.042 & -0.636 \\ 0.956 & -0.636 & 11.671 \end{pmatrix}$$

Clearly, the information matrix equality does not hold. More surprisingly, the same effect appears with the mixed logit formulation:

$$-H\left(\hat{\theta}\right) = \begin{pmatrix} 5.450 & 0.027 & 0.196 & 0.241 & 0.209 \\ 0.027 & 0.836 & -0.053 & -0.017 & -0.016 \\ 0.196 & -0.053 & 1.048 & -0.147 & -0.104 \\ 0.241 & -0.017 & -0.147 & 2.140 & -0.464 \\ 0.209 & -0.016 & -0.104 & -0.464 & 1.981 \end{pmatrix}$$

TABLE 2 Case Study 1: Correlation Matrices

	β_1	β_2	σ_2	β_3	σ_3
MNL					
β_1	1.000	–0.040		–0.083	
β_2	–0.040	1.000		0.053	
β_3	–0.083	0.053		1.000	
MMNL					
β_1	1.000	–0.111	–0.252	–0.475	–0.342
β_2	–0.111	1.000	0.388	0.147	0.003
σ_2	–0.252	0.388	1.000	0.240	0.041
β_3	–0.475	0.147	0.240	1.000	0.914
σ_3	–0.342	0.003	0.041	0.914	1.000

$$I\left(\hat{\theta}\right) = \begin{pmatrix} 3.530 & 0.033 & 0.096 & -0.700 & -0.138 \\ 0.033 & 1.988 & 0.741 & 0.172 & -0.277 \\ 0.096 & 0.741 & 3.639 & -0.094 & -0.856 \\ -0.700 & 0.172 & -0.094 & 2.868 & 2.035 \\ -0.138 & -0.277 & -0.856 & 2.035 & 2.291 \end{pmatrix}$$

Various additional experimentations were performed with different data schemes, which are not reported because of space considerations. Similar behavior was observed as soon as the variance of the random parameters was large, suggesting that the conditions of the information matrix equality are not satisfied, so it is safer to stick to the robust variance–covariance matrix estimation. Additional investigations on this issue are currently being pursued.

Case Study 2. Optimal Design for Mixed Logit on Panel Data

Because the information matrix indirectly provides the estimate of the variance–covariance matrix, several authors capitalize on the matrix to design stated-choice experiments. The aim here is not to explore how to construct a good design; the aim is simply to examine some methods proposed in the literature with respect to the inference theory, especially the induced correlations. In particular, Bliemer and Rose propose the use of the *D*-error measure when designing experiments for mixed logit models (*9*). Given some prior, they aim to minimize the *D*-error, defined as

$$\Delta\left(\mathrm{Cov}\left(\hat{\theta}\right)\right)^{1/K} \tag{7}$$

where $\Delta(A)$ designs the determinant of matrix *A,* and Bliemer and Rose estimate $\mathrm{Cov}(\hat{\theta})$ as $H_N^{-1}(\hat{\theta})$. Another popular measure is the *A*-measure, calculated as

$$\frac{\mathrm{tr}\left(\mathrm{Cov}\left(\hat{\theta}\right)\right)}{K}$$

where tr is trace.

D-error and *A*-error significantly differ in that the *A*-error relies on the matrix diagonal only, that is, the variances but not the covariances, whereas the *D*-error takes advantage of the entire covariance–variance matrix. As a result, it can be expected that a *D*-optimal design, minimizing Equation 7, tends to create correlations between the studied factors, as a way to decrease the total error. To explore that further, Case Study 1 as proposed in Bliemer and Rose is reproduced (*9*). They assumed a mixed logit model with panel formulation, developed the expression of the Hessian, and searched for a design minimizing Equation 7. The experiment furthermore supposes that each respondent is faced with two alternatives described by four attributes, each with three levels of variation, and has to answer to nine choice situations. The levels of the four attributes for both alternatives in the nine choice situations are presented in Table 3. The following prior values were chosen to derive the optimal design: $\beta_1 = \mathcal{N}(0.6, 0.04)$, $\beta_2 = \mathcal{N}(-0.9, 0.04)$, $\beta_3 = -0.2$, and $\beta_4 = 0.8$. The minimization procedure is not addressed, but the results, reproduced in Table 3, are taken for granted. Bliemer and Rose recommendations (*9*) are followed; a population of 100 individuals is generated when dealing with a mixed logit model and as few as 23 individuals are generated

TABLE 3 Case Study 2: *D*-Optimal Design

Choice Situations (s)	Attribute x_{j1}	x_{j2}	x_{j3}	x_{j4}
Choice *s*1				
Alternative *j*1	3	3	1	*2*
Alternative *j*2	1	1	*2*	*2*
Choice *s2*				
Alternative *j*1	2	1	1	1
Alternative *j*2	2	3	3	3
Choice *s*3				
Alternative *j*1	3	3	3	2
Alternative *j*2	1	1	1	1
Choice *s*4				
Alternative *j*1	3	2	2	3
Alternative *j2*	1	2	2	2
Choice *s*5				
Alternative *j*1	1	2	2	3
Alternative *j2*	3	2	3	1
Choice *s*6				
Alternative *j*1	1	1	3	2
Alternative *j*2	3	3	1	2
Choice *s*7				
Alternative *j*1	2	2	1	1
Alternative *j*2	2	2	3	3
Choice *s*8				
Alternative *j*1	2	1	3	1
Alternative *j*2	2	3	1	3
Choice *s*9				
Alternative *j*1	1	3	2	3
Alternative *j*2	3	1	2	1

for the multinomial version of the same model, obtained by setting all the β's to be constant among the individuals.

Estimation results are summarized in Table 4; 16,381 randomized lattice draws per individual were used for the mixed logit model. The logit model reproduces the average values of the normally distributed coefficients quite well, and reasonably well with the constant parameters. However, it can be seen that a sample of 23 individuals is too small a size to accommodate asymptotic analysis (valid for $N \to \infty$). A bootstrap procedure, resampling over the population, not the observations, was applied to take the dependencies into account. As expected, the variance–covariance matrices are not stable for such a small population, so the multinomial logit model with 100 individuals (900 observations) was also estimated, and the results have been found much more reliable (by using again a bootstrap procedure). Correlations are already very high (in absolute values) for logit estimates

TABLE 4 Case Study 2: MNL and MMNL Model Estimation

Variable	MNL, 23 individuals (*t*-stat.)	MNL, 100 individuals (*t*-stat.)	MMNL Lattice (*t*-stat.)
β_1	0.642 (4.61)	0.598 (8.59)	0.624 (8.45)
σ_1	—	—	0.217 (2.16)
β_2	–0.970 (5.74)	–0.905 (10.70)	–0.951 (10.64)
σ_2	—	—	0.263 (3.21)
β_3	–0.261 (2.73)	–0.162 (3.62)	–0.172 (3.64)
β_4	0.722 (4.41)	0.841 (10.68)	0.872 (10.85)

NOTE: — = not applicable; *t*-stat. = *t*-statistic.

(for which only the robust estimates are considered because misspecification is faced), as shown in Table 5. For instance, correlation between β_2 and β_4 is found to be close to –0.8, which is certainly not desirable in many practical cases. Correlations remain important for the mixed logit model. Substantially different values are noted between standard and robust estimates of correlation factors for the lowest values, but the highest values are similar. This similarity confirms the tendency of *D*-error minimization to create correlation structure.

As in the previous case study, the opposite of the Hessian is not valid for the multinomial logit because of the panel structure. As an illustration, the Frobenius-norm of the difference between the information matrix and the opposite of the Hessian at the estimated parameters were found to be 2.35 and 2.64 for the multinomial logit model, with 23 and 100 individuals, respectively, but only 1.39 for the mixed-logit model; in this case, one can estimate the matrices only by simulation, and there are larger matrices. The problem of consistency is alleviated, when compared with Case Study 1, because the variances are much smaller, which is also reflected by the values of the parameter means being close to their multinomial logit equivalent. In the following table, the *A*-error and *D*-error measures are also reported for logit and panel mixed logit, scaled with the number of individuals approximating the expectation by averaging over the synthetic population:

Error Measure	*MNL (23 individuals)*	*MNL (100 individuals)*	*MMNL*	*MMNL–Robust*
A-error	0.482	0.506	0.696	0.651
D-error	0.291	0.283	0.462	0.432

The values obtained are close to those calculated by Bliemer and Rose (*9*). The variations can easily be explained because the syn-

TABLE 5 Case Study 2: MNL and MMNL Correlation Matrices

	β_1	σ_1	β_2	σ_2	β_3	β_4
MNL (23 individuals)						
β_1	1.000		–0.667		–0.008	0.548
β_2	–0.667		1.000		0.240	–0.812
β_3	–0.008		0.240		1.000	–0.242
β_4	0.548		–0.812		–0.242	1.000
MNL (100 individuals)						
β_1	1.000		–0.700		–0.300	0.627
β_2	–0.700		1.000		0.387	–0.77
β_3	–0.300		0.387		1.000	–0.242
β_4	0.627		–0.777		–0.469	1.000
MMNL						
β_1	1.000	0.196	–0.683	0.106	–0.281	0.658
σ_1	0.196	1.000	–0.250	0.102	–0.097	0.219
β_2	–0.655	–0.250	1.000	–0.152	0.354	–0.790
σ_2	0.106	0.102	–0.152	1.000	–0.068	0.082
β_3	–0.281	–0.097	0.354	–0.068	1.000	–0.377
β_4	0.658	0.219	–0.790	0.082	–0.377	1.000
MMNL–Robust						
β_1	1.000	0.214	–0.684	–0.031	–0.275	0.622
σ_1	0.214	1.000	–0.196	0.180	0.046	0.070
β_2	–0.684	–0.196	1.000	–0.030	0.352	–0.763
σ_2	–0.031	0.180	–0.030	1.000	0.020	–0.015
β_3	–0.275	0.046	0.352	0.020	1.000	–0.439
β_4	0.622	0.070	–0.763	–0.015	–0.439	1.000

thetic populations have been independently generated; a different quasi–Monte Carlo construction is used for simulation, and the robust variance–covariance matrix estimation is used instead of the (opposite) Hessian of the log likelihood function.

Case Study 3. Nonparametric Mixed Logit Model

The third artificial data set is cross-sectional and simulates 2,000 individuals choosing once across a set of five discrete alternatives. Each alternative has three attributes drawn from $\mathcal{N}(0.5, 1.0)$; the three coefficients in the utility functions are all random and assumed to be normal with mean 0 and standard deviation 2.0, lognormal with parameters $\mu = 0$ and $\sigma = 1.0$, and spline with control points {−15.0, –3.0,–0.5,0.2, 0.5, 0.5, 3.0, 15.0}. The model was estimated by using 1,021 randomized lattice draws per individual. Results are reported in Table 6. Cross-sectional models are difficult models to calibrate because a large population size was needed to obtain significant parameters (*9*). The parameters are however well recovered, especially for the normal and lognormal distributions. The support of the spline is captured, and even if the coefficient does not directly correspond to the original control points, the global form of the distribution is better. No parameter associated with the spline appears to be significant, and although no results are reported here, the situation does not improve if the number of control points is varied or if Monte Carlo draws are used (which moreover do not reproduce the spline at the same extent). The coefficients are heavily dependent because of the monotonicity constraint, and the normality assumption does not hold any longer because, in particular, R1 is violated. The *t*-statistic test cannot be applied to decide which individual parameter is useful in the model specification. One can, however, turn to the likelihood ratio test to verify that the factor associated to the spline is significant. The likelihood ratio test serves the same function for maximum likelihood estimation that the *F*-test serves for least squares (*1*). The log likelihood function is compared for the unrestricted (L^u) and restricted models (L^r) under study. The hypothesis test is the following:

$$H_0:$$
$$\theta_i = 0 \qquad i \in C$$
$$H_1:$$
$$\theta_i \neq 0 \qquad i \in C$$

where C is a subset of $\{1, \ldots, K\}$. Then

$$-2\left(L^r - L^u\right)$$

is calculated, which asymptotically follows a χ^2 distribution with $K^u - K^r$ degrees of freedom. The null hypothesis H_0 is rejected if

$$-2\left(L^r - L^u\right) \geq \chi^2_{\left(1-\alpha, K^u - K^r\right)}$$

where $1 - \alpha$ is the significance level. In the case study, a rejection level of one is obtained for the null hypothesis, with optimal log likelihood values (scaled with the population size) of –2,378.67 and –2,419.41, depending on the inclusion of the spline parameters vector. In other words, the influence of the spline parameters vector is found to be highly significant.

The 95% confidence intervals on the estimated parameters are also reported in Table 7, assuming the asymptotic normality holds. Although acceptable results are obtained for the normal and lognormal factors, clearly those built on the spline coefficients cannot be trusted. An attempt is made to correct these intervals by estimating models constructed by using a bootstrap resampling of the populations [for a comprehensive coverage of bootstrap techniques, see Efron and Tibshirani (*24*)]. The bootstrap can indeed be used to construct approximate confidence intervals by using quartiles of the bootstrap. Because this approach requires many additional optimization runs, the number of replications has to be limited, and 40 bootstrap replications are performed, which is not enough to build a precise empirical cumulative distribution function. Therefore the extreme values among the bootstrap repetitions are simply taken to construct the new confidence intervals. Although the confidence intervals of the tails of the spline are large, reflecting the difficulty of capturing extreme behaviors, as few observations are then available, central intervals are much more reasonable and less conservative. Although the intervals overlap, lower and upper bounds are monotonically increasing, reflecting the spline conditions. To illustrate even more the influence of condition R1, Table 7 reports the robust standard deviation estimates and the standard deviation of the bootstrap parameters. For the normal and lognormal parameters, for which R1 holds, the standard deviation estimates are remarkably close, but when R1 is violated, that is, for the spline parameters, the

TABLE 6 Case Study 3: Experiment on Synthetic Nonparametric Data

Type	Estimator	*t*-Statistic	Normal CI	Bootstrap CI
Normal (μ)	0.032	0.38	(–0.131, 0.194)	(–0.178, 0.208)
Normal (σ)	2.023	7.73	(1.510, 2.536)	(1.492, 2.679)
Lognormal (μ)	0.052	0.48	(–0.159, 0.262)	(–0.246, 0.296)
Lognormal (σ)	0.992	4.59	(0.568, 1.416)	(0.320, 1.555)
Spline (q_1)	–17.632	2.54	(–31.259, –4.004)	(–35.266, –7.679)
Spline (q_2)	–1.050	0.28	(–8.378, 6.279)	(–5.234, –0.662)
Spline (q_3)	–1.050	0.52	(–4.950, 2.850)	(–2.255, –0.109)
Spline (q_4)	–1.050	1.16	(–2.817, 0.718)	(–1.474, 0.898)
Spline (q_5)	1.526	1.11	(–1.165, 4.216)	(–0.109, 2.082)
Spline (q_6)	1.526	0.67	(–2.942, 5.993)	(–0.898, 2.886)
Spline (q_7)	1.526	0.57	(–3.711, 6.763)	(0.898, 3.011)
Spline (q_8)	20.644	2.14	(1.774, 39.514)	(1.948, 133.698)

NOTE: CI = confidence interval.

TABLE 7 Case Study 3: Robust Versus Bootstrap Standard Deviation

Type	Estimator	Robust SD	Bootstrap SD
Normal (μ)	0.032	0.083	0.083
Normal (σ)	2.023	0.262	0.268
Lognormal (μ)	0.052	0.107	0.117
Lognormal (σ)	0.992	0.216	0.286
Spline (q_1)	−17.632	6.953	6.415
Spline (q_2)	−1.050	3.734	0.994
Spline (q_3)	−1.050	1.990	0.445
Spline (q_4)	−1.050	0.902	0.510
Spline (q_5)	1.526	1.373	0.491
Spline (q_6)	1.526	2.280	0.512
Spline (q_7)	1.526	2.672	0.574
Spline (q_8)	20.644	9.627	29.05

NOTE: SD = standard deviation.

usual standard deviations strongly differ from the bootstrap parameters, which are more reliable.

CONCLUSIONS

In this paper, the properties of the information matrix have been studied in the context of mixed logit estimation. The problem is relevant for transportation analysts given that the information matrix is at the heart of the derivation of asymptotic properties of the estimated parameters. From a theoretical perspective, the regularity conditions behind the information matrix equality have been reviewed, under which it is possible to calculate the information matrix as the opposite of the log likelihood Hessian evaluated at the estimates. An empirical analysis based on data with controlled characteristics has been conducted to demonstrate the importance of the correct use of the information matrix. It shows that for advanced models, especially mixed logit and models rooted in panel data, the information matrix equality can be violated, stressing that one should use the robust variance–covariance matrix estimates, not the opposite of the Hessian, to make inference. In addition, a large correlation between parameters estimated by using mixed logit has been calculated even though all the parameters were generated independently. The second case study, developed in the context of a stated choice experiment, suggests that *D*-optimal designs, focusing on individual parameter variance, tend to create correlations between the studied factors to decrease the total error. Finally, the case of mixed logit models specified with nonparametric coefficients has been studied, for which it is not possible to rely on the asymptotic analysis based on the information matrix. Under these conditions, the log likelihood ratio test has been proposed to assess the significance of parameters and bootstrap resampling techniques to calculate confidence intervals for them. This ratio test provides guidelines to researchers and practitioners to make better inference on advanced discrete choice models. However, more research remains necessary for complete understanding of the reasons behind information matrix violations.

ACKNOWLEDGMENT

The first author acknowledges the Natural Sciences and Engineering Research Council of Canada for partial support for this research.

REFERENCES

1. Ben-Akiva, M., and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press, Cambridge, Mass., 1985.
2. Train, K. *Discrete Choice Methods with Simulation.* Cambridge University Press, New York, 2003.
3. Fosgerau, M., and M. Bierlaire. A Practical Test for the Choice of Mixing Distribution in Discrete Choice Models. *Transportation Research Part B,* Vol. 41, No. 7, 2007, pp. 784–794.
4. Bastin, F., C. Cirillo, and Ph. L. Toint. Estimating Nonparametric Random Utility Models, with an Application to the Value of Time in Heterogeneous Populations. *Transportation Science,* Vol. 44, No. 4, 2010, pp. 537–549.
5. Revelt, D., and K. Train. Mixed Logit with Repeated Choices. *Review of Economics and Statistics: Households' Choices of Appliance Efficiency Effect,* Vol. 80, No. 4, 1998.
6. Train, K., and G. Sonnier. Mixed Logit with Bounded Distributions of Correlated Partworths. In *Applications of Simulations Methods in Environmental and Resource Economics* (A. Alberini and R. Scarpa, eds.), Kluwer Academics Publisher, Dordrecht, Netherlands, 2005, pp. 117–134.
7. Rose, J. M., and M. J. Bliemer. *Designing Efficient Data for Stated Choice Experiments.* Paper presented at 11th International Conference on Travel Behaviour Research, Kyoto, Japan, 2006.
8. Sándor, Z., and M. Wedel. Heterogeneous Conjoint Choice Designs. *Journal of Marketing Research,* Vol. 42, No. 2, 2005, pp. 210–218.
9. Bliemer, M. J., and J. M. Rose. Construction of Experimental Designs for Mixed Logit Models Allowing for Correlation Across Choice Observations. *Transportation Research Part B,* Vol. 44, 2010, pp. 720–734.
10. Devroye, L. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 1986.
11. Law, A. M. *Simulation Modeling & Analysis,* 4th ed. McGraw-Hill, Boston, Mass., 2007.
12. L'Ecuyer, P. Efficiency Improvement via Variance Reduction. *Proc., 1994 Winter Simulation Conference,* Orlando, Fla., 1994, pp. 122–132.
13. Newey, W. K., and D. McFadden. Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics,* Vol. IV, chap. 36 (R. Engle and D. McFadden, eds.), Elsevier, Amsterdam, Netherlands, 1994, pp. 2111–2245.
14. Spanos, A. *Econometric Modeling with Observational Data.* Cambridge University Press, Cambridge, United Kingdom, 1999.
15. Lehmann, E. L., and G. Casella. *Theory of Point Estimation,* 2nd ed. Springer-Verlag, New York, 1998.
16. Huber, P. J. The Behavior of Maximum Likelihood Estimates Under Nonstandard Condition. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (N. LeCam and J. Neyman, eds.), University of California Press, Berkeley, 1967.
17. White, H. Maximum Likelihood of Misspecified Models. *Econometrica,* Vol. 50, No. 1, 1982, pp. 1–25.
18. Munger, D., P. L'Ecuyer, F. Bastin, C. Cirillo, and B. Tuffin. Estimation Strategies for Complex Discrete Choice Models. *Transportation Research Part B* (in press).
19. Bastin, F., and C. Cirillo. Reducing Simulation Bias in Mixed Logit Model Estimation. *Journal of Choice Modelling,* Vol. 3, No. 2, 2010, pp. 71–88.
20. Bastin, F., C. Cirillo, and Ph. L. Toint. Convergence Theory for Nonconvex Stochastic Programming with an Application to Mixed Logit. *Mathematical Programming, Series B,* Vol. 108, No. 2/3, 2006, pp. 207–234.
21. Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman. Estimation and Inference in Nonlinear Structural Models. *Annals of Economic and Social Measurement,* Vol. 3/4, 1974, pp. 653–665.
22. Bastin, F., C. Cirillo, and Ph. L. Toint. Application of an Adaptive Monte Carlo Algorithm to Mixed Logit Estimation. *Transportation Research Part B,* Vol. 40, No. 7, 2006, pp. 577–593.
23. Bastin, F., C. Cirillo, and Ph. L. Toint. An Adaptive Monte Carlo Algorithm for Computing Mixed Logit Estimators. *Computational Management Science,* Vol. 3, No. 1, 2006, pp. 55–79.
24. Efron, B., and R. J. Tibshirani. *An Introduction to the Bootstrap.* No. 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Fla., 1993.

The Transportation Demand Forecasting Committee peer-reviewed this paper.