DRAFT

Specification, Identification, & Estimation of the Logit Kernel (or Continuous Mixed Logit) Model *

Moshe Ben-Akiva [‣], Denis Bolduc [▲], and Joan Walker [‣]

February 2001

Abstract

Logit kernel is a discrete choice model that has both probit-like disturbances as well as an additive i.i.d. extreme value (or Gumbel) disturbance à la multinomial logit. The result is an intuitive, practical, and powerful model that combines the flexibility of probit with the tractability of logit. For this reason, logit kernel has been deemed the "model of the future" and is becoming extremely popular in the literature. It has already been included in a recent edition of a widely used econometrics textbook.

While the basic structure of logit kernel models is well understood, there are important formulation and practical issues that are critical for estimation and yet are often overlooked. We aim to highlight some of these issues in the paper. One key point is that the addition of the Gumbel term is not necessarily innocuous, and thus the normalization required for logit kernel can be different than for an analogous pure probit model. Another point is that there are interesting and non-intuitive identification rules regarding nested structures and random coefficient models. Misunderstanding of these issues can lead to biased estimates as well as a significant loss of fit. A clear understanding of identification becomes even more critical given the fact that simulation, which is often used to estimate these models due to the high dimensionality of the integrals, has a tendency to cover up identification problems.

In the paper we present a general framework for specification, identification, and estimation of the logit kernel model. We specify the model using a general factor analytic error structure. We show that the factor analytic form includes all known (additive) error structures as special cases, including heteroscedasticity, error components, nesting structures, random coefficients, and auto correlation. We discuss in detail many of the special cases of the logit kernel model and highlight specification and identification issues related to each. Finally we demonstrate our findings with empirical examples using both simulated and real data. The objectives of the paper are to present our specific findings, as well as highlight the broader themes and provide tools for uncovering identification issues pertaining to logit kernel models.

---

* This work was partially supported by the Social Sciences and Humanities Research Council of Canada, Le Fonds FCAR, and a UPS fellowship. This paper is a major revision of the working paper by Ben-Akiva and Bolduc (1996), "Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure" based on recent work by Walker (2001).

[‣] Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, U.S.A.

[▲] Université Laval, Québec, Canada, G1K 7P4.

Introduction

The logit kernel model is a straightforward concept: it is a discrete choice model in which the disturbances (of the utilities) consist of both a probit-like portion and an additive i.i.d. Gumbel portion (i.e., a multinomial logit disturbance).

Multinomial logit (MNL) has its well-known blessing of tractability and its equally well-known curse of a rigid error structure leading to the IIA property. The nested logit model relaxes the rigidity of the MNL error structure and has the advantage of retaining a probability function in closed form. Nonetheless, nested logit is still limited and cannot capture many forms of unobserved heterogeneity, including, for example, random parameters. The logit kernel model with its probit-like disturbances completely opens up the specification of the disturbances so that almost any desirable error structure can be represented in the model. As with probit, however, this flexibility comes at a cost, namely that the probability functions consist of multi-dimensional integrals that do not have closed form solutions. Standard practice is to estimate such models by replacing the choice probabilities with easy to compute and unbiased simulators. The beauty of the additive i.i.d. Gumbel term is that it leads to a particularly convenient and attractive probability simulator, which is simply the average of a set of logit probabilities. The logit kernel probability simulator has all of the desirable properties of a simulator including being convenient, unbiased, and smooth.

Terminology

There are numerous terms floating around the literature that are related to the logit kernel model that we present here. McFadden and Train (2000) use the term "mixed logit" to refer to models that are comprised of a mixture of logit models. This is a broad class that encompasses any type of mixing distribution, including discrete distributions (for example, latent class) as well as continuous distributions. Within this reference, logit kernel is a special case of mixed logit in which the mixing distribution is continuous. There are also numerous terms that are used to describe various error specifications in discrete choice models, including error components, taste variation, random parameters (coefficients), random effects, unobserved heterogeneity, etc. When such models are specified in a form that includes an additive i.i.d. Gumbel term, then they fall within the logit kernel (as well as mixed logit) class of models. Many of these special cases are described later in the paper.

We choose to use the term *logit kernel*, because conceptually these models start with a logit model at the core and then are extended by adding a host of different error terms. In addition, the term is descriptive of the form of the likelihood function and the resulting logit kernel simulator.

Organization of the Paper

The paper is organized as follows. First, we introduce the logit kernel model and present a general discussion of identification. Then we discuss specification and identification of several important special cases, which are all based on a factor analytic representation of the error covariance structure. Next, we

focus on the estimation of logit kernel via maximum (simulated) likelihood. In the final section, we present empirical results that highlight some of the specification and identification issues.

Related Literature

There have been many previous efforts to extend the logit model to allow more flexible covariance structures. The most widely used extension is nested logit. The advantage of nested logit is that it relaxes the classic IIA assumption and yet has a closed form. Nonetheless it is still a fairly rigid model. Nested logit is not a logit kernel model, although it can be approximated in the logit kernel structure. In terms of logit kernel models, the earliest applications were in random parameter logit specifications, which appeared 20 years ago in the papers by Boyd and Mellman (1980) and Cardell and Dunbar (1980). The more general form of the model came about through researchers quest for smooth probability simulators for use in estimating probit models. McFadden's 1989 paper on the Method of Simulated Moments, includes a description of numerous smooth simulators, one of which involved probit with an additive i.i.d. Gumbel term. Stern (1992) described a similar simulator, which has an additive i.i.d. normal term instead of the Gumbel. At the time of these papers, there was a strong desire to retain the pure probit form of the model. Hence, the algorithms and specifications were designed to eventually remove the additive "contamination" element from the model (for example, McFadden, 1989) or ensure that it did not interfere with the pure probit specification (for example, Stern, 1992). Bolduc and Ben-Akiva (1991)[1] did not see the need to remove the added noise, and began experimenting with models that left the Gumbel term in tact, and found that the models performed well. There have been numerous relatively recent applications and investigations into the model, including Bhat (1997 & 1998), Bolduc, Fortin and Fournier (1996), Brownstone, Bunch and Train (2000), Brownstone and Train (1999), Goett, Hudson, and Train (2000), Gönül and Srinivasan (1993), Greene (2000), Mehndiratta and Hansen (1997), Revelt and Train (1998 & 1999), Srinivasan and Mahmassani (2000), and Train (1998). A very important recent contribution is McFadden and Train's (2000) paper on mixed logit, which both (i) proves that any well-behaved random utility consistent behavior can be represented as closely as desired with a mixed logit specification, and (ii) presents easy to implement specification tests for these models.

While logit kernel has strong computational advantages, it, like probit, does not have a closed form solution and can easily lead to high dimensional integrals. The well-known Gaussian Quadrature method of numerical integration is not computationally feasible for dimensionalities above 3 or so, and therefore estimation via simulation is a key aspect to applications of the logit kernel model. The basic idea behind simulation is to replace the multifold integral (the probability equations) with easy to compute probability simulators. Lerman and Manski (1981) introduced this concept and proposed the use of a frequency simulator to simulate probit probabilities. The frequency simulator was found to have poor computational properties primarily because it is not smooth (i.e., not continuous and not differentiable). Basically the frequency simulator maps each draw to a value of either 0 or 1, whereas a smooth simulator would map each draw to a value somewhere between 0 and 1 (and therefore retains more information). The result is

[1] Later generalized to Ben-Akiva and Bolduc (1996).

that discontinuous simulators require a prohibitively large number of simulation draws to obtain acceptable accuracy. In addition, a theoretical advantage of smoothness is that it greatly simplifies asymptotic theory. For these reasons, there has been a lot of research on various smooth simulators (see, for example, Börsch-Supan and Hajivassiliou, 1993; McFadden, 1989; Pakes and Pollard, 1989; and Stern, 1992). The discovery of the GHK simulator provided a smooth simulator for probit, which quickly became the standard for estimating probit models (see Hajivassiliou and Ruud, 1994). Now there is great interest in the logit kernel smooth simulator because it is conceptually intuitive, flexible, and relatively easy to program.

With simulation, the types and number of draws that are made from the underlying distribution to calculate the simulated probabilities are always important issues. Traditionally, simple pseudo-random draws (for example, Monte Carlo) have been used. Bhat (2000) and Train (1999) present an interesting addition to the econometric simulation literature, which is the use of intelligent drawing mechanisms (in many cases non-random draws known as Halton sequences). These draws are designed to cover the integration space in a more uniform way, and therefore can significantly reduce the number of draws required. We employ this approach for the empirical results presented later in this paper.

A final point is that we use Maximum Likelihood Estimation (ML) or Maximum Simulated Likelihood (MSL). An alternative to this is the Method of Simulated Moments (MSM) proposed by McFadden (1989) and Pakes and Pollard (1989). MSM is often favored over MSL because a given level of accuracy in model parameter estimation can be obtained with a fairly small number of replication draws. The accuracy of the MSL methodology critically depends on using a large number of simulation draws because the log-likelihood function is simulated with a non-negligible downward bias. For several reasons, we still stick to the MSL approach. First, MSL requires the computation of the probability of only the chosen alternative, while MSM needs all choice probabilities. With large choice sets this factor can be quite important. Second, the objective function associated with MSL is numerically better behaved than the MSM objective function. Third, with the increase in computational power and the implementation of intelligent drawing mechanisms, the number of draws issue is not as critical as it once was.

The Logit Kernel Model

The Discrete Choice Model

Consider the following discrete choice model. For a given individual n, $n = 1,...,N$ where N is the sample size, and an alternative i, $i = 1,...,J_n$ where J_n is the number of alternatives in the choice set C_n of individual n, the model is written as:

$$y_{in} = \begin{cases} 1 & \text{if } U_{in} \geq U_{jn}, \text{ for } j = 1,...,J_n \\ 0 & \text{otherwise} \end{cases},$$

$$U_{in} = X_{in} b + e_{in},$$

where y_{in} indicates the observed choice, and U_{in} is the utility of alternative i as perceived by individual n. X_{in} is a $(1\times K)$ vector of explanatory variables describing individual n and alternative i, including alternative-specific dummy variables as well as generic and alternative-specific attributes and their interactions with the characteristics of individual n. b is a $(K\times 1)$ vector of coefficients and e_{in} is a random disturbance. The assumption that the disturbances are i.i.d. Gumbel leads to the tractable, yet restrictive logit model. The assumption that the disturbances are multivariate normal distributed leads to the flexible, but computationally demanding probit model. The logit kernel model presented in this paper is a hybrid between logit and probit and represents an effort to incorporate the advantages of each.

In a more compact vector form, the discrete choice model can be written as follows:

$$y_n = [y_{1n},...,y_{J_n n}]' ,$$

$$U_n = X_n b + e_n , \qquad (1)$$

where y_n, U_n, and e_n are $(J_n\times 1)$ vectors and X_n is a $(J_n\times K)$ matrix.

The Logit Kernel Model with Factor Analytic Form

Model Specification

In the logit kernel model, the e_{in} random utility term is made up of two components: a probit-like component with a multivariate distribution, and an i.i.d. Gumbel random variate. The probit-like term captures the interdependencies among the alternatives. We specify these interdependencies using a factor analytic structure. The factor analytic structure was first proposed for probit by McFadden (1984) as a means of reducing the dimensionality of the integral. We use if here because it is a flexible specification that includes all known (additive) error structures as special cases, as we will show below.

Using the factor analytic form, the disturbance vector e_n is specified as follows:

$$e_n = F_n x_n + n_n , \qquad (2)$$

where x_n is an $(M\times 1)$ vector of M multivariate distributed latent factors, F_n is a $(J_n\times M)$ matrix of the factor loadings that map the factors to the error vector (F_n includes fixed and/or unknown parameters and may also be a function of covariates), and v_n is a $(J_n\times 1)$ vector of i.i.d. Gumbel random variates. For estimation, it is desirable to specify the factors such that they are independent, and we therefore decompose x_n as follows:

$$x_n = Tz_n , \qquad (3)$$

where z_n are a set of standard independent factors (often normally distributed), TT' is the covariance matrix of x_n, and T is the Cholesky factorization of it. The number of factors, M, can be less than, equal to, or greater than the number of alternatives. To simplify the presentation, we assume that the factors

have standard normal distributions, however, they can follow any number of different distributions, such as lognormal, uniform, etc.

Substituting Equations (2) and (3) into Equation (1), yields:

The Factor Analytic Logit Kernel Specification

$$U_n = X_n b + F_n T z_n + n_n \quad , \tag{4}$$

$$\text{cov}(U_n) = F_n T T' F_n' + (g/m^2) I_{J_n} \tag{5}$$

(which we denote as $\Omega_n = \Sigma_n + \Gamma_n$),

where: U_n is a $(J_n \times 1)$ vector of utilities;

X_n is a $(J_n \times K)$ matrix of explanatory variables;

b is a $(K \times 1)$ vector of unknown parameters;

F_n is a $(J_n \times M)$ matrix of factor loadings, including fixed and/or unknown parameters;

T is a $(M \times M)$ lower triangular matrix of unknown parameters, where $TT' = Cov(x_n = Tz_n)$;

z_n is a $(M \times 1)$ vector of i.i.d. random variables with zero mean and unit variance; and

n_n is a $(J_n \times 1)$ vector of i.i.d. Gumbel random variables with zero location parameter and scale equal to $m > 0$. The variance is g/m^2, where g is the variance of a standard Gumbel ($p^2/6$).

The unknown parameters in this model are m, b, those in F_n, and those in T. X_n are observed, whereas z_n and n_n are unobserved.

It is important to note that we specify the model in level form (i.e., U_{jn}, $j = 1,...,J_n$) rather than in difference form (i.e., $(U_{jn} - U_{J_n n})$, $j = 1,...,(J_n - 1)$). We do this for interpretation purposes, because it enables us to parameterize the covariance structure in ways that capture specific (and conceptual) correlation effects. Nonetheless, it is the difference form that is estimable, and there are multiple level structures that can represent any unique difference covariance structure. We return to this issue later in the paper.

Response Probabilities

As will become apparent later, a key aspect of the logit kernel model is that if the factors z_n are known, the model corresponds to a multinomial logit formulation:

$$\Lambda(i \mid z_n) = \frac{e^{m\left(X_{in}b + F_{in}Tz_n\right)}}{\sum_{j \in C_n} e^{m\left(X_{jn}b + F_{jn}Tz_n\right)}} \ , \tag{6}$$

where $\Lambda(i \mid z_n)$ is the probability that the choice is i given z_n, and F_{jn} is j^{th} row of the matrix F_n, $j = 1,..., J_n$.

Since the z_n is in fact not known, the unconditional choice probability of interest is:

$$P(i) = \int_z \Lambda(i \mid z)n(z, I_M)dz \ , \tag{7}$$

where $n(z, I_M)$ is the joint density function of z, which, by construction, is a product of standard univariate normals:

$$n(z, I_M) = \prod_{m=1}^{M} f(z_m) \ .$$

The advantage of the logit kernel model is that we can naturally estimate $P(i)$ with an unbiased, smooth, tractable simulator, which we compute as:

$$\hat{P}(i) = \frac{1}{\mathbb{D}} \sum_{d=1}^{\mathbb{D}} \Lambda(i \mid z_n^d) \ ,$$

where z_n^d denotes draw d from the distribution of z, thus enabling us to estimate high dimensional integrals with relative ease.

Finally, note that if $T = 0$ then the model reduces to logit.

Identification and Normalization

It is not surprising that the estimation of such models raises identification and normalization issues. There are two sets of relevant parameters that need to be considered: the vector b and the unrestricted parameters of the distribution of the disturbance vector e_n, which include F_n, T, and m. For the vector b, identification is identical to that for a multinomial logit model. Such issues are well understood, and the reader is referred to Ben-Akiva and Lerman (1985) for details.

The identification of the parameters in error structure is more complex, and will be discussed in detail in this paper.

Comments on Identification of Pure Probit versus Logit Kernel

Recall that the error structure of the logit kernel model consists of a probit-like component and an additive i.i.d. extreme value term (the Gumbel). Bolduc (1992), Bunch (1991), Dansie (1985) and others address identification issues for disturbance parameters in the multinomial probit model. Bunch (1991) presents clear guidelines for identification (consisting of Order and Rank conditions, which are described below) and provides examples of identified and unidentified error structures. He also provides a good literature review of the investigations into probit identification issues. For the most part, the identification guidelines for pure probit are applicable to the probit-like component of the logit kernel model. However, there are some differences, which are touched on here, and will be expanded on in the detailed discussion that follows.

We will see below that by applying the mechanics that are used to determine identification of a Probit model (Order and Rank) to the logit kernel model, effectively what happens is that the number of identifying restrictions that were necessary for a pure probit model are also required for the probit-like portion of the logit kernel model. However, there are some subtle, yet important, differences. Recall that one constraint is always necessary to set the scale of the model. In a pure probit model, this is done by setting at least one of the elements of the covariance structure[2] to some positive value (usually 1). Call this element that is constrained s_p. With logit kernel, on the other hand, the scale of the model is set as in a standard logit model by constraining the m parameter of the i.i.d. Gumbel term. Since the scale of the logit kernel model is set by m, the normalization of s_p is now a regular identifying restriction in the logit kernel model. One issue with the normalization of s_p for the logit kernel model is that in order to be able to trivially test the hypothesis that a logit kernel model is statistically different from a pure logit model, it is desirable to set s_p equal to zero so that pure logit is a special case of a logit kernel specification. A second difference is that while the specific element of the covariance matrix that is used to set the scale in a probit model is arbitrary, the selection of s_p is not necessarily arbitrary in the equivalent logit kernel model. This is due to the structure of the logit kernel model, and will be explained further below (in the discussion of the 'positive definiteness' condition.)

Finally, it turns out that the fact that s_p must be constrained in a logit kernel model is not exactly correct. In a *probit* kernel model (i.e., with an i.i.d. normal term), it is true that s_p must be constrained. In this case, there is a perfect trade-off between the multivariate normal term and the i.i.d. normal term. However, in the logit kernel model, this perfect trade-off does not exist because of the slight difference between the Gumbel and Normal distributions. Therefore, there will be an optimal combination of the Gumbel and Normal distribution, and this effectively allows another parameter to be estimated. This leads to somewhat surprising results. For example, in a heteroscedastic logit kernel model a variance term can be estimated for *each* of the alternatives, whereas probit, probit kernel, or extreme value logit requires that one of the variances be constrained. The same holds true for an unrestricted covariance structure. Nonetheless, the reality is that without the constraint, the model is nearly singular (i.e., the objective function is very flat at the optimum), as will be demonstrated in the estimation results that follow. Due to

[2] Technically, the constraint is on the covariance matrix of utility *differences*.

the near singularity, it is advisable to impose the additional constraint, and we proceed using this approach throughout the rest of the discussion.

Overview of Identification

The first step of identification is to determine the model of interest, that is, the disturbance structure that is a priori assumed to exist. For example, an unrestricted covariance matrix (of utility differences) or various restricted covariance matrices such as heteroscedasticity or nesting. Once that is determined, there are three steps to determining the identification and normalization of the hypothesized model. The first two have to do with identification. For the model to be identified, both the order condition (necessary) and the rank condition (sufficient) must hold. The order condition establishes the maximum number of parameters that can be estimated, which is based on the number of alternatives in the choice set. The rank condition establishes the actual number of parameters that can be estimated, which is based on the number of independent equations available. In cases in which the conclusion from the order and rank conditions is that additional restrictions are in order, then a third condition (which we refer to as the positive definiteness condition) is necessary to verify that the chosen normalization is valid. Recall that the reason that an identifying restriction is necessary is that there are an infinite number of solutions (i.e., parameter estimates) to match the given model structure. The point of an identifying restriction is to establish the existence of a single unique solution, but not change the underlying model in any way. The positive definiteness condition asks the question of whether the models true structure (i.e., the one on which the rank and order conditions were applied) is maintained given the chosen identifying restriction. This is not an important issue for probit, but, as we will see, it has important implications for logit kernel. Each of the conditions is expanded on below, and we use the heteroscedastic logit kernel model to illustrate each condition.

The Specification of the Heteroscedastic Logit Kernel Model

The heteroscedastic model, assuming a universal choice set ($C_n = C \ \ \forall n$), is written as:[3]

Vector notation: $U_n = X_n b + T z_n + n_n$, ($M = J$ and F_n equals the identity matrix I_J),

$$T = \begin{bmatrix} s_1 & & & \\ 0 & s_2 & & \\ 0 & 0 & \ddots & \\ 0 & 0 & 0 & s_J \end{bmatrix} \ (J \times J), \ z_n \ (J \times 1),$$

and, defining $s_{ii} = (s_i)^2$, the $Cov(U_n)$ is:

[3] Note that our notation for symmetric matrices is to show only the lower triangular portion.

$$\Omega = \begin{bmatrix} s_{11} + g/m^2 & & & \\ 0 & s_{22} + g/m^2 & & \\ 0 & 0 & \ddots & \\ 0 & 0 & 0 & s_{JJ} + g/m^2 \end{bmatrix} (J \times J).$$

Scalar notation: $U_{in} = X_{in}b + s_i z_{in} + n_{in}$, $i \in C$.

Note that for a heteroscedastic model with a universal choice set, the covariance matrix does not vary across the sample, and so we can drop the subscript n from Ω_n.

We carry the identification conditions through for a binary heteroscedastic model, a three alternative heteroscedastic model, and a four alternative heteroscedastic model, because the three models serve well to highlight various aspects of identification and normalization. The covariance structures for these three models are as follows:

$$J = 2: \ \Omega = \begin{bmatrix} s_{11} + g/m^2 & \\ 0 & s_{22} + g/m^2 \end{bmatrix},$$

$$J = 3: \ \Omega = \begin{bmatrix} s_{11} + g/m^2 & & \\ 0 & s_{22} + g/m^2 & \\ 0 & 0 & s_{33} + g/m^2 \end{bmatrix},$$

$$J = 4: \ \Omega = \begin{bmatrix} s_{11} + g/m^2 & & & \\ 0 & s_{22} + g/m^2 & & \\ 0 & 0 & s_{33} + g/m^2 & \\ 0 & 0 & 0 & s_{44} + g/m^2 \end{bmatrix}.$$

Setting the Location

The general approach to identification of the error structure is to examine the covariance matrix of utility differences, denoted in the general case as Ω_{n,Δ_j}. Taking the differences sets the "location" of the model, a necessity for random utility models. The covariance matrix of utility differences for any individual is:

$$\Omega_{n,\Delta_j} = Cov(\Delta_j U_n) = \Delta_j F_n T T' F_n' \Delta_j' + \Delta_j (g/m^2) I_J \Delta_j',$$

where Δ_j is the linear operator that transforms the J utilities into $(J-1)$ utility differences taken with respect to the j^{th} alternative. Δ_j is a $(J-1) \times J$ matrix that consists of a $(J-1) \times (J-1)$ identity matrix with a column vector of -1's inserted as the j^{th} column. We use the notation $\Omega_{n,\Delta}$ to denote the covariance matrix of utility differences taken with respect to the J^{th} alternative.

Setting the Location for the Heteroscedastic Model

For the example heteroscedastic models using J as the base, the covariance matrices of utility differences are as follows:

$$J=2:\ \Delta_J=\begin{bmatrix}1 & -1\end{bmatrix},\qquad \Omega_\Delta=\begin{bmatrix}s_{11}+s_{22}+2g/m^2\end{bmatrix},$$

$$J=3:\ \Delta_J=\begin{bmatrix}1 & 0 & -1\\ 0 & 1 & -1\end{bmatrix},\qquad \Omega_\Delta=\begin{bmatrix}s_{11}+s_{33}+2g/m^2 & \\ s_{33}+g/m^2 & s_{22}+s_{33}+2g/m^2\end{bmatrix},$$

$$J=4:\ \Delta_J=\begin{bmatrix}1 & 0 & 0 & -1\\ 0 & 1 & 0 & -1\\ 0 & 0 & 1 & -1\end{bmatrix},$$

$$\Omega_\Delta=\begin{bmatrix}s_{11}+s_{44}+2g/m^2 & & \\ s_{44}+g/m^2 & s_{22}+s_{44}+2g/m^2 & \\ s_{44}+g/m^2 & s_{44}+g/m^2 & s_{33}+s_{44}+2g/m^2\end{bmatrix}.$$

Order Condition

The first condition is the order condition, which is necessary for identification. When discussing the Order Condition, it is useful to separate the covariance matrix into that which is constant across the sample (called the 'alternative-specific' portion) and that which varies across the sample (for example, in the case of random parameters). The order condition only applies to the alternative-specific portion of the covariance matrix. It states that a maximum of $s=J(J-1)/2-1$ alternative-specific parameters are estimable in Ω, which is equal to the number of distinct cells in Ω_Δ (symmetric) minus 1 to set the scale (another necessity of random utility models). Therefore:

with 2 alternatives, no alternative-specific covariance terms can be identified;
with 3 alternatives, up to 2 terms can be identified;
with 4 alternatives, up to 5 terms can be identified;
with 5 alternatives, up to 9 terms can be identified;
etc.

When the error structure has parameters that are not alternative-specific, for example, random parameters, it is possible to estimate more than s parameters, because there is additional information derived from the variations of the covariance matrix across individuals. Technically, there still is an order condition, but the limit is large (related to the size of the sample) and is therefore never a limiting condition.

The Order Condition and the Heteroscedastic Model

The disturbance parameters in the heteroscedastic model are alternative-specific, so the order condition must hold. Each heteroscedastic model has $J+1$ unknown parameters: J s_{ii}'s and one m. The order condition then provides the following information regarding identification:

$J = 2$: $unknowns = \{s_{11}, s_{22}, m\}$; $s = 0$ → 0 variances are identified

$J = 3$: $unknowns = \{s_{11}, s_{22}, s_{33}, m\}$; $s = 2$ → up to 2 variances are identified

$J = 4$: $unknowns = \{s_{11}, s_{22}, s_{33}, s_{44}, m\}$; $s = 5$ → potentially *all* variances are identified

Note that there are published probit and logit kernel models in the literature that do not meet the order condition, see, for example, Greene (2000) Table 19.15 and Louviere et al. (2000) Table B.6. While the logit kernel models in Greene and Louviere do not meet the order condition, these models are nonetheless barely identified due to the slight difference between the normal and Gumbel distributions (as discussed earlier). However, the probit model does not have this luxury, and therefore the probit model reported in Greene is not identified (as will be demonstrated in the mode choice application).

While the order condition provides a quick check for identification, it is clearly shown in Bunch (1991) that the number of parameters that can be estimated is often less than s, depending on the covariance structure postulated. Therefore, the rank condition must also be checked, which is described next.

Rank Condition

The rank condition is more restrictive than the order condition, and it is a sufficient condition for identification. The order condition simply counts cells, and ignores the internal structure of Ω_Δ. The rank condition, however, counts the number of linearly independent equations available in Ω_Δ that can be used to estimate the parameters of the error structure. Bolduc (1992) and Bunch (1991) describe the mechanics of programming the rank condition. The basic idea behind determining this count is to examine the Jacobian matrix, which is equal to the derivatives of the elements in Ω_Δ with respect to the unknown parameters. The number of parameters that can be estimated is equal to the Rank of the Jacobian matrix minus 1 (to set the scale). These mechanics are demonstrated below with the heteroscedastic example.

The Rank Condition and the Heteroscedastic Model

The first step is to vectorize the unique elements of Ω_Δ into a column vector (we call this operator *vecu*):[4]

$$J = 3: \; vecu(\Omega_\Delta) = \begin{bmatrix} s_{11} + s_{33} + 2g/m^2 \\ s_{22} + s_{33} + 2g/m^2 \\ s_{33} + g/m^2 \end{bmatrix},$$

[4] Note that there's no need to continue with identification for the binary heteroscedastic case, since the order condition resolved that none of the error parameters are identified.

$$J=4:\ vecu(\Omega_{\Delta}) = \begin{bmatrix} s_{11}+s_{44}+2g/m^2 \\ s_{22}+s_{44}+2g/m^2 \\ s_{33}+s_{44}+2g/m^2 \\ s_{44}+g/m^2 \end{bmatrix}.$$

By examination, it is clear that we are short an equation in both cases. This is formally determined by examining the Rank of the Jacobian matrix of vecu(Ω_{Δ}) with respect to each of the unknown parameters $(s_{11},...,s_{JJ},g/m^2)$:

$$J=3:\ \begin{matrix} Jacobian \\ matrix\ of \\ vecu(\Omega_{\Delta}) \end{matrix} = \begin{bmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix},\ Rank=3 \quad \rightarrow \quad \begin{matrix} can\ estimate\ 2\ of\ the\ parameters; \\ must\ normalize\ m\ and\ one\ s_{ii}. \end{matrix}$$

$$J=4:\ \begin{matrix} Jacobian \\ matrix\ of \\ vecu(\Omega_{\Delta}) \end{matrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix},\ Rank=4 \ \rightarrow \ \begin{matrix} can\ estimate\ 3\ of\ the\ parameters; \\ must\ normalize\ m\ and\ one\ s_{ii}. \end{matrix}$$

So for both of these cases, the scale term m as well as one of the s_{ii}'s must be normalized.

Which s_{ii} should be fixed? And to what value? This is where the positive definiteness condition comes into play, and it turns out that the normalizations for logit kernel models are not always arbitrary or intuitive.

Positive Definiteness

When the conclusion from the order and rank conditions is that further identifying restrictions (normalizations) are required, the positive definiteness condition is used to determine the set of acceptable normalizations. Conceptually, the need for the positive definiteness condition is as follows. First note that the reason for the additional normalization is that there are infinite possible solutions that result in the hypothesized covariance structure. The normalization is necessary to establish the existence of a unique solution, but it does not change the underlying model structure (i.e., the covariance matrix of utility differences) in any way. The positive definiteness condition is necessary to verify that the chosen normalization is valid, i.e., that the remaining parameters that are estimated are able to replicate the underlying model structure. It turns out that with logit kernel models, there can be seemingly obvious normalizations that are not valid, because the structure of the model prevents the underlying covariance matrix of utility differences from being recovered.

To work through the details of the positive definiteness condition, we rephrase the above discussion as follows. There are two overriding issues behind the positive definiteness condition:

Statement 1: There are infinite possible normalizations that can be imposed to identify the model. However, note that all valid normalizations for a particular specification will result in identical $W_{n,D}$, that is, $\{W_{n,D}^{N1}$ from normalization 1$\}$ = $\{W_{n,D}^{N2}$ from normalization 2$\}$. For example, with this relationship, one can convert the estimated parameters from a particular normalization (say $s_{11}=0$) to the parameters that will be estimated if a different normalization (say $s_{11}=1$) is imposed (as long as both normalizations are valid).

Statement 2: The logit kernel covariance matrix is $\Omega_n = \Sigma_n + \Gamma_n$, where $\Sigma_n = (F_n T)(F_n T)'$ (Equation (5)). Therefore, by construction, Σ_n is necessarily positive semi-definite ('semi' because $F_n T$ can equal zero).

Given these two issues, any valid normalization must be such that both of the following conditions hold for all observations:

I. $\Omega_{n,\Delta}^{N} = \Omega_{n,\Delta} \quad \rightarrow \quad \Sigma_{n,\Delta}^{N} + \Gamma_{n,\Delta}^{N} = \Omega_{n,\Delta}$ (by definition of a normalization).

The covariance matrix of utility differences of the normalized model (denoted by N) equals the covariance matrix of utility differences of the non-normalized (theoretical) model.

II. Σ_n^N *is positive semi-definite* (by construction).

If the normalization is such that both Conditions I and II cannot be met, the parameter estimates will be inconsistent and result in a loss of fit. It turns out that for logit kernel, these conditions can impose restrictions on the feasible set of normalizations, as we describe below.

We have already stated that Condition II necessarily holds due to the construction of the model. Therefore, the issue is whether the imposed normalization is such that Condition I can be met, given the restriction that Σ_n^N is positive semi-definite. Problems can arise with logit kernel models due to the additive i.i.d. Gumbel portion of the covariance structure, Γ_n. Because of Γ_n, there can be normalizations for which satisfying Condition I requires a negative definite Σ_n^N. However, this conflicts with Condition II, and so any such normalization is not valid. Note that this issue actually arises with any model structure that includes an i.i.d. disturbance term along with a parameterized disturbance, for example, a probit kernel model.

Positive Definiteness and the Heteroscedastic Model

Looking at the heteroscedastic case, we will use the three alternative model as an example. It is useful in the analysis to deal directly with the estimated (i.e., scaled) parameters, so we introduce the notation $\dot{s}_{ii} = (ms_i)^2$. Say we impose the normalization that the third heteroscedastic term, $\dot{s}_{33}$, is constrained to some fixed value we denote as $\dot{s}_{ff}^{N}$. Condition I can then be written as:

$$\begin{bmatrix} (\dot{s}_{11}^{N} + \dot{s}_{ff}^{N} + 2g)/m_N^2 & \\ (\dot{s}_{ff}^{N} + g)/m_N^2 & (\dot{s}_{22}^{N} + \dot{s}_{ff}^{N} + 2g)/m_N^2 \end{bmatrix} = \begin{bmatrix} (\dot{s}_{11} + \dot{s}_{33} + 2g)/m^2 & \\ (\dot{s}_{33} + g)/m^2 & (\dot{s}_{22} + \dot{s}_{33} + 2g)/m^2 \end{bmatrix},$$

where the matrix on the left represents the normalized model $\left(\dot{s}_{ii}^{N} = (m_N \dot{s}_i^{N})^2\right)$ and the matrix on the right represents the theoretical (non-normalized) model. This relationship states that when the normalization is imposed, the remaining parameters in the normalized model will adjust such that the theoretical (or true) covariance matrix of utility differences is recovered. It also provides us with three equations:

$$(\dot{s}_{ff}^{N} + g)/m_N^2 = (\dot{s}_{33} + g)/m^2 \ , \tag{8}$$

$$(\dot{s}_{11}^{N} + \dot{s}_{ff}^{N} + 2g)/m_N^2 = (\dot{s}_{11} + \dot{s}_{33} + 2g)/m^2 \ , \text{ and} \tag{9}$$

$$(\dot{s}_{22}^{N} + \dot{s}_{ff}^{N} + 2g)/m_N^2 = (\dot{s}_{22} + \dot{s}_{33} + 2g)/m^2 \ . \tag{10}$$

Condition II states that Σ^N must be positive semi-definite, where:

$$\Sigma^N = \begin{bmatrix} \dot{s}_{11}^{N} & & \\ 0 & \dot{s}_{22}^{N} & \\ 0 & 0 & \dot{s}_{ff}^{N} \end{bmatrix} * \frac{1}{m_N^2} \ .$$

This matrix is positive semi-definite if and only if the diagonal entries are non-negative and m_N^2 is strictly positive, or:

$$m_N^2 > 0 \ , \tag{11}$$

$$\dot{s}_{11}^{N} \geq 0 \ , \tag{12}$$

$$\dot{s}_{22}^{N} \geq 0 \ , \text{ and} \tag{13}$$

$$\dot{s}_{ff}^{N} \geq 0 \ . \tag{14}$$

The positive definiteness condition requires that all valid normalizations satisfy the restrictions stated by Equations (8) to (14). The question is, what values of $\dot{s}_{ff}^{N}$ guarantee that these relationships hold?

To derive the restrictions on $\dot{s}_{ff}^{N}$, we first use Condition I (Equations (8) to (10)) to develop equations for the unknown parameters of the normalized model $(m_N^2, \ \dot{s}_{11}^{N}, \ and \ \dot{s}_{22}^{N})$ as functions of the normalized parameter $\dot{s}_{ff}^{N}$ and the theoretical parameters $(m^2, \ \dot{s}_{11}, \ \dot{s}_{22}, \ and \ \dot{s}_{33})$, which leads to:

$$m_N^2 = m^2(\dot{s}_{ff}^{N} + g)/(\dot{s}_{33} + g) \ , \tag{15}$$

$$\dot{s}_{11}^{N} = \left((\dot{s}_{11} + g)\dot{s}_{ff}^{N} + (\dot{s}_{11} - \dot{s}_{33})g\right)/\left(\dot{s}_{33} + g\right) \ , \text{ and} \tag{16}$$

$$\dot{s}_{22}^{N} = \left((\dot{s}_{22} + g)\dot{s}_{ff}^{N} + (\dot{s}_{22} - \dot{s}_{33})g\right)/\left(\dot{s}_{33} + g\right) \ . \tag{17}$$

Equations (11) to (14) impose restrictions on the parameters of the normalized model, and so we can combine them with Equations (15) to (17), which results in the following set of restrictions:

$$\dot{s}^{N}_{ff} \geq 0 \ , \qquad \text{(Eq. (14))} \qquad (18)$$

$$m^2(\dot{s}^{N}_{ff} + g)/(\dot{s}_{33} + g) > 0 \ , \qquad \text{(Eqs. (11) \& (15))} \qquad (19)$$

$$\left((\dot{s}_{11} + g)\dot{s}^{N}_{ff} + (\dot{s}_{11} - \dot{s}_{33})g\right)/\left(\dot{s}_{33} + g\right) \geq 0 \ , \text{ and} \qquad \text{(Eqs. (12) \& (16))} \qquad (20)$$

$$\left((\dot{s}_{22} + g)\dot{s}^{N}_{ff} + (\dot{s}_{22} - \dot{s}_{33})g\right)/\left(\dot{s}_{33} + g\right) \geq 0 \ . \qquad \text{(Eqs. (13) \& (17))} \qquad (21)$$

The other information we have is that Σ is positive semi-definite (by construction), and therefore:

$$m^2 > 0 \ , \ \dot{s}_{11} \geq 0 \ , \ \dot{s}_{22} \geq 0 \ , \text{ and } \ \dot{s}_{33} \geq 0 \ . \qquad (22)$$

So going back to restrictions (18)-(21), the first two restrictions are trivial: Equation (18) just states that the normalization has to be non-negative; and given Equations (18) and (22), Equation (19) will always be satisfied. Equations (20) and (21) are where it gets interesting, because solving for $\dot{s}^{N}_{ff}$ leads to the following restrictions on the normalization:

$$\dot{s}^{N}_{ff} \geq \left(\dot{s}_{33} - \dot{s}_{ii}\right) g/(g + \dot{s}_{ii}) \quad , \ i = 1,2 \ . \qquad (23)$$

($\dot{s}_{33}$ is the heteroscedastic term that is fixed.)

What does this mean? Note that as long as alternative 3 is the minimum variance alternative, the right hand side of Equation (23) is negative, and so the restriction is satisfied for any $\dot{s}^{N}_{ff} \geq 0$. However, when alternative 3 is not the minimum variance alternative, $\dot{s}^{N}_{ff}$ must be set "large enough" (and certainly above zero) such that Equation (23) is satisfied. This latter approach to normalization is not particularly practical since the $\dot{s}_{ii}$ are unknown (how large is large enough?), and it has the drawback that MNL is not a case nested within the logit kernel specification. Therefore, the following normalization is recommended:

The preferred normalization for the heteroscedastic logit kernel model is to constrain the heteroscedastic term of the <u>minimum</u> variance alternative to zero.

A method for implementing this normalization is described later in the section on heteroscedastic logit kernel models.

<u>Positive Definiteness and a Probit Model</u>

What about the positive definiteness condition for pure probit? Pure probit models also must satisfy a positive definiteness condition, but it turns out that these do not impose any problematic restrictions on the normalization. With pure probit, there is obviously no Gumbel term, so Condition I can be written as $\Sigma^{N}_{n,\Delta} = \Sigma_{n,\Delta}$. Condition II is similar to that for logit kernel, except that Σ^{N}_{n} must now be positive definite

(since it cannot equal zero). Since $\Sigma_{n,\Delta}$ is well-behaved (by construction), Condition I states that $\Sigma_{n,\Delta}^{N}$ will also be well-behaved, and, therefore, so will Σ_{n}^{N}. The result is that the positive definiteness condition automatically holds for normalizations that are intuitively applied to probit.

Positive Definiteness and a *Probit* Heteroscedastic Model

This can be demonstrated for the heteroscedastic pure probit case, Condition I is:

$$\begin{bmatrix} (\dot{s}_{11}^{N} + \dot{s}_{ff}^{N})/\tilde{m}_{N}^{2} & \\ (\dot{s}_{ff}^{N})/\tilde{m}_{N}^{2} & (\dot{s}_{22}^{N} + \dot{s}_{ff}^{N})/\tilde{m}_{N}^{2} \end{bmatrix} = \begin{bmatrix} (\dot{s}_{11} + \dot{s}_{33})/\tilde{m}^{2} & \\ (\dot{s}_{33})/\tilde{m}^{2} & (\dot{s}_{22} + \dot{s}_{33})/\tilde{m}^{2} \end{bmatrix},$$

where $\tilde{m}$ is the scale of the probit model (i.e., not the traditional Gumbel m).

Solving for the unknown parameters from the normalized model:

$$\tilde{m}_{N}^{2} = \tilde{m}^{2}\dot{s}_{ff}^{N}/\dot{s}_{33},$$
$$\dot{s}_{11}^{N} = \dot{s}_{11}\dot{s}_{ff}^{N}/\dot{s}_{33}, \text{ and}$$
$$\dot{s}_{22}^{N} = \dot{s}_{22}\dot{s}_{ff}^{N}/\dot{s}_{33}.$$

Condition II requires:

$$\tilde{m}_{N}^{2} > 0,$$
$$\dot{s}_{11}^{N} > 0,$$
$$\dot{s}_{22}^{N} > 0, \text{ and}$$
$$\dot{s}_{ff}^{N} > 0.$$

Given that the theoretical Σ_{Δ} is well behaved (i.e., all theoretical variances and scale are strictly positive), it is clear that any $\dot{s}_{ff}^{N} > 0$ will result in Conditions I and II being satisfied. So, the normalization is arbitrary, and the standard practice of normalizing any one of the terms to 1 is valid.

Examination of the normalization unrestricted probit and logit kernel models are provided in the Appendix. The heteroscedastic and unrestricted covariance matrix examples illustrate the nature of the problem. The issue arises due to the manner in which the normalized parameter estimates adjust to replicate the true covariance structure. With probit, the parameters shift in a simple multiplicative manner. However, with logit kernel, the parameters shift in an additive manner, and this can lead to infeasible 'negative' variances and a factor analytic term that is not positive definite.

The brief summary of identification is that the order and rank conditions need to be applied to verify that any estimated model is identified, and the positive definiteness condition needs to be applied to verify that a particular normalization is valid. It is critical to examine identification on a case-by-case basis, which is how we will proceed in the remainder of the paper. There is also an empirical issue concerning identification, which is whether or not the data provide enough information to estimate any given

theoretically identified structure. This is the usual multicollinearity problem, and it arises when there are too many parameters in the error structure and therefore the Hessian is nearly singular.

Special Cases

Many interesting cases can be embedded in the general factor analytic logit kernel specification presented in Equation (4). We will cover the following special cases in this section:

- *Heteroscedastic* – a summary and generalization of the discussion above.
- *Nested and Cross-nested* – analogous to nested and cross-nested logit.
- *Error Components* – a generalization of heteroscedastic and nested structures.
- *Factor Analytic* – a further generalization in which parameters in F_n are also estimated.
- *General Auto-Regressive* – particularly useful for large choice sets.
- *Random parameters* – where most of the current applications of logit kernel in the literature are focused.

This is not meant to be an exhaustive list. There are certainly other special cases of the logit kernel model, some of which are presented in papers listed in the references. The objective of this section is to show the flexibility of logit kernel, to provide specific examples of specification and identification, and to establish rules for identification and normalization for some of the most common special cases.

Heteroscedastic

The heteroscedastic model was presented above. The scalar notation form of the model is repeated here for convenience:

$$U_{in} = X_{in}\beta + \sigma_i \zeta_{in} + \varepsilon_{in} \ , \qquad i \in C_n \ .$$

Identification

Identification was already discussed above for $J = 2,\ 3,$ and 4. These results can be straightforwardly generalized to the following:

Identification

$J = 2$ none of the heteroscedastic variances can be identified.

$J \geq 3$ $J-1$ of the heteroscedastic variances can be identified.

Normalization

For $J \geq 3$, a normalization must be imposed on one of the variance terms, denote this as $\dot{s}_{jj} = \dot{s}^{N}_{ff}$ where $\dot{s}_{jj}$ is the true, albeit unknown, variance term that is fixed to the value $\dot{s}^{N}_{ff}$.

This normalization is not arbitrary, and must meet the following restriction:

$$\dot{s}^{N}_{ff} \geq \left(\dot{s}_{jj} - \dot{s}_{ii}\right) g \Big/ (g + \dot{s}_{ii}) \quad , \; i = 1,..., J \; .$$

This restriction shows that the natural tendency to normalize an arbitrary heteroscedastic term to zero is incorrect. If the alternative does not happen to be the minimum variance alternative, the parameter estimates will be inconsistent, there can be a significant loss of fit (as demonstrated in the application section), and it can lead to the incorrect conclusion that the model is homoscedastic. This is an important issue, which, as far as we can tell, is ignored in the literature. It appears that arbitrary normalizations are being made for models of this form (see, for example Gönül and Srinivasan, 1993, and Greene, 2000, Table 19.15). Therefore, there is a chance that a non-minimum variance was normalized to zero, which would mean that the model is misspecified. It is important to note that it is the addition of the i.i.d. disturbance that causes the identification problem. Therefore, heteroscedastic pure probit models as well as the heteroscedastic extreme value models (see, for example, Bhat, 1995, and Steckel and Vanhonacker, 1988) do not exhibit this property.

Ideally, we would like to impose a normalization such that MNL is a special case of the model. Therefore, the best normalization is to fix the minimum variance alternative to zero. However, there is in practice no prior knowledge of the minimum variance alternative. A brute force solution is to estimate J versions of the model, each with a different heteroscedastic term normalized; the model with the best fit is the one with the correct normalization. This is obviously cumbersome as well as time consuming. Alternatively, one can estimate the unidentified model with all J heteroscedastic terms. Although this model is not identified, it will pseudo-converge to a point that reflects the true covariance structure of the model. The heteroscedastic term with minimum estimated variance in the unidentified model is the minimum variance alternative, thus eliminating the need to estimate J different models. Examples of this method are provided in the applications section.

Nesting & Cross-Nesting Error Structures

Nesting and cross-nesting logit kernel is another important special case, and is analogous to nested and cross-nested logit. The nested logit kernel model is specified as follows:

$$U_n = X_n b + F_n T z_n + n_n \; ,$$

where: z_n is $(M \times 1)$, M is the number of nests, and one factor is defined for each nest.

F_n is $(J_n \times M)$, $f_{jm} = \begin{cases} 1 & \textit{if alternative } j \textit{ is a member of nest } m \\ 0 & \textit{otherwise} \end{cases}$

T is $(M \times M)$ diagonal, which contains the standard deviation of each factor.

In a strictly hierarchical nesting structure, the nests do not overlap, and $F_n F_n'$ is block diagonal. In a cross-nested structure, the alternatives can belong to more than one group.

Identification

As usual, the order and rank conditions are checked for identification. The order condition states that at most $J(J-1)/2-1$ nesting parameters can be identified. However, the rank condition leads to further restrictions as described below.

Models with 2 Nests

The summary of identification for a 2 nest structure is that only 1 of the nesting parameters is identified. Furthermore, the normalization of the nesting parameter is arbitrary. This is best shown by example. Take a 5 alternative case (with universal choice set) in which the first 2 alternatives belong to one nest, and the last 3 alternatives belong to a different nest. The model is written as:

$$\begin{aligned} U_{1n} &= \ldots + s_1 z_{1n} + n_{1n} \\ U_{2n} &= \ldots + s_1 z_{1n} + n_{2n} \\ U_{3n} &= \ldots + s_2 z_{2n} + n_{3n} \\ U_{4n} &= \ldots + s_2 z_{2n} + n_{4n} \\ U_{5n} &= \ldots + s_2 z_{2n} + n_{5n} \end{aligned} \quad , \quad \text{where: } F = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \text{ and } T = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix}.$$

We denote this specification as 1, 1, 2, 2, 2 (a shorthand notation of the matrix F). The covariance matrix of utility differences (with alternative 5 as the base) is as follows:

$$\Omega_\Delta = \begin{bmatrix} s_{11} + s_{22} + 2g/m^2 & & & \\ s_{11} + s_{22} + g/m^2 & s_{11} + s_{22} + 2g/m^2 & & \\ g/m^2 & g/m^2 & 2g/m^2 & \\ g/m^2 & g/m^2 & g/m^2 & 2g/m^2 \end{bmatrix}.$$

It can be seen from this matrix that only the sum $(s_{11} + s_{22})$ can be identified. This is verified by the rank condition as follows:

$$vecu(\Omega_\Delta) = \begin{bmatrix} s_{11} + s_{22} + 2g/m^2 \\ s_{11} + s_{22} + g/m^2 \\ g/m^2 \\ 2g/m^2 \end{bmatrix} \rightarrow \underset{matrix}{Jacobian} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \end{bmatrix} \rightarrow \text{RANK=2}$$

$\rightarrow$ can estimate 1 of the parameters; must normalize m and one s_{ii}.

Furthermore, unlike the heteroscedastic logit kernel model, either one of the variance terms can be normalized to zero (i.e., the normalization is arbitrary). This can be seen intuitively by noticing that only the sum $(s_{11} + s_{22})$ appears in Ω_{Δ}, and so it is always this sum that is estimated regardless of which term is set to zero. This can also be verified via the positive definiteness condition, as follows. Say we impose the normalization $\dot{s}^{N}_{22} = 0$. Condition I leads to the relationships $m^2_N = m^2$ and $\dot{s}^{N}_{11} = (\dot{s}_{11} + \dot{s}_{22})$. Condition II states that Σ^N must be positive semi-definite, where:

$$\Sigma^N = \begin{bmatrix} \dot{s}^{N}_{11} & & & & \\ \dot{s}^{N}_{11} & \dot{s}^{N}_{11} & & & \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \frac{1}{m^2_N} .$$

A matrix is positive semi-definite if all of its eigenvalues are non-negative. The eigenvalues for Σ^N shown above are: $2\dot{s}^{N}_{11} / m^2_N,\ 0,\ 0,\ 0,\ 0$. We know from Condition I that $m^2_N > 0$ and $\dot{s}^{N}_{11} \geq 0$, which means $2s^{N}_{11} / m^2_N \geq 0$, Σ^N is positive semi-definite, and the normalization $\dot{s}^{N}_{22} = 0$ is valid. Similarly, it can be shown that the normalization $\dot{s}^{N}_{11} = 0$ is also valid.

While it is not possible to estimate both variance parameters of the 1, 1, 2, 2, 2 structure, the following structures are all identified and result in *identical* covariance structures (i.e., identical models):

{ 1, 1, 0, 0, 0 } = { 0, 0, 2, 2, 2 } = { 1, 1, 2, 2, 2 with $s_1 = s_2$ } .

These results straightforwardly extend to all two nest structures regardless of the number of alternatives (as long as at least one of the nests has 2 or more alternatives).

Models with Three or More Nests

The summary of identification for models with 3 or more nests is that *all* of the nesting parameters are identified. To show this, we will again look at a 5 alternative model, this time imposing a 3 nest structure (1, 1, 2, 3, 3):

$$\begin{aligned} U_{1n} &= \ldots + s_1 z_{1n} + n_{1n} \\ U_{2n} &= \ldots + s_1 z_{1n} + n_{2n} \\ U_{3n} &= \ldots + s_2 z_{2n} + n_{3n} \\ U_{4n} &= \ldots + s_3 z_{3n} + n_{4n} \\ U_{5n} &= \ldots + s_3 z_{3n} + n_{5n} \end{aligned} \quad , \quad \text{where: } F = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } T = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix} .$$

The covariance matrix of utility differences is:

$$\Omega_{\Delta} = \begin{bmatrix} s_{11}+s_{33}+2g/m^2 & & & \\ s_{11}+s_{33}+g/m^2 & s_{11}+s_{33}+2g/m^2 & & \\ s_{33}+g/m^2 & s_{33}+g/m^2 & s_{22}+s_{33}+2g/m^2 & \\ g/m^2 & g/m^2 & g/m^2 & 2g/m^2 \end{bmatrix}.$$

A check of the rank condition verifies that all three variance parameters are identified:

$$vecu(\Omega_{\Delta}) = \begin{bmatrix} s_{11}+s_{33}+2g/m^2 \\ s_{11}+s_{33}+g/m^2 \\ s_{33}+g/m^2 \\ s_{22}+s_{33}+2g/m^2 \\ g/m^2 \\ 2g/m^2 \end{bmatrix} \rightarrow \underset{matrix}{Jacobian} = \begin{bmatrix} 1 & 0 & 1 & 2 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix} \rightarrow \text{RANK=4}$$

$\rightarrow$ can estimate 3 of the parameters; only need to normalize m.

It is an interesting result that 1, 1, 0, 2, 2 structure results in both variance parameters being identified (by virtue of having a 3 nest structure) whereas only one parameter of the 1, 1, 2, 2, 2 structure is identified.

Conceptually, the number of estimable parameters can be thought of in terms of the number of differences and number of covariances that are left in the utility differences. In a two nest structure, only one difference remains and no covariances and therefore one parameter is estimable. Whereas in a three nest structure, there are two differences, plus the covariance between these two differences, and so three parameters are estimable.

This finding can be extended to any model with 3 or more nests (where 'nests' can have only 1 alternative, as long as at least one nest has 2 or more alternatives) as follows. Without loss of generality, assume that the base alternative is a member of a nest with 2 or more alternatives (as in the example above). Define m_b as the group to which the base alternative belongs, and s_{bb} as the variance associated with this base. Recall that M is the number of nests. The covariance matrix of utility differences has the following elements:

On the diagonal:

$s_{ii}+s_{bb}+2g/m^2 \quad \forall\, i \notin m_b$,	*M-1* equations,	(24)
$2g/m^2$,	*1* equation.	(25)

On the off-diagonal:

$s_{bb}+g/m^2$,	*1* equation,	(26)
g/m^2,	irrelevant: a dependent equation,	
$s_{ii}+s_{bb}+g/m^2$ for some $i \notin m_b$,	irrelevant: a dependent equation.	

Equations (24) through (26) provide identification for all nesting parameters, and the remaining equations are dependent. In the two-nest case, Equation (26) does not exist, and thus is an equation short of identification.

Cross-Nested Models

There are no general rules for identification and normalization of cross-nested structures, and one has to check the rank condition on a case-by-case basis. For example, in the five alternative case in which the third alternative belongs to both nests (1, 1, 1-2, 2, 2), the (non-differenced) covariance matrix is:

$$\Omega = \begin{bmatrix} \sigma_{11} + \gamma/\mu^2 & & & & \\ \sigma_{11} & \sigma_{11} + \gamma/\mu^2 & & & \\ \sigma_{11} & \sigma_{11} & \sigma_{11} + \sigma_{22} + \gamma/\mu^2 & & \\ 0 & 0 & \sigma_{22} & \sigma_{22} + \gamma/\mu^2 & \\ 0 & 0 & \sigma_{22} & \sigma_{22} & \sigma_{22} + \gamma/\mu^2 \end{bmatrix}.$$

A check of the order and rank conditions would find that both of the parameters in this cross-nested structure are identified. However, note that the cross-nesting specification can have unintended consequences on the covariance matrix. For example, in the (1, 1, 1-2, 2, 2) specification shown above, the third alternative is forced to have the highest variance. There are numerous possible solutions. One is to add a set of heteroscedastic terms, another is to add factors such that all the alternative-specific variances are identical as with the following specification:

$$F = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \text{ and } T = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_2 \end{bmatrix}.$$

The covariance matrix of utility differences for this structure is as follows:

$$\Omega_\Delta = \begin{bmatrix} 2\sigma_{11} + 2\sigma_{22} + 2\gamma/\mu^2 & & & \\ 2\sigma_{11} + \sigma_{22} + \gamma/\mu^2 & 2\sigma_{11} + 2\sigma_{22} + 2\gamma/\mu^2 & & \\ 2\sigma_{11} + \gamma/\mu^2 & 2\sigma_{11} + \gamma/\mu^2 & 2\sigma_{11} + 2\gamma/\mu^2 & \\ \sigma_{11} + \gamma/\mu^2 & \sigma_{11} + \gamma/\mu^2 & \sigma_{11} + \gamma/\mu^2 & 2\sigma_{11} + 2\gamma/\mu^2 \end{bmatrix}.$$

A check of the rank condition verifies that both variance parameters are identified for this specification.

$$vecu(\Omega_{\Delta}) = \begin{bmatrix} 2s_{11} + 2s_{22} + 2g/m^2 \\ 2s_{11} + s_{22} + g/m^2 \\ 2s_{11} + g/m^2 \\ 2s_{11} + 2g/m^2 \\ s_{11} + g/m^2 \\ 2s_{11} + 2g/m^2 \end{bmatrix} \rightarrow \underset{matrix}{Jacobian} = \begin{bmatrix} 2 & 2 & 2 \\ 2 & 1 & 1 \\ 2 & 0 & 1 \\ 2 & 0 & 2 \\ 1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \rightarrow \text{RANK=3}$$

$\rightarrow$ can estimate 2 of the parameters, only need to normalize m.

Extensions to Nested Models

There are various complexities that can be introduced to the nesting structure, including multi-level nests, cross-nested structures with multiple dimensions, and unknown parameters in the loading matrix (F). While we have investigated various special cases of these extended models, we have not yet derived general rules for identification. We recommend that identification be performed automatically on a case-by-case basis by programming the rank and order conditions into the estimation program.

Error Components

The error component formulation is a generalization that includes the heteroscedastic, nested, and cross-nested structures. The model is specified as follows:

$$U_n = X_n b + F_n T z_n + n_n \ ,$$

where F_n, z_n, and T are defined as in the general case, and F_n is a matrix of fixed factor loadings equal to 0 or 1. If T is diagonal (as it often is), then the disturbances in scalar form are:

$$e_{in} = \sum_{m=1}^{M} f_{imn} s_m z_{mn} + n_{in}, \qquad i \in C_n \ ,$$

where:

$$f_{imn} = \begin{cases} 1 & \text{if the } m^{th} \text{ element of } z_n \text{ applies to alternative i for individual n,} \\ 0 & \text{otherwise .} \end{cases}$$

The number of factors can be less than, equal to, or greater than the number of alternatives.

Identification

The order condition states that up to $J(J-1)/2-1$ parameters in T are identified. However, it is always necessary to check the rank condition for the particular specification and the positive definiteness condition for valid normalizations. Examples were provided above for the special cases of heteroscedastic, nesting, and cross-nesting specifications. Note that the rank condition should always be checked when any

combination of nesting, cross-nesting, and heteroscedasticity are applied. That is, the identification rules cannot be independently applied for combinations.

Factor Analytic

The Factor Analytic specification is a further generalization in which the F_n matrix contains unknown parameters. The model is written as in the general case:

$$U_n = X_n b + F_n T z_n + n_n .$$

If T is diagonal, the disturbances can be written in scalar form as follows:

$$e_{in} = \sum_{m=1}^{M} f_{imn} s_m z_{mn} + n_{in}, \qquad i \in C_n ,$$

where both the f_{imn}'s and s_m's are unknown parameters.

Identification

This is a very broad class of models. Therefore, it is difficult to go beyond the rank and order generalizations of identification. However, note that some constraints must be imposed on F_n and T in order to achieve identification. For alternative-specific error structures, the minimum number of necessary constraints can be determined from the order condition: a maximum of $J(J-1)/2-1$ parameters can be estimated and there are up to $M(J+1)+1$ unknown parameters (M in T diagonal, JM in F_n, plus the scale term m). Once the order condition is met, the rank condition needs to be checked on a case-by-case basis. Finally, it must be verified that any imposed normalization satisfies the positive definiteness condition.

General Autoregressive Process

A fully unrestricted error correlation structure in models with large choice sets is problematic as the dimension of the integral is on the order of the number of alternatives and the number of parameters grows quadratically with the number of alternatives. A generalized autoregressive framework is attractive in these situations, because it allows one to capture fairly general error correlation structures using parsimonious parametric specifications. The key advantage of the method is that the number of parameters in the error structure grows linearly with the size of the choice set.

The disturbances $\dot{x}_n = (\dot{x}_{1n},...,\dot{x}_{J_n n})'$ [5] of a first-order generalized autoregressive process [GAR(1)] is defined as follows:

$$\dot{x}_n = r W_n \dot{x}_n + T_n z_n , \qquad z_n \sim N(0, I_{J_n}) , \tag{27}$$

[5] $\dot{x}_n$ has a slightly different interpretation than the x_n used elsewhere in the paper.

where W_n is a $(J \times J)$ matrix of weights $w_{i,j,n}$ describing the influence of each $\dot{x}_{jn}$ error upon the others, r is an unknown parameter, and $T_n z_n$ allows for heteroscedastic disturbances, where T_n is $(J_n \times J_n)$ diagonal (the subscript n is included to allow for different sized choice sets). Using a general notation, we write $w_{i,j,n}$ as:

$$w_{i,j,n} = \frac{w^*_{i,j,n}}{\sum_{k=1}^{J_n} w^*_{ik,n}} \quad , \qquad \forall \quad j \neq i \quad and \quad w_{i,j,n} = 0 \quad \forall \; i=j \; , \tag{28}$$

where $w^*_{i,j,n}$ is a function of unknown parameters and observable explanatory variables, which describe the correlation structure in effect. Solving for $\dot{x}_n$ in Equation (27) and incorporating it into Equation (4), leads to a logit kernel form of the GAR[1] specification:

$$U_n = X_n b + F_n T_n z_n + n_n, \qquad \text{where } F_n = (I - rW_n)^{-1}.$$

The normalization applied in Equation (28) ensures that the process is stable for values of r in the $(-1,1)$ interval. The interpretation and the sign of r, usually referred to as the correlation coefficient, depend on the definition of proximity embodied in w^*_{ij}.

In practice, the parameters in $w^*_{i,j,n}$ could be estimated. However, there are important special cases in which they are fixed. For example, spatial studies often use spatial autoregressive of order 1 [SAR(1)] error processes, which define the contiguity structure through a Boolean contiguity matrix. In this case, $w^*_{ij} = 1$ if i and j are contiguous and $w^*_{ij} = 0$ otherwise. For this specification, a $r > 0$ implies that errors of the same sign are grouped together. A slightly more complex specification, which requires estimation of a single parameter q, is to set $w^*_{ij} = (d_{ij})^{-q}$, in which the distance d_{ij} plays the role of a contiguity or proximity measure between pairs of alternatives. For examples of SAR(1) see Anselin (1989), and Cliff and Ord (1981). For an application of SAR(1) processes in economics, see Case (1991). Bolduc, Fortin, and Fournier (1996) use an SAR(1) process to estimate a logit kernel model with 18 alternatives.

For more details on GAR(1), including a discussion on identification issues, see Bolduc (1992).

Random Parameters

The MNL formulation with normally distributed random taste parameters can be written as:

$$U_n = X_n b_n + n_n \; , \text{ where } b_n \sim N(b, \Sigma_b).$$

b_n is a K-dimensional random normal vector with mean vector b and covariance matrix Σ_b. Replacing b_n with the equivalent relationship: $b_n = b + T z_n$, where T is the lower triangular Cholesky matrix such that $TT' = \Sigma_b$, leads to a general factor analytic logit kernel specification where $F_n = X_n$:

$$U_n = X_n b + X_n T z_n + n_n \; .$$

The parameters that need to be estimated in this model are b and those present in T. T is usually specified as diagonal, but it does not have to be (see, for example, Train, 1998, and Walker, 2001). Independently distributed parameters are probably a questionable assumption when variables are closely related, for example in-vehicle and out-of-vehicle travel time.[6] Also, note that the distribution does not have to be normal. For example, parameters with sign constraints should be specified with a lognormal distribution. See the telephone case study presented later for an example of a model with a lognormally distributed b_n parameter.

Identification

For identification of random parameter models, it is useful to separate the random parameters into two groups: those that are applied to alternative-specific constants and those applied to variables that vary across the sample.

Alternative-specific constants

When alternative-specific zero/one dummy variables have randomly distributed parameters, this is identical to the heteroscedastic, nested, and error component structures. In such cases, the order and rank conditions as discussed earlier hold.

Variables that include variation across the sample

As pointed out in the general discussion on identification, the order condition does not hold for the portion of the covariance matrix that varies across the sample. Rather, as many parameters as the data will support (without running into multicollinearity problems) can be estimated.

Continuous Attributes of the Alternatives

When random parameters are specified for continuous attributes of the alternatives, there are no identification issues per se. Data willing, the full covariance structure (i.e., variances for each parameter as well as covariances across parameters) can be estimated.

Categorical Attributes of the Alternatives

An interesting and unintuitive identification issue arises when categorical variables[7] are specified with *independently* distributed random parameters. Say there are M categories for a variable. Then there is theoretically a b_m and s_m for each category m, $m = 1,...,M$. It is well known that for the systematic terms (the b_m's), only $(M-1)$ b_m's can be identified and therefore a base must be arbitrarily selected. However, this is not necessarily true for the disturbance terms. To do the analysis,

[6] Note that if a subset of the covariances are estimated, then one has to be careful about the way the structural zeros are imposed on the Cholesky. In order for the structure of the Cholesky *T* (i.e., the location of the structural zeros) to be transferred to the covariance structure *TT'*, the structural zeros must be in the left-most cells of each row in the Cholesky. See Walker (2001) for more discussion.

[7] An example of a categorical variable in a housing choice context is *X*={street parking only, reserved parking space in a lot, private garage}, where each alternative has exactly one of the possible *X*'s associated with it.

the rank condition comes into play. Identification of the s_m's can be thought of as identification for a nested structure (think of it as examining the covariance structure for a particular individual). Therefore, if there are only 2 categories, then only one random parameter is identified and the normalization is arbitrary; if there are 3 or more categories, then a random parameter for each of the categories is identified. The key here being that, unlike the systematic portion of the utility function, it is incorrect to set one of the s_m's as a base when there are 3 or more categories. Unlike the identification analysis for a nested structure, the number of alternatives J does not impact the number of s_m's that can be estimated, because of the variation across observations. Note that this analysis applies for a single categorical variable, and it is not immediately apparent that the conclusion translates to the case when random parameters are specified for multiple categorical variables in the model. The issue of identification for categorical variables is not addressed in the literature, see, for example, Goett, Hudson, and Train (2000), who include random parameters on several categorical variables in their empirical results.

When covariances are estimated (as they probably should be), then a full set of variances and covariances can be estimated for the $M-1$ b_m's estimated in the systematic utility.

Characteristics of the Decision-maker

If a random parameter is placed on a variable that is a characteristic of the decision-maker (for example, years employed), it necessarily must be interacted with an alternative-specific variable (otherwise it will cancel out when the differences are taken). The normalization or such parameters then depends on the type of variable with which it interacts. If it interacts with alternative-specific dummy variables, then the heteroscedastic rules apply (i.e., $J-1$ variance terms can be estimated, and the minimum variance term must be constrained to zero). If it interacts with nest-specific constants, then the rules for nested error structures apply, etc. Furthermore, we suspect that if the characteristic is a categorical variable (for example, low income, medium income, high income), then the rules we presented for categorical attributes also apply (although this hasn't been verified).

Identification of Lognormally Distributed Parameters

Our application of the Order and Rank conditions for identification assume that the disturbance component of the utility can be separated from the systematic portion of the utility. With lognormally distributed parameters, the mean and variance of the distribution are a function of both of the disturbance parameters and therefore this separability does not exist. While the identification rules described above cannot be strictly applied, they provide guidelines for identification. And, as always, empirical tests such as examining the Hessian should also be applied.

As long as the identification restrictions described above are imposed, the number of random parameters that can be identified is dependent on the data itself in terms of the variation and the collinearity present in the explanatory variables. Therefore, empirical methods are used to verify identification of random parameter models, for example, verifying that the Hessian is non-singular at the convergence point. An issue with simulation is that identification issues often do not present themselves empirically unless a large

number of draws are used. Therefore, other useful methods are to constrain one or more parameters and observe whether the likelihood changes, or to test the impact of different starting values. Also, it is particularly important in random parameter models to verify stability of parameter estimates as the number of draws increases.

McFadden and Train (2000) note the inherent difficulty of identifying the factor structure for random parameter models, because many different factor combinations will fit the data approximately as well.

Parameter Estimation

We now describe the method that we use to estimate the joint vector of parameters $d = (b', y')'$, where b is the vector of unknown parameters in the systematic portion of the utility and y is the vector of unknown parameters in the error structure. For example, in the heteroscedastic model, only the alternative-specific standard deviations are included in y. In the GAR(1) version based on a Boolean contiguity matrix, the same standard deviations are estimated in addition to r (the correlation coefficient). The factor analytic and the random parameter structures can potentially have a very large number of unknown parameters.

The approach is to employ probability simulators within a maximum likelihood framework, which leads to Maximum Simulated Likelihood (MSL). The application of this method is straightforward and provides great flexibility in terms of the structure of the covariance matrix.

Maximum Likelihood

The log-likelihood of the sample is:

$$L(d) = \sum_{n=1}^{N} \ln P(i_n \mid d) \ ,$$

where $P(i_n \mid d)$ is the probability associated with the choice made by individual n. The score vector is:

$$\frac{\partial L(d)}{\partial d} = \sum_{n=1}^{N} \frac{1}{P(i_n \mid d)} \frac{\partial P(i_n \mid d)}{\partial d} .$$

Inserting the probability equations for the logit kernel model (Equations (6) and (7)) leads to the score for the logit kernel model:

$$\frac{\partial L(d)}{\partial d} = \sum_{n=1}^{N} \frac{1}{P(i_n \mid d)} \int_z \Lambda(i_n \mid d, z) \frac{\partial \ln \Lambda(i_n \mid d, z)}{\partial d} n(z, I_M) dz \ . \qquad (29)$$

Note that we also use the relationship $\partial X / \partial q = X \left(\partial \ln(X) / \partial q \right)$ in Equation (29) in order to make the derivative tractable: $\ln \Lambda(i_n \mid d, C_n) = X_{in} b + F_{in} T z_n - \ln \sum_{j \in C_n} e^{X_{jn} b + F_{jn} T z_n}$, which is easy to differentiate.

Each factor z introduces a dimension to the integral. Unless the dimension of z is small (≤ 3), the Maximum Likelihood (ML) estimator just described cannot be computed in a reasonable amount of time. For models with z of larger dimension, we use the Maximum Simulated Likelihood (MSL) methodology, described next.

Maximum Simulated Likelihood

The response probability for alternative i is replaced with the unbiased, smooth, tractable simulator:

$$\hat{P}(i \mid d) = \frac{1}{\mathbb{D}} \sum_{d=1}^{\mathbb{D}} \Lambda(i \mid d, z_n^d) \ , \tag{30}$$

where z_n^d denotes draw d from the distribution of z_n (each draw consists of M elements). Thus, the integral is replaced with an average of values of the function computed at discrete points. There has been a lot of research concerning how best to generate the set of discrete points (see Bhat, 2000, for a summary and references). The most straightforward approach is to use pseudo-random sequences. However, variance reduction techniques (for example, antithetic draws) and quasi-random approaches (for example, the Halton draws, which are used in the empirical results in this paper) have been found to cover the dimension space more evenly and thus are more efficient.

Incorporating the simulated probability, the simulated log-likelihood is then:

$$\hat{L}(d) = \sum_{n=1}^{N} \ln \hat{P}(i_n \mid d) \ , \tag{31}$$

and the simulated score is:

$$\frac{\partial \hat{L}(d)}{\partial d} = \sum_{n=1}^{N} \frac{1}{\hat{P}(i_n \mid d)} \frac{1}{\mathbb{D}} \sum_{d=1}^{\mathbb{D}} \Lambda(i_n \mid d, z_n^d) \frac{\partial \ln \Lambda(i_n \mid d, z_n^d)}{\partial d} . \tag{32}$$

A well-known result previously obtained in Börsch-Supan and Hajivassiliou (1993), among others, indicates that the log-likelihood function, although consistent, is simulated with a downward bias for finite number of draws. The issue is that while the probability simulator (30) is unbiased, the log-simulated-likelihood (31) is biased due to the log transformation. This can be seen by Jensen's inequality and the concavity of the log function. It can also be seen by taking a second degree Taylor's expansion of $\ln(\hat{P}(i))$ around $P(i)$, which gives:

$$\ln(\hat{P}(i)) \approx \ln(P(i)) + \frac{1}{P(i)}(\hat{P}(i) - P(i)) - \frac{1}{2P(i)^2}(\hat{P}(i) - P(i))^2 \ .$$

Taking the expected value of this relationship implies that:

$$\hat{L}(d) - L(d) \approx -\frac{\text{var}(\hat{P}(i \mid d))}{2P(i \mid d)^2} \leq 0 \ . \tag{33}$$

This suggests that in order to minimize the bias in simulating the log-likelihood function, it is important to simulate the probabilities with good precision. The precision increases with the number of draws, as well as with the use of efficient methods to generate the draws. The number of draws necessary to sufficiently remove the bias cannot be determined a priori; it depends on the type of draws, the model specification, and the data.

Applications

In this section, we consider four applications: two based on synthetic data and two on real data. The first sample concerns a hypothetical choice situation among three alternatives; the focus is on the parameter identification issues of heteroscedastic models. The second sample, also using synthetic data, has 5 alternatives and focuses on identification issues of categorical variables with random parameter. The third application uses a mode choice dataset that is used for logit kernel models that appear in two recent textbooks (Greene, 2000, and Louviere, Hensher, and Swait, 2000). We replicate the models presented in the texts, and use them to highlight practical issues that arise in estimating logit kernel models. The fourth application is based on a survey collected to predict residential telephone demand. We estimate several error structures for the telephone data, including heteroscedasticity, nesting, cross-nesting, and random parameter, and highlight many of the important identification and estimation issues of logit kernel models.

Estimation Notes & Practical Issues

Optimization Algorithm

While the likelihood function for linear in the parameters logit models is strictly concave, this is not true for logit kernel models (note that it is also not true for the nested logit model). Furthermore, the simple Newton methods that are used for MNL estimation tend to lose their robustness when the optimization function is not concave. Therefore, modified Newton methods, which address non-concavity with techniques such as trust regions, should be used for logit kernel models. For details on these methods, see Dennis and Schnabel (1983). In the applications presented in this paper, we use the `DUMIAH` routine provided in Fortran's IMSL Libraries. The `maxlik` routine provided in Gauss could also be used.[8]

Direction Matrix

To decrease estimation time, we analytically program the derivatives and approximate the matrix of second derivatives (the Hessian) with first order information. The most straightforward approximation of the Hessian is the BHHH technique (Berndt et al., 1974), which is computed as:

[8] Note that Kenneth Train of UC Berkeley provides Gauss-based estimation code for logit kernel (a.k.a. mixed logit) models from his website: http://emlab.berkeley.edu/users/train/index.html

$$\mathbf{R} = \sum_{n=1}^{N} \left(\frac{\partial L_n(d)}{\partial d} \right) \left(\frac{\partial L_n(d)}{\partial d} \right)', \tag{34}$$

where the score is defined as in Equation (29) (evaluated per sample observation). For Maximum Simulated Likelihood, it is computed with the simulated scores (32).

Under certain regularity conditions, BHHH can be shown to be a consistent estimator of the covariance matrix of parameters at the maximum likelihood estimate. There are also numerous other approximations that can be used, see Dennis and Schnabel (1983) for further discussion.

Standard Errors at Convergence

For a finite number of simulation draws, BHHH may substantially underestimate the covariance of the estimator due to simulation error (see McFadden and Train, 2000, for a discussion). BHHH (or some other approximation) is still preferred for the direction matrix due to the low cost of estimating the matrix as well as the robustness of estimation with regards to the direction matrix. However, it is advisable to use robust standard errors to generate the test statistics at convergence. A robust asymptotic covariance matrix estimator is $\mathbf{H}^{-1}\mathbf{R}\mathbf{H}^{-1}$ (Newey and McFadden, 1994), where $\mathbf{H}$ is the Hessian, calculated numerically or analytically, and $\mathbf{R}$ is defined as in Equation (34). When simulation is used, the simulated Hessian and Score are used. We report robust t-statistics (calculated using a numerical Hessian) for all estimation results.

Simulation Draws

We primarily use Halton draws for the simulation; however, some of the specifications are also estimated using pseudo-random draws for comparison. (See Bhat, 2000, and Train, 1999, for more information on Halton draws.) We have found the Halton draws to be more efficient than pseudo-random draws. For each observation, we draw $\mathbb{D}$ random vectors ($z_n^1, \ldots, z_n^{\mathbb{D}}$, each $(M \times 1)$) from the given multivariate distribution of the factors, and these draws are kept constant across iterations so that the simulator does not "chatter" as d changes (see McFadden and Train, 2000, for more information). The probability is then simulated using Equation (30), the log-likelihood using Equation (31), and the derivatives using Equation (32).

Simulation Bias and Identification

Two issues critical to estimating logit kernel models are simulation bias and identification.

As noted above, the number of draws, $\mathbb{D}$, must be large enough to sufficiently reduce the bias shown in Equation (33). The problem is that there is no way to know a priori how large is large enough, because this depends on the particular model structure and data. Therefore it is always necessary, as we do in these applications, to verify that the estimated parameters remain stable as the number of draws is increased.

The number of draws also plays an important role in testing for identification. Note that there are two forms of unidentification: structural, as indicated by the order and rank conditions, and informational, which

is when the data do not provide enough information to support the given structure (i.e., multicollinearity). It turns out that identification problems often do not appear (via a singular Hessian) when a small number of draws is used. For example, in the most extreme case, any specification (whether identified or not) will always appear identified when only 1 draw is used, because this is equivalent to adding explanatory variables to the systematic portion of the utility. This issue also emphasizes the importance of checking the rank condition prior to estimation, and of verifying robustness of estimates using different starting values.

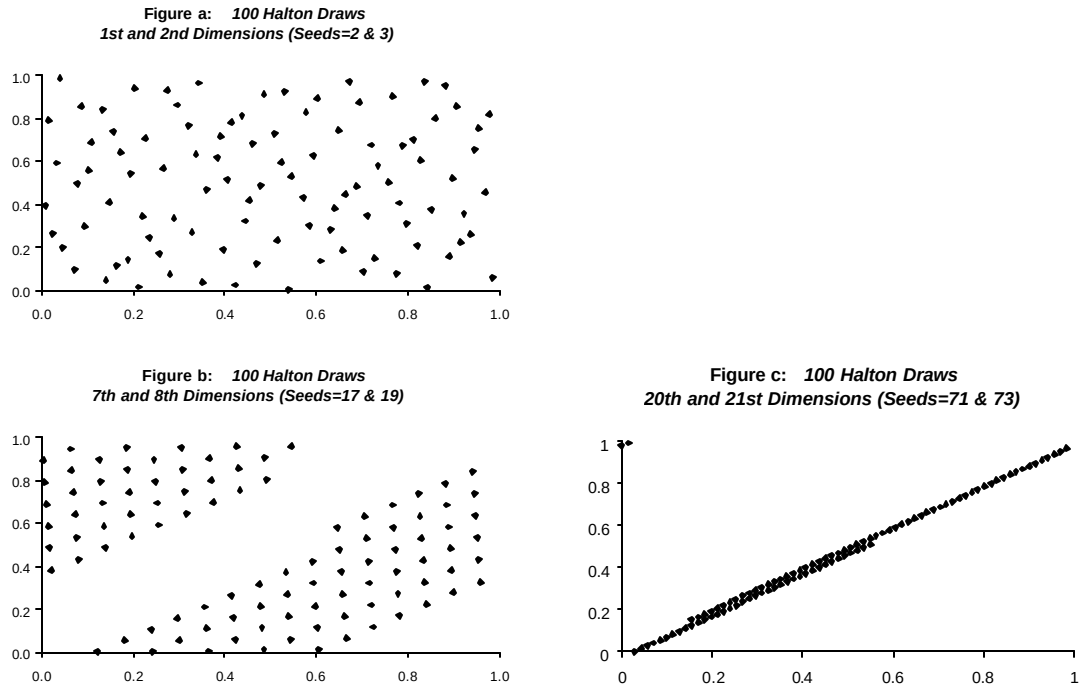

Figure 1: 100 Halton Draws for Different Dimensions of the Integral

Another issue with the number of draws is that as the dimension of the problem increases the number of draws necessary to estimate the model also increases. Conceptually, the issue is that it takes more draws to adequately cover the dimension space; this applies to all methods used to integrate non-closed form functions (for example, Gaussian quadrature or simulation via pseudo-random or quasi-random methods). It is interesting to note that with Halton draws, planes develop when small numbers of draws are used for high dimensional integrals. The generation of Halton draws is presented very clearly in Train (1999). Briefly, to implement Halton draws, a non-random series is developed for each dimension, each series is seeded with a prime number, and the seeds are implemented in order (2, 3, 5, 7, etc.). As an example of the problem with planes developing, take an extreme case: 100 draws are often sufficient to estimate a two dimensional model. As shown in Figure 1a, examination of a sample of Halton draws for a particular observation shows that the draws cover the 1st and 2nd dimensions of the sample space quite well. However, Figure 1b indicates that 100 draws for the 7th and 8th dimensions do not cover the space well, and Figure 1c shows that the 100 draws for the 20th and 21st dimensions are even worse.

To summarize, due to the issues of bias and identification, it is critical to empirically verify on a case-by-case basis that a sufficient number of draws are being used to estimate the model.

Synthetic Data I: Heteroscedasticity

The first application concerns a hypothetical choice situation among three alternatives. The model specification is as follows.

$$U_{1n} = a_1 + X_{1n}b + s_1 z_{1n} + n_{1n} \quad ,$$
$$U_{2n} = a_2 + X_{2n}b + s_2 z_{2n} + n_{2n} \, ,$$
$$U_{3n} = \quad X_{3n}b + s_3 z_{3n} + n_{3n} \, .$$

The true parameter values used to generate the synthetic data are:

$$a_1 = 1.5, \;\; a_2 = 0.5, \;\; b = -1, \;\; s_1 = 3, \;\; s_2 = 2, \;\; s_3 = 1, \; and \; m = 1.$$

The explanatory variable, X, is simulated as a normal variable with a standard deviation of 3, independent across alternatives and observations. The utilities for each observation are generated by drawing a single random draw for each z_{jn} from independent standard normal distributions and each n_{jn} from independent standard Gumbel distributions. The utilities are calculated, and the alternative with the highest utility is then the chosen alternative.

Estimation results using the synthetic data are provided in Table 1. Table 1a presents estimation results regarding selecting and setting the base heteroscedastic term. Recall that only $J-1$ heteroscedastic terms are identified, and that it is necessary to either set the minimum variance term to zero, or set any of the other variance terms high enough according to the equation derived earlier (Equation (23)):

$$\dot{s}_{ff}^{N} \geq \left(\dot{s}_{jj} - \dot{s}_{ii}\right) \Big/ (g + \dot{s}_{ii}) \quad , \; i = 1, \ldots, J \; ,$$

where $\dot{s}_{jj}$ is the theoretical (true) variance that is fixed to the value $\dot{s}_{ff}^{N}$.

All of the models in Table 1a are estimated with 10,000 observations and 500 Halton draws. The first model shows estimation results for an unidentified model; this model is used to determine the minimum variance alternative, and it correctly identifies the third alternative as having minimum variance.[9] Models 2 through 4 show identified models in which the minimum variance alternative is constrained to different values (0, 1, and 2); as expected, the log-likelihoods of these models are basically equivalent and all of these represent correct specifications. Models 5 through 10 show identified models in which the maximum variance alternative is constrained to different values (0, 1, 1.5, 2.25, 3, and 4). Applying Equation (23) (repeated above), the model specification will be correct as long as s_1 is constrained to a value above

[9] We were able to calculate t-statistics for the unidentified model here (and elsewhere) for two reasons. First, simulation has the tendency to mask identification issues, and therefore does not always result in a singular Hessian for a finite number of draws. Second, the slight difference between the Gumbel and Normal distributions makes the unidentified model only 'nearly' singular, and not perfectly singular.

2.2. The empirical results verify this. First, there is a severe loss of fit when the s_1 is constrained below 2.2. Second, the parameter estimates for the mis-specified models are biased. This can be seen by examining the ratio of the systematic parameters (for example, b / a_1) across models. While the scale shifts for various normalizations (and therefore the parameter estimates also shift), the ratio of systematic parameters should remain constant across normalizations. A cursory examination of the estimation results shows that these ratios begin to drift with successively invalid normalizations. Finally, note that these results indicate a slight loss of fit when the base alternative is constrained to a high value (s_3 =2 and s_1 =4), and this is due to the issue addressed earlier regarding the slight difference between the Gumbel and normal distributions. It must be emphasized that the normalization in heteroscedastic logit kernel models is not arbitrary.

Table 1: Synthetic Data I - Heteroscedastic Models (3 Alternatives)

Table a: Selecting and Setting the Base Heteroscedastic Term (10,000 Observations & 500 Halton Draws)

	True	Unidentified		Identified: Minimum Variance Base						Identified: Maximum Variance Base									
Parameter	Value	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
α_1	1.5	1.27	*(3.4)*	1.24	*(15.7)*	1.51	*(15.9)*	2.18	*(15.9)*	0.97	*(29.1)*	1.02	*(27.9)*	1.08	*(23.4)*	1.24	*(5.8)*	1.57	*(17.2)*	2.03	*(17.4)*
α_2	0.5	0.43	*(2.6)*	0.42	*(8.9)*	0.53	*(9.2)*	0.76	*(9.2)*	0.37	*(11.1)*	0.40	*(11.5)*	0.41	*(10.4)*	0.42	*(2.2)*	0.54	*(6.8)*	0.70	*(7.0)*
β	-1.0	-0.80	*(3.8)*	-0.78	*(14.6)*	-0.94	*(14.1)*	-1.36	*(13.7)*	-0.51	*(55.5)*	-0.57	*(65.0)*	-0.64	*(39.1)*	-0.78	*(16.0)*	-0.98	*(37.1)*	-1.27	*(37.1)*
σ_1	3.0	2.32	*(2.9)*	2.24	*(9.7)*	2.84	*(10.3)*	4.30	*(11.0)*	*0.00*	---	*1.00*	---	*1.50*	---	*2.25*	–	*3.00*	---	*4.00*	---
σ_2	2.0	1.27	*(1.9)*	1.21	*(4.7)*	1.69	*(5.9)*	2.80	*(7.7)*	0.06	*(0.1)*	0.03	*(0.3)*	0.50	*(1.8)*	1.22	*(6.6)*	1.82	*(11.7)*	2.58	*(14.5)*
σ_3	1.0	0.35	*(0.2)*	0.00	–	1.00	–	2.00	---	0.00	*(0.9)*	0.00	*(1.6)*	0.01	*-(0.5)*	0.16	*(0.0)*	1.07	*(4.4)*	1.78	*(7.6)*
(Simul.) Log-Likelihood:		-6837		-6837		-6837		-6838		*-6907*		*-6865*		-6845		-6837		-6837		-6838	
	Model:	1		2		3		4		5		6		7		8		9		10	

Table b: Varying the Numbers and Types of Draws (10,000 Observations)

			Halton Draws								Pseudo-Random Draws							
	True	True with	200 Halton		1000 Halton		2000 Halton		4000 Halton		500 'Random'		1000 'Random'		5000 'Random'		10000 'Random'	
Parameter	Value	σ_3=0	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
α_1	1.5	1.18	1.22	*(16.5)*	1.24	*(15.4)*	1.24	*(15.5)*	1.24	*(14.5)*	1.20	*(16.5)*	1.21	*(16.2)*	1.23	*(15.6)*	1.24	*(15.7)*
α_2	0.5	0.39	0.42	*(9.1)*	0.42	*(8.8)*	0.42	*(8.8)*	0.42	*(8.9)*	0.42	*(9.3)*	0.42	*(9.1)*	0.42	*(8.9)*	0.42	*(8.8)*
β	-1.0	-0.79	-0.77	*(15.6)*	-0.78	*(14.2)*	-0.78	*(14.3)*	-0.78	*(13.0)*	-0.75	*(15.6)*	-0.76	*(15.3)*	-0.78	*(14.4)*	-0.78	*(14.6)*
σ_1	3.0	2.23	2.19	*(10.2)*	2.25	*(9.5)*	2.26	*(9.5)*	2.25	*(8.7)*	2.14	*(10.2)*	2.15	*(10.0)*	2.23	*(9.5)*	2.26	*(9.7)*
σ_2	2.0	1.37	1.14	*(4.6)*	1.22	*(4.5)*	1.23	*(4.6)*	1.23	*(4.2)*	1.06	*(4.0)*	1.10	*(4.2)*	1.19	*(4.4)*	1.22	*(4.7)*
σ_3	1.0	*0.00*	*0.00*	---	*0.00*	---	*0.00*	---	*0.00*	---	*0.00*	---	*0.00*	---	*0.00*	–	*0.00*	---
(Simul.) Log-Likelihood:			-6837		-6837		-6837		-6836		-6835		-6839		-6838		-6836	

Table c: Varying the Number of Observations (500 Halton Draws)

	True	1000 Obs		5000 Obs		10000 Obs		40000 Obs		80000 Obs	
Parameter	Value	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
α_1	1.5	2.27	*(2.1)*	1.64	*(9.6)*	1.51	*(15.9)*	1.45	*(32.1)*	1.54	*(38.4)*
α_2	0.5	0.91	*(2.4)*	0.68	*(8.4)*	0.53	*(9.2)*	0.53	*(18.4)*	0.52	*(23.7)*
β	-1.0	-1.69	*(1.9)*	-0.99	*(8.3)*	-0.94	*(14.1)*	-0.95	*(29.2)*	-1.02	*(33.2)*
σ_1	3.0	5.64	*(1.7)*	3.13	*(6.5)*	2.84	*(10.3)*	2.85	*(21.3)*	3.05	*(24.8)*
σ_2	2.0	3.58	*(1.5)*	1.62	*(3.2)*	1.69	*(5.9)*	1.72	*(12.3)*	2.08	*(17.4)*
σ_3	1.0	*1.00*	---	*1.00*	---	*1.00*	---	*1.00*	---	*1.00*	---
(Simul.) Log-Likelihood:		-655		-3369		-6837		-27499		-54944	

The models shown in Table 1b were estimated to investigate the impact of the number and types of draws. All of these models are estimating using the normalization s_3 =0, and so we report the true parameters as calculated given this normalization (using Equations (15) to (17)). The model estimates verify that the 500 Halton draws used for the models in Table 1a are sufficient. The results also show that the Halton draws are more efficient then pseudo-random draws, as the parameter estimates stabilize for a lower number of Halton draws. Table 1c is provided to show that as the number of observations increases, the estimated parameters converge on their true values. Note that a potentially large number of observations is required to accurately reproduce the parameters of the population. However, the required number of observations is highly dependent on the model specification and data, and generalizations cannot be drawn.

Synthetic Data II: Random parameters on Categorical Variables

The second application, which also involves synthetic data, concerns the issue of identification of random parameters for categorical variables. Recall that if the variable has two categories (i.e., a 0/1 dummy) then one systematic parameter and one random parameter are identified, and the normalization of each is arbitrary. For variables with 3 (or more) categories, two systematic parameters are identified but all 3 random parameters (one per category) are identified. Empirical results are shown in Table 2. Table 2a, b, and c all use slightly different datasets and model specifications. The general specification is as follows:

$$U_{in} = a_i + \begin{bmatrix} X_{1in} & X_{2in} & X_{3in} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} X_{1in} & X_{2in} & X_{3in} \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix} \begin{bmatrix} z_{1n} \\ z_{2n} \\ z_{3n} \end{bmatrix} + n_{in}$$

$$, \forall\ i = 1,...,5\ ;\ n\ ,$$

where $a_5 = 0$ (the base alternative-specific constant) and X is a categorical variable, that is $X_{kin} = \{0,1\}$ & $X_{1in} + X_{2in} + X_{3in} = 1,\ \forall\ i;\ k = 1,\ldots,3; n$. The data are generated using the same approach as described in the synthetic data above, i.e., a X, z, $and n$ are sampled for each person, the utilities are calculated according to the model and parameters above, and the alternative with the highest utility is the chosen alternative. 10,000 observations are used for all of the models.

The dataset for the models in 2a includes a categorical variable with 2 categories ($X_{3in} = 0\ \forall\ i, n$). While the covariance structure varies across individuals, identification is analogous to a nested structure with two nests, for example, 1, 1, 2, 2, 2 or 1, 2, 2, 2, 2 or 1, 2, 1, 2, 1, etc. depending on the values of X for observation n.[10] Therefore, 1 systematic parameter (b) and 1 random parameter (s) can be estimated. Furthermore, the normalization of the random parameter is arbitrary. These statements are supported by the estimation results. The first two models show that the model with

[10] This concept of a categorical variable being analogous to a 2-nest nesting structure is denoted as *"~1, 1, 2, 2, 2"* in Table 2.

Table 2: Synthetic Data II – Categorical Variables with Random Parameters (5 Alternatives; 10,000 Observations)

Table a: Categorical variables with 2 categories, each enters all 5 utilities (~1, 1, 2, 2, 2)

		Unidentified		Unidentified		Identified: Base 1		Identified: Base 2		Identified: Base 2	
	True	500 Halton		500 Halton		500 Halton		500 Halton		1000 Halton	
Parameter	Value	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
α_1	0.5	0.48	*(11.2)*	0.48	*(11.2)*	0.48	*(11.2)*	0.48	*(11.2)*	0.48	*(11.2)*
α_2	0.5	0.44	*(10.2)*	0.44	*(10.2)*	0.44	*(10.2)*	0.44	*(10.2)*	0.44	*(10.2)*
α_3	1.0	0.92	*(22.7)*	0.92	*(22.7)*	0.92	*(22.7)*	0.92	*(22.7)*	0.92	*(22.7)*
α_4	1.0	0.98	*(24.2)*	0.98	*(24.2)*	0.98	*(24.2)*	0.98	*(24.2)*	0.98	*(24.2)*
β_1	0.5	0.50	*(7.9)*	0.50	*(7.9)*	0.50	*(7.9)*	0.50	*(7.9)*	0.50	*(7.9)*
σ_1	2.0	0.84	*(2.3)*	3.91	*(13.9)*			3.94	*(14.4)*	3.94	*(14.4)*
σ_2	4.0	3.85	*(13.6)*	0.47	*(0.7)*	3.94	*(14.4)*				
$(\sigma_1^2+\sigma_2^2)^{1/2}$	4.5	3.94		3.94		3.94		3.94		3.94	
(Simul.) Log-Likelihood:		-15310		-15310		-15310		-15310		-15310	
	Model:	*1*		*2*		*3*		*4*		*5*	

Table b: Categorical variables with 2 categories, each enters 4 of 5 utilities (~1, 1, 2, 2, 0)

		Misspecified 1		Misspecified 2		Identified		Identified	
	True	500 Halton		500 Halton		500 Halton		1000 Halton	
Parameter	Value	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
α_1	0.5	0.10	*(1.5)*	0.41	*(9.6)*	0.47	*(5.1)*	0.47	*(5.1)*
α_2	0.5	0.04	*(0.6)*	0.35	*(8.2)*	0.41	*(4.4)*	0.41	*(4.5)*
α_3	1.0	0.52	*(7.8)*	0.80	*(19.5)*	0.90	*(9.7)*	0.90	*(9.8)*
α_4	1.0	0.57	*(8.7)*	0.86	*(21.0)*	0.95	*(10.3)*	0.96	*(10.4)*
β_1	0.5	0.53	*(8.7)*	0.11	*(2.8)*	0.50	*(7.3)*	0.50	*(7.3)*
σ_1	2.0			2.29	*(16.0)*	1.73	*(8.4)*	1.73	*(8.5)*
σ_2	4.0	3.45	*(15.1)*			3.55	*(13.2)*	3.55	*(13.2)*
(Simul.) Log-Likelihood:		-15398		-15537		-15378		-15378	

Table c: Categorical variables with 3 categories, each enters all utilities (~1, 1, 2, 2, 3)

		Misspecified		Identified		Identified	
	True	500 Halton		500 Halton		1000 Halton	
Parameter	Value	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
α_1	0.5	0.36	*(7.7)*	0.36	*(7.7)*	0.36	*(7.7)*
α_2	0.5	0.40	*(8.5)*	0.40	*(8.5)*	0.40	*(8.5)*
α_3	1.0	0.93	*(20.5)*	0.93	*(20.6)*	0.93	*(20.6)*
α_4	1.0	0.92	*(20.2)*	0.92	*(20.3)*	0.92	*(20.3)*
β_1	1.0	1.06	*(6.4)*	1.06	*(6.4)*	1.06	*(6.7)*
β_2	0.5	1.06	*(7.0)*	0.69	*(4.4)*	0.70	*(4.4)*
σ_1	2.0	3.47	*(12.2)*	2.75	*(7.5)*	2.77	*(8.1)*
σ_2	3.0			2.52	*(6.8)*	2.49	*(6.7)*
σ_3	4.0	4.74	*(11.1)*	4.37	*(10.7)*	4.38	*(10.9)*
(Simul.) Log-Likelihood:		-15376		-15368		-15368	

both random parameters is unidentified, as the fit is identical for very different estimates of the random parameters. The third and fourth models show that the normalization is arbitrary: the parameter and fit are the same for either normalization. The fifth model verifies that enough draws are being used for estimation.

The dataset used for the models in Table 2b is similar to that used in Table 2a, with the exception that the categorical variable only applies to the first four alternatives ($X_{k5n} = 0 \ \forall \ k, n$). In this case, identification is related to a nested structure with three nests (for example, 1, 1, 2, 2, 0); therefore, 1 systematic parameter is estimable and *both* of the random parameters are estimable. This is shown in the estimation results, where the models with either of the systematic terms fixed to 0 results in a significant loss of fit.

In Table 2c, the categorical variable contains three categories. Identification here is also related to a nested model with 3 nests (for example, 1, 1, 2, 2, 3), and therefore 2 systematic parameters are identified and all 3 random parameters are identified. This is supported by the estimation results, in which constraining one of the random terms to zero results in a significant loss of fit.

Empirical Application I: Mode Choice

The logit kernel formulation is now making its way into econometric textbooks. In this section, we investigate the identification issues of logit kernel models that appear in Greene (2000, Table 19.15) and Louviere, Hensher and Swait (2000, Table B6.5). Both texts make use of the same data and present similar model specifications.

The Data

This is a revealed choice dataset containing mode choices for travel between Sydney and Melbourne, Australia. The choices available are air, train, bus, and car.[11] There are 210 observations in the sample, and the explanatory variables are[12]:

GCost: Generalized cost ($00)
= in vehicle cost + in vehicle time*value of travel time savings.

TTime: Terminal waiting time for plane, train and bus (hours). Auto terminal time is zero.

Income: Household income ($00,000), which is interacted with the 'air' alternative specific dummy variable.

[11] The dataset is actually a choice-based sample, and therefore the weighted exogenous sample maximum likelihood estimator (WESML, see Ben-Akiva and Lerman, 1985) should be used for the logit-based models (and the probit-equivalent for the probit models, see Imbens, 1992) to obtain consistent estimates. However, we did not use WESML in order to replicate the models as reported in the textbooks.

[12] Note: *(i)* The Louviere, Swait, and Hensher model also included a 'party size' explanatory variable. We based our models on the more parsimonious specification used in Greene. *(ii)* We scaled the data differently than that used for the models reported in the textbooks.

Models

In this section, we use the models presented in Greene and Louviere et al. to highlight various practical issues in model estimation. Greene estimated a series of models including probit as well as several logit kernel specifications (an unrestricted covariance structure, a heteroscedastic model, and a more general random parameter model). Louviere et al. present an even more general random parameter model.

Table 3: Mode Choice Model – Probit

Specification:	Unidentified				Identified			
Draws:	1000 'Random'		1000 'Random'		1000 'Random'		5000 'Random'	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants								
Air (1)	0.270	n/a	0.968	n/a	0.456	(1.2)	0.377	(0.6)
Train (2)	0.579	n/a	2.10	n/a	0.959	(4.8)	0.917	(3.5)
Bus (3)	0.486	n/a	1.76	n/a	0.805	(4.4)	0.768	(3.1)
GCost ($00)	-0.468	n/a	-1.70	n/a	-0.772	(4.0)	-0.747	(4.6)
TTime (hours)	-0.662	n/a	-2.39	n/a	-1.10	(3.8)	-1.03	(2.3)
Income ($00,000) - Air (1)	0.700	n/a	2.54	n/a	1.15	(2.0)	1.16	(2.5)
T11	0.608	n/a	2.20	n/a	1.00	----	1.00	----
T21	0.131	n/a	0.476	n/a	0.216	(0.9)	0.224	(2.3)
T31	0.0736	n/a	0.267	n/a	0.121	(0.5)	0.132	(1.5)
T22	0.246	n/a	0.888	n/a	0.407	(3.0)	0.381	(2.9)
T32	0.113	n/a	0.408	n/a	0.186	(1.5)	0.175	(2.9)
T33	0.130	n/a	0.471	n/a	0.216	(2.7)	0.202	(2.4)
Log Likelihood (simul.):	-197.727		-197.727		-197.727		-197.784	

Unrestricted Probit

The first model we present is a probit model in which the covariance matrix of utility differences (Ω_Δ) is unrestricted. In this case, the parameters of the Cholesky decomposition of Ω_Δ are estimated, or:

$$T = \begin{bmatrix} T_{11} & 0 & 0 \\ T_{21} & T_{22} & 0 \\ T_{31} & T_{32} & T_{33} \end{bmatrix}, \text{ where } TT' = \Omega_\Delta .$$

Note that even with probit, one has to be careful about identification. The Order Condition states that only five of the six parameters can be estimated. (Greene indirectly estimates all six, and therefore reports results for an unidentified model.) The need for this restriction can be verified empirically, and we present the results in Table 3. These were obtained using the GHK simulator with pseudo-random draws. First we report two sets of estimation results for the unidentified model. The two models have identical fits and yet different parameter estimates (note that the difference is a scale shift). The models also have a singular Hessian and therefore t-stats could not be generated. We also report estimation results for the identified model (setting $T_{11} = 1$). The model is now identified: the fit is identical to the unidentified models and the Hessian is not singular. The 5,000 draw result is provided to verify stability.

Unrestricted Logit Kernel

Greene also presents a logit kernel version of the probit model presented Table 3 (which he calls a 'constants random parameters logit model'). For the logit kernel version, the disturbance parameters include the six T_{ij} parameters as well as the logit scale parameter m. The identification of this model presents some interesting issues. First, an application of the order condition suggests that the m as well as one of the T_{ij}'s must be normalized for identification. However, as we will show empirically, this is not exactly the case. The reason is due to the slight difference between the Normal and Gumbel distribution. Since there is not an exact trade-off between the probit-like term and the Gumbel, there is an optimal weighting between the two distributions that make up the disturbance, and this allows an extra term to be estimated. Nonetheless, the model is nearly singular without a constraint on a T_{ij}, and so it is advisable to impose a normalization.

The second issue relates to the manner in which T_{ij} is normalized. The covariance matrix of utility differences for this model is:

$$\begin{bmatrix} T_{11}^2 + 2g/m^2 & & \\ T_{11}T_{21} + g/m^2 & T_{21}^2 + T_{22}^2 + 2g/m^2 & \\ T_{11}T_{31} + g/m^2 & T_{21}T_{31} + T_{22}T_{32} + g/m^2 & T_{31}^2 + T_{32}^2 + T_{33}^2 + 2g/m^2 \end{bmatrix}$$

We want to impose a normalization such that the model can reduce to a pure MNL. Therefore we want to normalize some $T_{ij} = 0$. Note that we cannot set $T_{11} = 0$, because this will restrict two of the covariance terms in the probit portion to be zero. We have also found empirical evidence that it is not always valid to set $T_{22} = 0$ due to the positive definiteness condition. However, it appears that the normalization $T_{33} = 0$ (or, more generally normalizing the lowest diagonal element of the cholesky matrix) is a valid normalization, and this is what we apply for this model. (See the Appendix for more information.)

The empirical results for the unrestricted logit kernel model are provided in Table 4. The first two columns provide estimation results for the case in which all six T_{ij}'s are estimated. The model is identified as suggested by a non-singular Hessian and stable parameter estimates as the number of draws is increased. The middle columns provide estimation results for models in which T_{33} is normalized to various values. There is marginal loss of fit due to the normalizations, but the likelihood function is fairly flat across the normalizations. The final column is provided to verify the stability of the normalized model with a high number of draws.

Heteroscedastic Logit Kernel

Greene also reports a heteroscedastic logit kernel model (which he calls an 'uncorrelated random parameters logit model'). As with the unrestricted logit kernel model discussed above, the rank and order conditions suggest a normalization is necessary when this is not exactly the case. Nonetheless, a normalization is advisable since the model is otherwise nearly singular. Furthermore, as we emphasized earlier, if a normalization is imposed, the selection of the base alternative to normalize is not arbitrary.

Table 4: Mode Choice Model – Unrestricted Logit Kernel

Specification:	Multinomial Logit		'Unidentified' (Nearly Singular)				Identified with Various Normalizations								Identified	
Draws:			2000 Halton		40.000 Halton		2000 Halton		2000 Halton		2000 Halton		2000 Halton		4000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants																
Air (1)	5.21	*(5.3)*	4.42	*(1.4)*	4.41	*(1.5)*	4.42	*(1.4)*	4.76	*(0.8)*	8.28	*(0.3)*	25.3	*(0.9)*	4.41	*(1.4)*
Train (2)	3.87	*(7.5)*	6.09	*(1.2)*	6.02	*(1.5)*	6.09	*(1.4)*	8.28	*(2.5)*	19.0	*(2.6)*	41.4	*(5.5)*	6.09	*(1.1)*
Bus (3)	3.16	*(5.8)*	5.00	*(1.1)*	4.93	*(1.4)*	5.00	*(1.4)*	6.92	*(2.5)*	15.9	*(3.0)*	35.1	*(5.8)*	5.00	*(1.0)*
GCost ($00)	-1.55	*(3.1)*	-4.04	*(0.7)*	-3.97	*(0.8)*	-4.04	*(0.8)*	-6.22	*(1.3)*	-15.4	*(1.5)*	-33.1	*(1.7)*	-4.04	*(0.6)*
TTime (hours)	-5.77	*(6.4)*	-7.50	*(1.8)*	-7.43	*(2.3)*	-7.50	*(2.2)*	-9.73	*(3.5)*	-21.5	*(4.6)*	-48.9	*(5.6)*	-7.50	*(1.7)*
Income ($00,000) - Air (1)	1.33	*(1.4)*	5.55	*(0.5)*	5.44	*(0.6)*	5.55	*(0.6)*	8.91	*(0.8)*	23.5	*(0.5)*	40.5	*(0.7)*	5.55	*(0.5)*
T11			4.85	*(0.6)*	4.76	*(0.7)*	4.85	*(0.7)*	7.78	*(1.0)*	20.3	*(0.8)*	40.8	*(1.5)*	4.86	*(0.5)*
T21			0.934	*(0.4)*	0.904	*(0.5)*	0.933	*(0.5)*	1.59	*(0.9)*	4.35	*(0.6)*	7.83	*(1.1)*	0.928	*(0.4)*
T31			0.554	*(0.4)*	0.538	*(0.5)*	0.554	*(0.5)*	0.913	*(0.7)*	2.50	*(0.6)*	4.30	*(0.5)*	0.551	*(0.4)*
T22			1.25	*(0.3)*	1.18	*(0.3)*	1.25	*(0.3)*	2.81	*(1.2)*	7.79	*(3.5)*	17.9	*(3.1)*	1.25	*(0.2)*
T32			0.711	*(0.3)*	0.681	*(0.4)*	0.711	*(0.4)*	1.30	*(1.4)*	3.44	*(1.4)*	7.55	*(2.2)*	0.709	*(0.3)*
T33			5.12E-03	*(0.1)*	-7.88E-05	*(0.0)*	0.000	----	1.00	----	4.00	----	10.0	----	0.00	----
Log Likelihood (simul.):	-199.128		-195.466		-195.491		-195.466		-196.500		-197.713		-197.647		-195.481	

Table 5: Mode Choice Model – Heteroscedastic Logit Kernel

Specification:	Multinomial Logit		Heteroscedastic Models									
			'Unidentified'		Identified: Base 1		Identified: Base 3		Identified: Base 4		Identified: Base 4	
Draws:			1000 Halton		1000 Halton		1000 Halton		1000 Halton		5000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants												
Air (1)	5.21	*(5.3)*	4.65	*(3.1)*	5.21	*(6.4)*	4.65	*(3.1)*	4.62	*(3.6)*	4.69	*(3.7)*
Train (2)	3.87	*(7.5)*	5.19	*(4.6)*	3.87	*(7.9)*	5.19	*(4.8)*	5.07	*(6.8)*	5.08	*(7.2)*
Bus (3)	3.16	*(5.8)*	4.20	*(3.9)*	3.16	*(6.4)*	4.21	*(4.0)*	4.11	*(5.4)*	4.12	*(5.8)*
GCost ($00)	-1.55	*(3.1)*	-3.27	*(3.2)*	-1.55	*(3.7)*	-3.27	*(3.3)*	-3.17	*(4.3)*	-3.15	*(4.6)*
TTime (hours)	-5.77	*(6.4)*	-6.90	*(5.4)*	-5.77	*(10.8)*	-6.90	*(5.7)*	-6.78	*(7.0)*	-6.78	*(7.8)*
Income ($00,000) - Air (1)	1.33	*(1.4)*	3.68	*(1.4)*	1.33	*(1.1)*	3.68	*(1.4)*	3.53	*(1.4)*	3.45	*(1.5)*
σ_1			3.38	*(3.1)*	0.00	–	3.38	*(3.2)*	3.27	*(3.4)*	3.18	*(3.6)*
σ_2			0.143	*(0.0)*	0.0414	*(0.0)*	0.143	*(0.0)*	0.128	*(0.0)*	0.029	*(0.0)*
σ_3			0.00206	*(0.0)*	0.0181	*(0.0)*	0.00	---	0.00266	*(0.0)*	0.00584	*(0.0)*
σ_4			0.432	*(0.2)*	0.0558	*(0.0)*	0.434	*(0.2)*	0.00	---	0.00	---
Log Likelihood (simul.):	-199.128		-196.751		-199.118		-196.751		-196.768		-196.255	

The empirical results for the Mode Choice dataset are provided in Table 5. We estimate the 'unidentified' model to determine the parameters that are candidates for normalization. The results suggest that train, bus, or car can be used as the base (Greene normalizes the car alternative). We then report several identified models with different base alternatives normalized, and show that the model in which the air heteroscedastic term is the base is a mis-specified model (as indicated by the loss of fit).

Random Parameter Logit Kernel

Greene also reports a model that expands the unrestricted logit kernel model presented in Table 4 by including normally distributed random parameters for the cost, time, and income variables.[13] The primary issue here is that there are only 210 observations in the sample, and it is not a rich enough dataset to support the estimation of a large number of disturbance parameters. This is demonstrated with the empirical results reported in Table 6, in which we present a series of random parameter models starting with more parsimonious specifications.

The first model is the multinomial logit model, provided for comparison. Model 1-2 (estimated with 2000 and 4000 Halton draws) includes independent random parameters on the cost, time, and income variables. This model appears identified, and results in a large improvement in fit over the multinomial logit model.[14] The t-stats are low here due to the correlation among the parameter estimates. Model 4 shows that allowing for a single random parameter on the time variable achieves much of the total improvement in fit. Model 5-6 (estimated with 2000 and 4000 Halton draws) allows for a full set of correlations among the random parameters, and this results in a marginal improvement in fit over the independent model. (Note that the Cholesky parameters and not the variances and covariances are reported). Model 7 is estimated with a more parsimonious correlated structure. So far, these models all appear to be identified and provide significant (and similar) explanation of the disturbances. This is not the case for the remaining models. Model 8-9 includes the three independent random parameters along with heteroscedasticity, and the model appears unidentified. Model 10 is the model reported in Greene (although we normalized T_{33}). It includes an unrestricted covariance structure as well as the three independent random parameters, and the model appears unidentified. Louviere, Hensher and Swait report estimation results for a model similar to Greene (i.e., an unrestricted covariance structure with additional random parameters), and their model, too, appears unidentified.

The important points of these random parameter results are that, first, there are often several specifications that result in a similar improvement in fit. Second, that it is important not to overdue the specification, because it is easy to end up with an unidentified model.

[13] Note that since the time and cost parameters have a sign constraint, they should be specified with log-normally distributed parameters.

[14] Note that we achieved a much larger improvement in fit than any of the models reported in Greene and Louviere et al., even with this more parsimonious specification.

Table 6: Mode Choice Model – Random Parameters

Specification:	Multinomial Logit		Independent Random Parameters						Correlated Random Parameters					
Draws:			2000 Halton		4000 Halton		4000 Halton		2000 Halton		4000 Halton		4000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Systematic Parameters:														
Altern. Specific constants														
Air (1)	5.21	*(5.3)*	12.0	*(3.6)*	11.8	*(2.9)*	9.49	*(5.7)*	17.8	*(2.5)*	17.6	*(2.6)*	10.8	*(3.8)*
Train (2)	3.87	*(7.5)*	12.9	*(3.1)*	12.7	*(2.5)*	9.65	*(5.5)*	18.4	*(2.4)*	18.3	*(2.5)*	10.7	*(3.6)*
Bus (3)	3.16	*(5.8)*	11.6	*(3.2)*	11.5	*(2.6)*	8.69	*(5.5)*	16.7	*(2.4)*	16.5	*(2.5)*	9.7	*(3.7)*
GCost ($00)	-1.55	*(3.1)*	-4.21	*(2.0)*	-4.14	*(1.6)*	-2.57	*(3.3)*	-6.71	*(1.6)*	-6.53	*(1.8)*	-4.02	*(1.9)*
TTime (hours)	-5.77	*(6.4)*	-16.7	*(3.3)*	-16.5	*(2.7)*	-12.5	*(5.8)*	-24.1	*(2.4)*	-24.1	*(2.5)*	-13.4	*(3.9)*
Income ($00,000) - Air (1)	1.33	*(1.4)*	9.61	*(1.9)*	9.48	*(1.7)*	5.93	*(2.5)*	14.4	*(1.6)*	14.3	*(1.7)*	5.5	*(2.0)*
Nonzero Cholesky: T11 (σ_1)														
T21														
T31														
T22 (σ_2)														
T32														
T33 (σ_3)														
GCost			0.493	*(0.4)*	0.332	*(0.1)*			4.99	*(0.9)*	4.86	*(1.1)*	3.00	*(1.3)*
TTime			10.7	*(2.5)*	10.6	*(2.1)*	7.9	*(3.7)*	13.6	*(2.0)*	14.1	*(2.0)*	3.86	*(0.4)*
Income - Air			8.34	*(1.3)*	8.18	*(1.1)*			6.94	*(1.0)*	5.56	*(1.3)*		
GCost - TTime									9.21	*(1.5)*	8.13	*(1.8)*	7.70	*(2.0)*
GCost - (Income-Air)									6.57	*(0.6)*	9.03	*(0.9)*		
TTime - (Income-Air)									-13.6	*(1.3)*	-14.6	*(1.5)*		
Log Likelihood (simul.):	-199.128		-177.523		-177.640		-178.680		-174.419		-174.420		-176.816	
Model:	*1*		*2*		*3*		*4*		*5*		*6*		*7*	

Specification:	Random Parameters & Heteroscedasticity				Random Param. & Unconstrained	
Draws:	2000 Halton		4000 Halton		2000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Systematic Parameters:						
Altern. Specific constants						
Air (1)	25.7	*n/a*	28.2	*n/a*	44.1	*n/a*
Train (2)	31.3	*n/a*	34.2	*n/a*	56.0	*n/a*
Bus (3)	27.8	*n/a*	30.4	*n/a*	48.4	*n/a*
GCost ($00)	-13.4	*n/a*	-14.6	*n/a*	-23.0	*n/a*
TTime (hours)	-39.5	*n/a*	-43.3	*n/a*	-69.9	*n/a*
Income ($00,000) - Air (1)	25.5	*n/a*	28.7	*n/a*	48.6	*n/a*
Nonzero Cholesky: T11 (σ_1)	12.4	*n/a*	11.7	*n/a*	24.3	*n/a*
T21					2.69	*n/a*
T31					-0.389	*n/a*
T22 (σ_2)	2.16	*n/a*	2.07	*n/a*	4.90	*n/a*
T32					2.68	*n/a*
T33 (σ_3)	0.57	*n/a*	1.60	*n/a*	0.00	----
GCost	0.10	*n/a*	2.16	*n/a*	-2.67	*n/a*
TTime	25.5	*n/a*	28.1	*n/a*	45.8	*n/a*
Income - Air	6.69	*n/a*	18.69	*n/a*	13.1	*n/a*
GCost - TTime						
GCost - (Income-Air)						
TTime - (Income-Air)						
Log Likelihood (simul.):	-176.072		-176.036		-175.393	
Model:	*8*		*9*		*10*	

Empirical Application II: Telephone Service

In this section, we apply these methods to residential telephone demand analysis. The model involves a choice among five residential telephone service options for local calling. A household survey was conducted in 1984 for a telephone company and was used to develop a comprehensive model system to predict residential telephone demand (Train, McFadden and Ben-Akiva, 1987). Below we use part of the data to estimate a model that explicitly accounts for inter-dependencies between residential telephone service options. We first describe the data. Then we present estimation results using a variety of error structures.

The Data

Local telephone service typically involves the choice between flat (i.e., a fixed monthly charge for unlimited calls within a specified geographical area) and measured (i.e., a reduced fixed monthly charge for a limited number of calls plus usage charges for additional calls) services. In the current application, five services are involved, two measured and three flat. They can be described as follows:

- *Budget measured* - no fixed monthly charge; usage charges apply to each call made.
- *Standard measured* - a fixed monthly charge covers up to a specified dollar amount (greater that the fixed charge) of local calling, after which usage charges apply to each call made.
- *Local flat* - a greater monthly charge that may depend upon residential location; unlimited free calling within local calling area; usage charges apply to calls made outside local calling area.
- *Extended area flat* - a further increase in the fixed monthly charge to permit unlimited free calling within an extended area.
- *Metro area flat* - the greatest fixed monthly charge that permits unlimited free calling within the entire metropolitan area.

The sample concerns 434 households. The availability of the service options of a given household depends on its geographical location. Details are provided in Table 7. In Table 8, we summarize the service option availabilities over the usable sample.

Table 7: Telephone Data - Availability of Service Options

Service Options	Geographic Location		
	Metropolitan Areas	Perimeter Exchanges Adjacent to Metro Areas	All Other
Budget Measured	Yes	Yes	Yes
Standard Measured	Yes	Yes	Yes
Local Flat	Yes	Yes	Yes
Extended Flat	No	Yes	No
Metro Flat	Yes	Yes	No

Table 8: Telephone Data - Summary Statistics on Availability of Service Options

Service Options	Chosen	Percent	Total Available
Budget Measured	73	0.168	434
Standard Measured	123	0.283	434
Local Flat	178	0.410	434
Extended Flat	3	0.007	13
Metro Flat	57	0.131	280
Total :	434	1.000	1595

Models

The model that we use in the present analysis is intentionally specified to be simple. The explanatory variables used to explain the choice between the five service options are four alternative-specific constants, which correspond to the first four service options, and a generic cost variable (the natural log of the monthly cost of each service options expressed in dollars). We investigated three types of error structures: heteroscedasticity, nested and cross-nested structures, and taste heterogeneity (random parameters).

Heteroscedastic

The results for the heteroscedastic case are provided in Table 9 and Table 10. Table 9 displays results from the unidentified model. To explore the issue of normalization of the minimum variance alternative, we estimated the unidentified model for various numbers of Halton draws and pseudo-random draws. The results suggest that there is no strong base alternative, and it could be either alternative 1, 2, 4, or 5. Table 10 provides estimation results for identified heteroscedastic models. Again, to explore the issue of the minimum variance alternatives, 5 identified models were estimated, each one with a different base heteroscedastic term. (Note that this defeats the purpose of estimating the unidentified model, but was done for illustration purposes only.) As indicated by the unidentified models, the identified model estimation results support the conclusion that any of alternatives 1, 2, 4, or 5 could be set as the base. However, constraining s_3 to zero results in a significant loss of fit, whereas constraining it to 4.0 brings it in line with the correctly specified model. Comparing the correctly specified heteroscedastic models with the MNL model, there is an obvious gain in likelihood from incorporating heteroscedasticity, primarily due to capturing the high variance of alternative 3.

Table 9: Telephone Model - Heteroscedastic Unidentified Models to Determine Base

	100 Halton		200 Halton		400 Halton		1000 Halton		2000 Halton		5000 'Random'		10000 'Random'	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants														
Budget Measured (1)	-3.30	*(6.9)*	-163.39	*n/a*	-3.28	*(7.5)*	-3.28	*(7.7)*	-3.27	*(7.6)*	-3.32	*(7.2)*	-3.29	*(7.7)*
Standard Measured (2)	-2.55	*(5.5)*	-126.84	*n/a*	-2.53	*(6.3)*	-2.53	*(6.4)*	-2.52	*(6.8)*	-2.55	*(6.4)*	-2.53	*(6.5)*
Local Flat (3)	-1.38	*(3.5)*	-78.09	*n/a*	-1.37	*(3.6)*	-1.37	*(3.6)*	-1.36	*(3.6)*	-1.38	*(3.7)*	-1.37	*(3.6)*
Extended Flat (4)	-1.07	*(1.3)*	-44.31	*n/a*	-1.04	*(1.3)*	-1.04	*(1.3)*	-1.04	*(1.5)*	-1.06	*(1.5)*	-1.04	*(1.4)*
Log Cost	-2.70	*(7.2)*	-145.18	*n/a*	-2.68	*(7.9)*	-2.68	*(8.2)*	-2.67	*(8.4)*	-2.70	*(8.1)*	-2.69	*(7.6)*
σ_1	0.10	*(0.3)*	60.29	*n/a*	0.06	*(0.3)*	0.03	*(0.2)*	0.00	*(0.1)*	0.31	*(0.5)*	0.13	*(0.4)*
σ_2	0.30	*(0.3)*	61.19	*n/a*	0.21	*(0.3)*	0.14	*(0.4)*	0.06	*(0.3)*	0.20	*(0.2)*	0.08	*(0.2)*
σ_3	2.91	*(3.2)*	196.53	*n/a*	2.88	*(3.3)*	2.88	*(3.4)*	2.87	*(3.6)*	2.91	*(4.3)*	2.91	*(3.1)*
σ_4	0.39	*(0.3)*	16.18	*n/a*	0.01	*(0.0)*	0.04	*(0.1)*	0.01	*(0.0)*	0.11	*(0.2)*	0.07	*(0.3)*
σ_5	0.22	*(0.2)*	81.36	*n/a*	0.01	*(0.1)*	0.09	*(0.3)*	0.01	*(0.0)*	0.05	*(0.1)*	0.26	*(0.2)*
(Simul.) Log-Likelihood:	-471.09		-468.27		-471.16		-471.20		-471.19		-470.89		-471.38	

Table 10: Telephone Model - Identified Heteroscedastic Models

	MNL		Identified Heteroscedastic Model													
			1000 Halton		1000 Halton		1000 Halton		1000 Halton		1000 Halton		1000 Halton		5000 'Random'		10000 'Random'	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants																		
Budget Measured (1)	-2.46	*(8.4)*	-3.27	*(7.9)*	-3.27	*(7.1)*	-5.03	*(2.4)*	-3.28	*(6.0)*	-3.27	*(7.8)*	-3.91	*(2.2)*	-3.28	*(7.6)*	-3.28	*(6.5)*
Standard Measured (2)	-1.74	*(6.6)*	-2.53	*(6.6)*	-2.52	*(6.2)*	-3.85	*(2.2)*	-2.53	*(6.1)*	-2.52	*(6.5)*	-3.02	*(2.4)*	-2.53	*(6.5)*	-2.53	*(5.0)*
Local Flat (3)	-0.54	*(2.7)*	-1.37	*(3.8)*	-1.36	*(3.2)*	-1.09	*(2.1)*	-1.37	*(3.6)*	-1.36	*(3.7)*	-1.67	*(3.3)*	-1.37	*(3.8)*	-1.37	*(3.4)*
Extended Flat (4)	-0.74	*(1.1)*	-1.04	*(1.3)*	-1.04	*(1.3)*	-1.37	*(1.5)*	-1.04	*(1.4)*	-1.04	*(1.4)*	-1.10	*(1.2)*	-1.05	*(1.3)*	-1.04	*(1.4)*
Log Cost	-2.03	*(9.6)*	-2.68	*(8.2)*	-2.67	*(4.9)*	-3.24	*(3.1)*	-2.68	*(6.2)*	-2.67	*(8.2)*	-3.33	*(2.9)*	-2.68	*(8.1)*	-2.69	*(7.6)*
σ_1					0.02	*(0.1)*	2.77	*(1.8)*	0.03	*(0.0)*	0.03	*(0.3)*	0.76	*(0.4)*				
σ_2			0.13	*(0.3)*			3.27	*(1.6)*	0.14	*(0.1)*	0.14	*(0.3)*	0.70	*(0.3)*	0.11	*(0.2)*	0.10	*(0.2)*
σ_3			2.88	*(4.9)*	2.88	*(2.4)*			2.88	*(3.3)*	2.87	*(3.8)*	4.00	----	2.89	*(4.7)*	2.91	*(2.9)*
σ_4			0.04	*(0.1)*	0.04	*(0.1)*	1.14	*(0.5)*			0.04	*(0.1)*	0.11	*(0.1)*	0.12	*(0.2)*	0.07	*(0.1)*
σ_5			0.09	*(0.3)*	0.09	*(0.2)*	0.01	*(0.0)*	0.10	*(0.0)*			1.33	*(1.3)*	0.03	*(0.1)*	0.26	*(0.2)*
(Simul.) Log-Likelihood:	-477.56		-471.20		-471.20		-476.66		-471.20		-471.20		-471.42		-470.92		-471.39	

Nested & Cross-Nested Structures

In Table 11, the estimation results of various nested and cross-nested specifications are provided. Table 11a reports results for identified model structures (as can be verified by the rank condition). The best specification is model 3, in which the first two alternatives are nested, the last two alternatives are nested, and the third term has a heteroscedastic term. This provides a significant improvement in fit over the MNL specification shown in the first column, and also provides a better fit than the heteroscedastic models in Table 10. The poor fit for many of the nesting and cross-nesting specifications is due to the fact that the variance for alternative 3 is constrained to be in line with the other variances. The heteroscedastic models

indicated that it has a much higher variance, and when this was added to the nested and cross-nested models (see Table 11b) the fit improved dramatically.[15]

Table 11c provides results for the unidentified model in which the first two alternatives are nested and the last 3 alternatives are nested, and we attempt (incorrectly) to estimate both error parameters. The first model, estimated with 1,000 Halton draws, appears to be identified. However, the second model, estimated using different starting values, shows that this is not the case; it has an identical fit, but very different estimates of the error parameters. This is as expected, because only the sum of the variances $(s_1^2 + s_2^2)$ can be identified. The remaining columns show that it can take a very large number of draws to get the telltale sign of an unidentified model, the singular Hessian – in this case, 80,000 Halton draws. (Again, the actual number depends on the specification and the data.) Table 11d shows that the normalization for the 2 nest model is arbitrary. The table presents three normalizations resulting in identical fits where:

$$\{ 1, 1, 0, 0, 0 \} = \{ 0, 0, 2, 2, 2 \} = \{ 1, 1, 2, 2, 2 \text{ with } s_1 = s_2 \}.$$

Table 11: Telephone Model - Nested & Cross-Nested Error Structures

Table a: Identified Nesting & Cross-Nesting Error Structures

	Nested Structures										Cross-Nested Structures					
Specification*:	1, 1, 2, 2, 0		1, 1, 2, 2, 3		1, 1, 2, 3, 3		1, 1, 2, 3, 3		1, 1, 2, 2, 2 ($s_1 = s_2$)		1, 1, 1-2, 2, 2		1-2, 2-3, 3-4, 4-5, 5-6 (all s equal)		1-2, 2-3, 3-4, 4-5, 5-6 (all s equal)	
Draws:	1000 Halton		1000 Halton		1000 Halton		2000 Halton		1000 Halton		1000 Halton		1000 Halton		5000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants																
Budget Measured (1)	-3.63	*(5.0)*	-3.63	*(5.0)*	-3.79	*(5.4)*	-3.80	*(5.3)*	-3.80	*(5.7)*	-3.80	*(5.7)*	-2.83	*(2.4)*	-2.72	*(3.1)*
Standard Measured (2)	-2.85	*(4.3)*	-2.85	*(4.3)*	-3.00	*(4.6)*	-3.01	*(4.6)*	-3.01	*(4.9)*	-3.00	*(4.9)*	-1.90	*(3.1)*	-1.85	*(3.9)*
Local Flat (3)	-1.48	*(3.1)*	-1.48	*(3.1)*	-1.63	*(3.1)*	-1.64	*(3.1)*	-1.09	*(3.6)*	-1.09	*(3.5)*	-0.55	*(2.3)*	-0.54	*(2.4)*
Extended Flat (4)	-1.52	*(1.5)*	-1.52	*(1.5)*	-1.18	*(1.3)*	-1.18	*(1.3)*	-1.19	*(1.4)*	-1.19	*(1.4)*	-0.76	*(1.0)*	-0.75	*(1.0)*
Log Cost	-3.05	*(4.5)*	-3.05	*(4.5)*	-3.19	*(5.0)*	-3.20	*(5.0)*	-3.25	*(6.1)*	-3.25	*(6.1)*	-2.40	*(2.1)*	-2.29	*(2.6)*
σ1	1.32	*(1.1)*	1.32	*(1.1)*	1.55	*(1.5)*	1.55	*(1.6)*	2.16	*(3.0)*	0.01	*(0.8)*	0.65	*(0.6)*	0.53	*(0.6)*
σ2	3.02	*(2.9)*	3.02	*(2.9)*	3.34	*(2.9)*	3.37	*(2.8)*			3.04	*(3.0)*				
σ3			0.00	*(0.0)*	0.01	*(0.1)*	0.01	*(0.2)*								
(Simul.) Log-Likelihood:	-471.26		-471.26		-470.70		-470.64		-473.04		-473.05		-477.48		-477.51	

[15] Therefore, the problem identified earlier with the cross-nested *1, 1, 1-2, 2, 2* structure does not apply to this dataset. In fact, as shown by the models in Table 11c, alternative 3 has an even larger relative variance than the *1, 1, 1-2, 2, 2* structure provides.

Table b: Nesting / Cross-Nesting plus Heteroscedasticity (0, 0, 1, 0, 0)

	Combined Models					
Specification*:	2, 2, 1-3, 3, 3 ($s_2 = s_3$)		2, 2, 2-1-3, 3, 3		2-3, 3-4, 4-1-5, 5-6, 6-7 ($s_2 ... s_7$ equal)	
Draws:	1000 Halton		1000 Halton		1000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants						
Budget Measured (1)	-3.81	*(5.5)*	-3.80	*(5.3)*	-3.28	*(7.3)*
Standard Measured (2)	-3.02	*(4.7)*	-3.01	*(4.6)*	-2.53	*(6.3)*
Local Flat (3)	-1.64	*(3.1)*	-1.64	*(3.1)*	-1.37	*(3.5)*
Extended Flat (4)	-1.19	*(1.3)*	-1.18	*(1.3)*	-1.04	*(1.3)*
Log Cost	-3.21	*(5.2)*	-3.20	*(5.0)*	-2.68	*(8.0)*
σ_1	3.37	*(2.8)*	3.38	*(2.8)*	2.88	*(3.3)*
σ_2	1.11	*(1.6)*	0.03	*(0.3)*	0.09	*(0.2)*
σ_3			1.55	*(1.6)*		
(Simul.) Log-Likelihood:	-470.64		-470.69		-471.22	

Table c: Unidentified Nested Error Structures

Specification*:	1, 1, 2, 2, 2 (Unidentified - can only estimate $(\sigma_1^2+\sigma_2^2)$)											
Draws:	1000 Halton		1000 Halton		10000 Halton		40000 Halton		40000 'Random'		80000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants												
Budget Measured (1)	-3.80	*(5.7)*	-3.80	*(5.7)*	-3.80	*(5.7)*	-3.80	*(5.8)*	-3.81	*(5.7)*	-3.80	*n/a*
Standard Measured (2)	-3.01	*(4.9)*	-3.01	*(4.9)*	-3.01	*(4.9)*	-3.01	*(4.9)*	-3.01	*(4.8)*	-3.01	*n/a*
Local Flat (3)	-1.09	*(3.6)*	-1.09	*(3.6)*	-1.09	*(3.6)*	-1.09	*(3.6)*	-1.09	*(3.5)*	-1.09	*n/a*
Extended Flat (4)	-1.19	*(1.4)*	-1.19	*(1.4)*	-1.19	*(1.4)*	-1.19	*(1.4)*	-1.19	*(1.4)*	-1.19	*n/a*
Log Cost	-3.25	*(6.1)*	-3.25	*(6.1)*	-3.25	*(6.1)*	-3.25	*(6.1)*	-3.25	*(6.0)*	-3.25	*n/a*
σ_1	2.65	*(3.1)*	0.78	*(0.5)*	2.55	*(2.5)*	2.56	*(1.5)*	1.83	*(1.1)*	1.93	*n/a*
σ_2	1.51	*(2.2)*	2.95	*(3.3)*	1.67	*(3.8)*	1.68	*(0.4)*	2.45	*(1.9)*	2.36	*n/a*
$(\sigma_1^2+\sigma_2^2)^{1/2}$	*3.05*		*3.05*		*3.05*		*3.06*		*3.06*		*3.05*	
(Simul.) Log-Likelihood:	-473.02		-472.99		-473.02		-473.02		-472.95		-473.02	

Table d: Identical (Identified) Nested Error Structures

Specification*:	1, 1, 0, 0, 0		0, 0, 2, 2, 2		1, 1, 2, 2, 2 ($\sigma_1=\sigma_2$)			
Draws:	1000 Halton		1000 Halton		1000 Halton		2000 Halton	
Parameter	*Est*	*T-stat*	*Est*	*T-stat*	*Est*	*T-stat*	*Est*	*T-stat*
Altern. Specific constants								
Budget Measured (1)	-3.80	*(5.7)*	-3.80	*(5.7)*	-3.80	*(5.7)*	-3.80	*(5.8)*
Standard Measured (2)	-3.01	*(4.9)*	-3.01	*(4.9)*	-3.01	*(4.9)*	-3.01	*(4.9)*
Local Flat (3)	-1.09	*(3.6)*	-1.09	*(3.6)*	-1.09	*(3.6)*	-1.09	*(3.6)*
Extended Flat (4)	-1.19	*(1.4)*	-1.19	*(1.4)*	-1.19	*(1.4)*	-1.19	*(1.4)*
Log Cost	-3.25	*(6.1)*	-3.25	*(6.1)*	-3.25	*(6.1)*	-3.25	*(6.1)*
σ_1	3.05	*(3.0)*			2.16	*(3.0)*	2.15	*(3.0)*
σ_2			3.05	*(3.0)*	2.16	---	2.15	---
$(\sigma_1^2+\sigma_2^2)^{1/2}$	*3.05*		*3.05*		*3.05*		*3.04*	
(Simul.) Log-Likelihood:	-473.02		-473.03		-473.04		-473.01	

* the specification lists the factors (and sigmas) that apply to each of the five alternatives

Random Parameters

We also considered unobserved taste heterogeneity for the parameter on log of cost. Since the parameter has a sign constraint, a lognormal distribution is used. (Draws from a lognormal distribution are generated by exponentiating draws taken from a normal distribution.) The results are shown in Table 12. The first model shows that when there are no other covariance parameters specified, the heterogeneity on log cost is insignificant. However, the second model shows that heterogeneity does add slightly to the explanatory power of the best nested model as specified in Table 11a. The remaining 4 models report specifications with both heterogeneity and taste variation. While the rank and order conditions suggest that a model with 4 heteroscedastic parameters and the lognormal parameter is identified, the estimation results show that there is a multicollinearity problem. Note that when only 200 pseudo-random draws are used, this model appears, incorrectly, to be identified.

Table 12: Telephone Model - Taste Variation, Lognormal Parameter for Log(Cost)

Specification*:	Taste Variation		1,1,2,3,3 & Taste Variation				1,2,3,4,5 & Taste Variation							
Draws:	1000 Halton		1000 Halton		2000 Halton		1000 Halton		200 'Random'		1000 Halton		1000 Halton	
Parameter	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Altern. Specific constants														
Budget Measured (1)	-2.46	*(8.2)*	-3.48	*(5.7)*	-3.50	*(4.3)*	-24.20	*n/a*	-4.06	*(2.6)*	-30.36	*n/a*	-26.84	*n/a*
Standard Measured (2)	-1.74	*(6.5)*	-2.68	*(4.7)*	-2.70	*(3.5)*	-16.75	*n/a*	-3.06	*(2.8)*	-22.03	*n/a*	-19.41	*n/a*
Local Flat (3)	-0.54	*(2.7)*	-1.44	*(3.1)*	-1.45	*(2.7)*	-7.57	*n/a*	-1.57	*(2.4)*	-10.72	*n/a*	-9.77	*n/a*
Extended Flat (4)	-0.74	*(1.0)*	-0.98	*(1.1)*	-0.98	*(1.1)*	-3.33	*n/a*	-1.07	*(1.1)*	-5.11	*n/a*	-4.75	*n/a*
Log Cost **	-2.03	*(9.6)*	-3.17	*(5.6)*	-3.18	*(5.1)*	-23.30	*n/a*	-3.69	*(2.7)*	-28.38	*n/a*	-26.02	*n/a*
σ Log Cost **	0.00	*(0.1)*	1.18	*(1.1)*	1.16	*(1.0)*	18.39	*n/a*	1.65	*(1.4)*	18.85	*n/a*	18.54	*n/a*
σ_1			0.40	*(0.1)*	0.50	*(0.1)*	12.38	*n/a*	1.00	*(0.6)*	13.72	*n/a*	12.19	*n/a*
σ_2			3.56	*(3.0)*	3.58	*(3.0)*	9.06	*n/a*	0.72	*(0.5)*	11.34	*n/a*	9.02	*n/a*
σ_3			0.05	*(0.8)*	0.01	*(0.1)*	24.50	*n/a*	4.13	*(2.3)*	30.45	*n/a*	28.96	*n/a*
σ_4							0.49	*n/a*						
σ_5							0.88	*n/a*	0.24	*(0.6)*	1.26	*n/a*		
Log Likelihood (simul.):	-477.56		-470.36		-470.28		-469.15		-470.74		-468.69		-469.47	

** the mean and standard deviation of the lognormal are reported

Summary of Telephone Data Models

By far the most important part of the error structure for the telephone dataset is that the Local Flat Alternative (3) has a significantly higher variance than the other alternatives. Note that a simple heteroscedastic model outperforms the most obvious nested structure in which the measured alternatives are nested together and the flat alternatives are nested together. Marginal improvements can be achieved by incorporating nesting, cross-nesting or taste variation as long as alternative 3 is allowed a free variance. While this dataset served its purpose in highlighting specification and identification issues, one would ideally like to estimate such logit kernel models with larger datasets.

Conclusion

In this paper we presented general rules for specification, identification, and estimation via maximum simulated likelihood for the logit kernel model. We presented guidelines for examining identification and normalization, which consisted of three conditions: order, rank, and positive definiteness. The positive definiteness condition is not an issue for probit models. However, as the heteroscedastic case highlights, it can have important consequences for logit kernel. We emphasized that identification must be examined on a case-by-case basis, and that it is not necessarily intuitive. Furthermore, given the fact that simulation has a tendency to mask identification problems, it becomes even more critical that identification is well understood.

We discussed in detail the specification and identification of many of the special cases, all within a general factor analytical framework, including:

Heteroscedasticity:	F_n diagonal (fixed) ; T diagonal.
Nesting (Cross-Nesting):	$F_n F_n'$ block-diagonal (fixed) ; T diagonal.
Error Components:	F_n fixed to 0/1 ; T (usually) diagonal.
Factor Analytic:	F_n unknown ; T triangular.
Autoregressive Process:	F_n moving average form of a GAR(1) process ; T diagonal.
Random parameters:	F_n a function of explanatory variables (fixed) ; T triangular.

Just as there are well-known standard rules for identification for the systematic parameters in a multinomial logit, we aimed to develop identification rules for the disturbance parameters of the logit kernel model. There are critical differences between the identification of these parameters and the identification of their counterparts in both the systematic portion of the utility as well as their counterparts in a probit model. The following summarizes these identification rules:

Heteroscedasticity

$J = 2$ alternatives:	0 parameters identified.
$J \geq 3$ alternatives:	$J-1$ parameters identified & must constrain the minimum variance term to 0.

Nesting

$M = 2$ nests:	$M-1$ parameters identified & normalization is arbitrary.
$M \geq 3$ nests:	M parameters identified.

Random parameters

Beyond the specific rules listed below, can estimate as many random parameters as the data will support.

Alternate-specific variables

Rules for heteroscedasticity, nesting, and error components apply.

Categorical variables with independently distributed parameters

$M = 2$ categories:	$M - 1$ parameters identified & normalization is arbitrary.
$M \geq 3$ or more categories:	M parameters identified. (Includes a binary categorical variable that does not enter all utilities.)

Characteristics of the Decision-maker with independently distributed parameters

Interacts with alternative-specific constants:	Analogous to the heteroscedastic case: $J - 1$ parameters identified & must constrain the minimum variance term to 0.
Interacts with nest-specific constants:	Analogous to nested case:
$M = 2$ nests:	$M - 1$ parameters identified.
$M \geq 3$ nests:	M parameters identified.

Our objectives were that through examination of the special cases we would be able to establish some identification and specification rules, and also highlight some of the broad themes and provide tools for uncovering other potential issues pertaining to logit kernel models. Clearly there are numerous identification issues that are not covered by the above list. Therefore, models have to be examined on a case-by-case basis. For the alternative-specific portion of the disturbance, it is recommended that the rank and order conditions be programmed into the estimation program. When the positive definiteness condition comes into play, it is recommended to examine the problem analytically, where possible, or empirically (by investigating various normalizations). For random parameter models, it is recommended to use the above identification rules as guidelines, and then empirically establish identification by (1) verifying that the parameter estimates are stable as the number of draws are increased and (2) checking that the Hessian is non-singular at the convergence point.

One of the most important points of the paper is that there are critical aspects to the logit kernel specification that are often overlooked in the literature. It must be remembered that this is a relatively new methodology, and there are numerous aspects that warrant further research, including:

- More testing and experience with applications,
- Further exploration of identification and normalization issues,
- Continued compilation and analysis of special cases and rules of identification,
- Better understanding of the impact on analysis of different factor specifications (particularly since often several factor specification will provide similar fit to the data),
- Investigation of analogous specifications estimated via different methods (for example, logit kernel versus probit, nested logit, cross-nested logit, heteroscedastic extreme value, etc.)
- Additional comparisons with GHK and other smooth simulators, and

- Further examination of Halton draws as well as other pseudo- and quasi-random drawing methods.

Finally, we also may need to look at modifying the specification of the logit kernel model to alleviate some of the complications. One of the issues with the logit kernel specification is that while pure logit is a special case of the model, pure probit is not. Our analysis assumes that it is acceptable to include the Gumbel term in the model. However, the Gumbel term may, in fact, have no business being in the model. For this reason, we would ideally want to specify and estimate the model in a way that allows the Gumbel term to disappear. Conceptually, such a model could be specified as a linear combination of the two error terms, so Equation (4) (assuming a universal choice set) would become:

$$U_n = X_n b + \sqrt{(g/m^2)(1-l^2)}F_n T z_n + l n_n \ ,$$

where l is an unknown parameter. The covariance of the model is then a linear combination of the two covariance matrices:

$$\text{cov}(U_n) = \left((1-l^2)F_n T T' F_n' + l^2 I_J\right)\left(g/m^2\right) \ .$$

Conceptually this Combined Logit-Probit (CLP) specification is an appealing model. Note that a strict application of the order and rank conditions lead to the conclusion that the model is not identified. However, as we described in the section on identification, the slight difference between the Gumbel and Normal distributions makes the model identified (albeit, nearly singular).

To summarize, the logit kernel formulation has a tremendous amount of potential, because it can replicate any desirable error structure and is straightforward to estimate via maximum simulated likelihood. However, it also has some issues that must be understood for proper specification. As increased computational power and readily available software open up these techniques for widespread use, it is a critical time to understand and address the nuances of the logit kernel model.

Appendix

Normalization of Unrestricted Probit and Logit Kernel Covariance Structures

This appendix examines the normalization of unrestricted probit and logit kernel models. The important point is that while the normalization of pure probit leads to straightforward scale shifts of all of the parameter estimates, this is not the case for logit kernel.

Case 1: Probit with 4 Alternatives

The unrestricted four alternative probit model written in differenced form has the error structure Tz_n, where:

$$\mathrm{T} = \begin{bmatrix} a_{11}/\tilde{m} & 0 & 0 \\ a_{21}/\tilde{m} & a_{22}/\tilde{m} & 0 \\ a_{31}/\tilde{m} & a_{32}/\tilde{m} & a_{33}/\tilde{m} \end{bmatrix}$$

Note that we use a's instead of s's since these aren't variance terms. Also $\tilde{m}$ is the scale of the probit model (i.e., not the traditional Gumbel m).

The covariance structure is then (using new notation):

$$\mathrm{TT'}\ \textit{theoretical}: \begin{bmatrix} (a_{11}^2)/\tilde{m}^2 & & \\ (a_{11}a_{21})/\tilde{m}^2 & (a_{21}^2 + a_{22}^2)/\tilde{m}^2 & \\ (a_{11}a_{31})/\tilde{m}^2 & (a_{21}a_{31} + a_{22}a_{32})/\tilde{m}^2 & (a_{31}^2 + a_{32}^2 + a_{33}^2)/\tilde{m}^2 \end{bmatrix}$$

A normalization must be made in order to achieve identification. Normalizing $a_{33} = a_{ff}^N$, and noting the unknown parameters as a and m, then the normalized covariance structure is:

$$\mathrm{TT'}\ \textit{normalized}: \begin{bmatrix} \left(a_{11}^N\right)^2/\tilde{m}_N^2 & & \\ \left(a_{11}^N a_{21}^N\right)/\tilde{m}_N^2 & \left(\left(a_{21}^N\right)^2 + \left(a_{22}^N\right)^2\right)/\tilde{m}_N^2 & \\ \left(a_{11}^N a_{31}^N\right)/\tilde{m}_N^2 & \left(a_{21}^N a_{31}^N + a_{22}^N a_{32}^N\right)/\tilde{m}_N^2 & \left(\left(a_{31}^N\right)^2 + \left(a_{32}^N\right)^2 + \left(a_{ff}^N\right)^2\right)/\tilde{m}_N^2 \end{bmatrix}$$

Setting $TT'\ \textit{normalized} = TT'\ \textit{theoretical}$, leads to the following equations:

$$\left(a_{11}^N\right)^2/\tilde{m}_N^2 = \left(a_{11}\right)^2/\tilde{m}^2$$

$$\left(a_{11}^N a_{21}^N\right)/\tilde{m}_N^2 = \left(a_{11}a_{21}\right)/\tilde{m}^2$$

$$\left(a_{11}^{N}a_{31}^{N}\right)/\tilde{m}_N^2=\left(a_{11}a_{31}\right)/\tilde{m}^2$$

$$\left(\left(a_{21}^{N}\right)^2+\left(a_{22}^{N}\right)^2\right)/\tilde{m}_N^2=\left(\left(a_{21}\right)^2+\left(a_{22}\right)^2\right)/\tilde{m}^2$$

$$\left(a_{21}^{N}a_{31}^{N}+a_{22}^{N}a_{32}^{N}\right)/\tilde{m}_N^2=\left(a_{21}a_{31}+a_{22}a_{32}\right)/\tilde{m}^2$$

$$\left(\left(a_{31}^{N}\right)^2+\left(a_{32}^{N}\right)^2+\left(a_{ff}^{N}\right)^2\right)/\tilde{m}_N^2=\left(\left(a_{31}\right)^2+\left(a_{32}\right)^2+\left(a_{33}\right)^2\right)/\tilde{m}^2$$

And solving for each of the unknown parameters in the normalized model leads to:

Solution: $\left(a_{11}^{N}\right)^2=\left(a_{11}\right)^2\frac{\tilde{m}_N^2}{\tilde{m}^2}$ → $a_{11}^{N}=a_{11}\frac{\tilde{m}_N}{\tilde{m}}$

$a_{21}^{N}=\frac{a_{11}a_{21}}{a_{11}^{N}}\frac{\tilde{m}_N^2}{\tilde{m}^2}$ → $a_{21}^{N}=a_{21}\frac{\tilde{m}_N}{\tilde{m}}$

$a_{31}^{N}=\frac{a_{11}a_{31}}{a_{11}^{N}}\frac{\tilde{m}_N^2}{\tilde{m}^2}$ → $a_{31}^{N}=a_{31}\frac{\tilde{m}_N}{\tilde{m}}$

$\left(a_{22}^{N}\right)^2=\left(\left(a_{21}\right)^2+\left(a_{22}\right)^2\right)\frac{\tilde{m}_N^2}{\tilde{m}^2}-\left(a_{21}^{N}\right)^2$ → $a_{22}^{N}=a_{22}\frac{\tilde{m}_N}{\tilde{m}}$

$a_{32}^{N}=\frac{1}{a_{22}^{N}}\left(\left(a_{21}a_{31}+a_{22}a_{32}\right)\frac{\tilde{m}_N^2}{\tilde{m}^2}-a_{21}^{N}a_{31}^{N}\right)$ → $a_{32}^{N}=a_{32}\frac{\tilde{m}_N}{\tilde{m}}$

$\frac{\left(a_{31}^{N}\right)^2+\left(a_{32}^{N}\right)^2+\left(a_{ff}^{N}\right)^2}{\tilde{m}_N^2}=\frac{\left(a_{31}\right)^2+\left(a_{32}\right)^2+\left(a_{33}\right)^2}{\tilde{m}^2}$ → $\tilde{m}_N=\frac{a_{ff}^{N}}{a_{33}}\tilde{m}$

Therefore, for probit, the normalization just scales all of the parameters, and any positive normalization is acceptable.

Case 2: Logit Kernel with 4 Alternatives

Now, we will show that the equivalent logit kernel case is not so straightforward. Following the same process, the covariance matrix of utility differences for the four alternative unrestricted logit kernel model is:

$$TT'+G\ theoretical:\ \begin{bmatrix}\left(a_{11}^2+2g\right)/m^2 & & \\ \left(a_{11}a_{21}+g\right)/m^2 & \left(a_{21}^2+a_{22}^2+2g\right)/m^2 & \\ \left(a_{11}a_{31}+g\right)/m^2 & \left(a_{21}a_{31}+a_{22}a_{32}+g\right)/m^2 & \left(a_{31}^2+a_{32}^2+a_{33}^2+2g\right)/m^2\end{bmatrix}$$

Imposing the normalization $a_{33} = a_{ff}$ leads to:

$TT' + G$

normalized :

$$\begin{bmatrix} \left(\left(a_{11}^{N}\right)^2 + 2g\right)/m_N^2 & & \\ \left(a_{11}^{N}a_{21}^{N} + g\right)/m_N^2 & \left(\left(a_{21}^{N}\right)^2 + \left(a_{22}^{N}\right)^2 + 2g\right)/m_N^2 & \\ \left(a_{11}^{N}a_{31}^{N} + g\right)/m_N^2 & \left(a_{21}^{N}a_{31}^{N} + a_{22}^{N}a_{32}^{N} + g\right)/m_N^2 & \left(\left(a_{31}^{N}\right)^2 + \left(a_{32}^{N}\right)^2 + \left(a_{ff}^{N}\right)^2 + 2g\right)/m_N^2 \end{bmatrix}$$

Setting the normalized covariance structure to the normalized structure leads to the following equations (the C notation is just to clean up the math later on):

$$\left(\left(a_{11}^{N}\right)^2 + 2g\right)/m_N^2 = \left(a_{11}^{2} + 2g\right)/m^2 \equiv C_1$$

$$\left(a_{11}^{N}a_{21}^{N} + g\right)/m_N^2 = \left(a_{11}a_{21} + g\right)/m^2 \equiv C_2$$

$$\left(a_{11}^{N}a_{31}^{N} + g\right)/m_N^2 = \left(a_{11}a_{31} + g\right)/m^2 \equiv C_3$$

$$\left(\left(a_{21}^{N}\right)^2 + \left(a_{22}^{N}\right)^2 + 2g\right)/m_N^2 = \left(a_{21}^{2} + a_{22}^{2} + 2g\right)/m^2 \equiv C_4$$

$$\left(a_{21}^{N}a_{31}^{N} + a_{22}^{N}a_{32}^{N} + g\right)/m_N^2 = \left(a_{21}a_{31} + a_{22}a_{32} + g\right)/m^2 \equiv C_5$$

$$\left(\left(a_{31}^{N}\right)^2 + \left(a_{32}^{N}\right)^2 + \left(a_{ff}^{N}\right)^2 + 2g\right)/m_N^2 = \left(a_{31}^{2} + a_{32}^{2} + a_{33}^{2} + 2g\right)/m^2 \equiv C_6$$

And solving for the estimated parameters in the normalized model leads to:

$$\left(a_{11}^{N}\right)^2 = C_1 m_N^2 - 2g$$

$$a_{21}^{N} = \frac{C_2 m_N^2 - g}{\sqrt{C_1 m_N^2 - 2g}}$$

$$a_{31}^{N} = \frac{C_3 m_N^2 - g}{\sqrt{C_1 m_N^2 - 2g}}$$

$$\left(a_{22}^{N}\right)^2 = C_4 m_N^2 - 2g - \frac{\left(C_2 m_N^2 - g\right)^2}{C_1 m_N^2 - 2g}$$

$$a_{32}^{N} = \frac{C_5 m_N^2 - g - \frac{\left(C_2 m_N^2 - g\right)\left(C_3 m_N^2 - g\right)}{\left(C_1 m_N^2 - 2g\right)}}{\sqrt{C_4 m_N^2 - 2g - \frac{\left(C_2 m_N^2 - g\right)^2}{\left(C_1 m_N^2 - 2g\right)}}}$$

$$m_N^2 = \frac{1}{C_6}\left(\frac{\left(C_3 m_N^2 - g\right)^2}{C_1 m_N^2 - 2g} + \frac{\left(C_5 m_N^2 - g - \frac{\left(C_2 m_N^2 - g\right)\left(C_3 m_N^2 - g\right)}{\left(C_1 m_N^2 - 2g\right)} \right)^2}{C_4 m_N^2 - 2g - \frac{\left(C_2 m_N^2 - g\right)^2}{\left(C_1 m_N^2 - 2g\right)}} + \left(a_{ff}^{N}\right)^2 + 2g \right)$$

Unlike probit, this is not a simple scale shift, i.e., the model must adjust to the normalization in complex, non-linear ways. Furthermore, it is not clear from these equations what the potential restrictions are on the normalization.

Empirical results exploring the normalization issue for a 4 alternative unrestricted logit kernel model are shown in Table A-13. The table includes estimation results using two different synthetic datasets (the true parameters vary across the datasets). There are 4 alternatives, and the model is specified with three alternative specific dummy parameters, one explanatory variable, and then an unrestricted covariance structure. The final column in the first table shows that, under some circumstances, restricting a_{22} to zero is an invalid normalization. The remaining estimation results suggest that restricting a_{33} to zero is a valid normalization regardless of the true parameter estimates. However, these results are not conclusive.

Table A-13: Normalization of Unrestricted Logit Kernel Model
(2 Synthetic Datasets; 4 Alternatives; 10,000 Observations; 1,000 Halton draws)

			Unidentified		Valid Normalizations						Invalid Normalization	
	Parameter	True	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Systematic	Alt. 1 dummy	1.0	1.38	*(2.8)*	0.93	*(11.5)*	1.02	*(11.8)*	1.31	*(12.1)*	0.76	*(12.4)*
	Alt. 2 dummy	1.0	1.28	*(2.8)*	0.85	*(10.2)*	0.94	*(10.3)*	1.21	*(10.5)*	0.67	*(11.0)*
	Alt. 3 dummy	0.0	0.03	*(0.3)*	0.04	*(0.5)*	0.04	*(0.5)*	0.03	*(0.3)*	0.02	*(0.3)*
	Variable 1	-1.0	-1.37	*(2.9)*	-0.93	*(23.5)*	-1.02	*(25.6)*	-1.30	*(28.8)*	-0.76	*(38.5)*
Distributions	α_{11}	**2.0**	3.16	*(2.1)*	1.60	*(9.1)*	1.96	*(11.3)*	2.94	*(15.7)*	-0.34	*(3.1)*
	α_{21}	1.0	1.75	*(2.1)*	0.86	*(3.7)*	1.09	*(4.7)*	1.63	*(6.2)*	-2.39	*(15.1)*
	α_{31}	2.0	2.86	*(2.7)*	2.01	*(9.1)*	2.13	*(9.4)*	2.70	*(10.9)*	-1.12	*(8.9)*
	α_{22}	**3.0**	4.62	*(2.6)*	2.89	*(14.6)*	3.25	*(16.2)*	4.35	*(19.1)*	**0.00**	---
	α_{32}	1.0	1.79	*(2.5)*	1.16	*(6.9)*	1.27	*(7.8)*	1.69	*(9.3)*	-0.01	*(0.0)*
	α_{33}	**1.0**	2.20	*(1.7)*	**0.00**	---	**1.00**	---	**2.00**	---	0.00	*(0.0)*
	(Simul.) Log-Likelihood:		-7973.176		-7974.867		-7973.843		-7973.187		-7998.768	

		True	Unidentified		Valid Normalizations							
	Parameter	Value	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*	*Est*	*t-stat*
Systematic	Alt. 1 dummy	1.0	0.94	*(8.5)*	0.92	*(9.4)*	0.92	*(9.4)*	0.92	*(9.4)*	0.94	*(8.9)*
	Alt. 2 dummy	1.0	0.95	*(8.2)*	0.93	*(9.1)*	0.92	*(9.1)*	0.93	*(9.1)*	0.96	*(8.4)*
	Alt. 3 dummy	0.0	0.18	*(1.5)*	0.17	*(1.5)*	0.17	*(1.5)*	0.17	*(1.5)*	0.18	*(1.5)*
	Variable 1	-1.0	-0.86	*(17.1)*	-0.85	*(31.8)*	-0.85	*(31.8)*	-0.85	*(31.6)*	-0.87	*(27.7)*
Distributions	α_{11}	**2.0**	1.43	*(5.3)*	1.37	*(6.9)*	1.37	*(6.9)*	-1.38	*(7.0)*	1.45	*(7.2)*
	α_{21}	1.0	0.79	*(4.6)*	0.76	*(5.0)*	0.76	*(5.0)*	-0.76	*(5.0)*	0.80	*(5.3)*
	α_{31}	2.0	2.53	*(3.9)*	2.50	*(3.8)*	2.48	*(3.8)*	-2.50	*(3.9)*	2.56	*(3.9)*
	α_{22}	**1.0**	0.39	*(0.9)*	-0.22	*(1.6)*	-0.22	*(1.6)*	-0.25	*(1.6)*	0.43	*(1.9)*
	α_{32}	1.0	3.19	*(1.2)*	-4.87	*(14.2)*	-4.78	*(13.8)*	-4.46	*(12.0)*	3.03	*(5.4)*
	α_{33}	**6.0**	3.83	*(1.5)*	**0.00**	---	**1.00**	---	**2.00**	---	**4.00**	---
	(Simul.) Log-Likelihood:		-8983.725		-8984.556		-8984.62		-8984.222		-8983.735	

Case 3: Logit Kernel with 3 Alternatives

The three alternative logit kernel case is a bit easier to compute. Following the same process as above:

$$T: \quad \begin{bmatrix} a_{11}/m & 0 \\ a_{21}/m & a_{22}/m \end{bmatrix}$$

$$TT'+G \quad \textit{theoretical}: \quad \begin{bmatrix} \left((a_{11})^2 + 2g\right)/m^2 & \\ \left(a_{11}a_{21}+g\right)/m^2 & \left((a_{21})^2 + (a_{33})^2 + 2g\right)/m^2 \end{bmatrix}$$

$$TT'+G \quad \text{normalized}: \quad \begin{bmatrix} \left(\left(a_{11}^N\right)^2+2g\right)/m_N^2 & \\ \left(a_{11}^N a_{21}^N+g\right)/m_N^2 & \left(\left(a_{21}^N\right)^2+\left(a_{ff}^N\right)^2+2g\right)/m_N^2 \end{bmatrix}$$

$$\rightarrow \quad \left(\left(a_{11}^N\right)^2+2g\right)/m_N^2=\left(\left(a_{11}\right)^2+2g\right)/m^2\equiv C_1$$

$$\left(a_{11}^N a_{21}^N+g\right)/m_N^2=\left(a_{11}a_{21}+g\right)/m^2\equiv C_2$$

$$\left(\left(a_{21}^N\right)^2+\left(a_{ff}^N\right)^2+2g\right)/m_N^2=\left(\left(a_{21}\right)^2+\left(a_{33}\right)^2+2g\right)/m^2\equiv C_3$$

Solution

$$a_{11}^N=\sqrt{C_1m_N^2-2g} \qquad \text{… or …} \qquad a_{11}^N=\frac{C_2m_N^2-g}{\sqrt{C_3m_N^2-a_{33}^2-2g}}$$

$$a_{21}^N=\frac{C_2m_N^2-g}{\sqrt{C_1m_N^2-2g}} \qquad \text{… or …} \qquad a_{21}^N=\sqrt{C_3m_N^2-\left(a_{ff}^N\right)^2-2g}$$

$$m_N^2=\frac{\left(\begin{array}{l}-\left(2g(C_1-C_2+C_3)+\left(a_{ff}^N\right)^2C_1\right)\\ \qquad \pm\sqrt{\left(2g(C_1-C_2+C_3)+\left(a_{ff}^N\right)^2C_1\right)-4\left(C_2^2-C_1C_3\right)\left(-2g\left(a_{ff}^N\right)^2-3g^2\right)}\end{array}\right)}{2\left(C_2^2-C_1C_3\right)}$$

Here, the restrictions are

$$\left(2g(C_1-C_2+C_3)+\left(a_{ff}^N\right)^2C_1\right)^2-4\left(C_2^2-C_1C_3\right)\left(-2g\left(a_{ff}^N\right)^2-3g^2\right)\geq 0\ ,$$

$$m^2>0\ ,$$

$$C_1m^2-2g>0 \ \text{… or …}\ C_3m^2-\left(a_{ff}^N\right)^2-2g>0\ ,$$

$$a_{11}^2\left(a_{21}^2+\left(a_{ff}^N\right)^2\right)-\left(a_{11}^2a_{21}^2\right)\geq 0\ ,\ \text{where}\ a_{11}=f(a_{ff}^N)\ \text{and}\ a_{21}=f(a_{ff}^N)\ ,$$

and only 1 of the two possible m^2 satisfies the conditions.

Again, it's not clear in which cases these restrictions become limiting. Our empirical tests suggests that the normalization of the lowest diagonal element in the cholesky matrix is, in fact, a valid normalization regardless of the true parameters (unlike, for example, the heteroscedastic case).

References

Anselin, L. (1989) *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers.

Ben-Akiva, M. and D. Bolduc (1996) "Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure", working paper, Massachusetts Institute of Technology.

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

Berndt, E.R., B.H. Hall, R.E. Hall, and J.A. Hausman (1974) "Estimation and Inference in Nonlinear Structural Models", *Annals of Economic & Social Measurement* **3**, 653-665.

Bhat, C.R. (1995) "A Heteroscedastic Extreme Value Model of Intercity Travel Mode Choice", *Transportation Research B* **29(6)**, 471-483.

Bhat, C.R. (1997) "Accommodating Flexible Substitution Patterns in Multi-dimensional Choice Modeling: Formulation and Application to Travel Mode and Departure Time Choice", *Transportation Research B* **32(7)**, 455-466.

Bhat, C.R. (1998) "Accommodating Variations in Responsiveness to Level-of-Service Measures in Travel Mode Choice Modeling", *Transportation Research A* **32(7)**, 495-507.

Bhat, C.R. (2000) "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model", forthcoming, *Transportation Research*.

Bolduc, D. (1992) "Generalized Autoregressive Errors in the Multinomial Probit Model", *Transportation Research B* **26(2)**, 155-170.

Bolduc, D. and M. Ben-Akiva (1991) "A Multinominal Probit Formulation for Large Choice Sets", Proceedings of the 6th International Conference on Travel Behaviour **2**, 243-258.

Bolduc, D., B. Fortin and M.A. Fournier (1996) "The Impact of Incentive Policies to Influence Practice Location of General Practioners: A Multinomial Probit Analysis", *Journal of Labor Economics* **14**, 703-732.

Börsch-Supan, A. and V. Hajivassiliou (1993) "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models", *Journal of Econometrics* **58**, 347-368.

Boyd, J.H. and R.E. Mellman (1980) "The Effect of Fuel Economy Standards on the U.S. Automotive Market: An Hedonic Demand Analysis", *Transportation Research A* **14**, 367-378.

Brownstone, D., D.S. Bunch and K. Train (2000) "Joint Mixed Logit Models of Stated and Revealed Preferences for Alternative-fuel Vehicles", *Transportation Research B* **34**, 315-338.

Brownstone, D. and K. Train (1999) "Forecasting New Product Penetration with Flexible Substitution Patterns", *Journal of Econometrics* **89**, 109-129.

Bunch, D.A. (1991) "Estimability in the Multinomial Probit Model", *Transportation Research B* **25**, 1-12.

Cardell, N.S. and F.C. Dunbar (1980) "Measuring the Societal Impacts of Automobile Downsizing" *Transportation Research A* **14**, 423-434.

Case, A. (1991) "Spatial Correlation in Household Demand", *Econometrica* **59(4)**, 953-965.

Cliff, A.D, and J.K. Ord (1981) *Spatial Processes, Models and Application*, Pion, London.

Dansie, B.R. (1985) "Parameter Estimability in the Multinomial Probit Model", *Transportation Research B* **19(6)**, 526-528.

Dennis, J.E. and R.B. Schnabel (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs.

Goett, A., K. Hudson and K. Train (2000) "Customers' Choice Among Retail Energy Suppliers: The Willingness-to-Pay for Service Attributes", working paper, AAG Associates and University of California at Berkeley.

Gönül, F. and K. Srinivasan (1993) "Modeling Multiple Sources of Heterogeneity in Multinomial Logit Models: Methodological and Managerial Issues", *Marketing Science* **12(3)**, 213-229.

Greene, W.H. (2000) *Econometric Analysis Fourth Edition*, Prentice Hall, Upper Saddle River, New Jersey.

Hajivassiliou, V. and P. Ruud (1994) "Classical Estimation Methods for LDV Models using Simulation", *Handbook of Econometrics* **IV**, R. Engle and D. McFadden, Eds., 2384-2441.

Lerman and Manski (1981) "On the Use of Simulated Frequencies to Approximate Choice Probabilities", *Structural Analysis of Discrete Data with Econometric Applications*, C.F. Manski and D. McFadden, Eds., The MIT Press, Cambridge, Massachusetts, 305-319.

Louviere, J.J., D.A. Hensher and J.D. Swait (2000) *Stated Choice Methods: Analysis and Application*, Cambridge University Press.

McFadden, D. (1984) "Econometric Analysis of Qualitative Response Models", *Handbook of Econometrics* **II**, Z. Friliches and M.D. Intriligator, Eds., Elsevier Science Publishers.

McFadden, D. (1989) "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration", *Econometrica* **57(5)**, 995-1026.

McFadden, D. and K. Train (2000) "Mixed MNL Models for Discrete Response", *Journal of Applied Econometrics* **15(5)**, 447-470.

Mehndiratta, R.M. and M. Hansen (1997) "Analysis of Discrete Choice Data with Repeated Observations: Comparison of Three Techniques in Intercity Travel Case", *Transportation Research Record* **1607**, 69-73.

Newey, W. and D. McFadden (1994) "Large Sample Estimation and Hypothesis Testing", *Handbook of Econometrics* **IV**, R. Engle and D. Mcfadden, Eds., 2111-2245.

Pakes, A. and D. Pollard (1989) "Simulation and the Asymptotics of Optimization Estimators", *Econometrica* **57(5)**, 1027-1057.

Revelt, D. and K. Train (1998) "Mixed Logit with Repeated Choices: Households' Choice of Appliance Efficiency Level", *Review of Economics and Statistics* **80(4)**, 647-657.

Revelt, D. and K. Train (1999) "Customer-Specific Taste Parameters and Mixed Logit", working paper, University of California at Berkeley.

Srinivasan, K.K. and H.S. Mahmassani (2000) "Dynamic Kernel Logit Model for the Analysis of Longitudinal Discrete Choice Data: Properties and Computational Assessment", presented at the International Association of Travel Behavior Research (IATBR) Conference, Gold Coast, Queensland, Australia.

Steckel, J.H. and W.R. Vanhonacker (1988) "A Heterogeneous Conditional Logit Model of Choice", *Journal of Business & Economic Statistics* **6(3)**, 391-398.

Stern, S. (1992) "A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models", *Econometrica* **60(4)**, 943-952.

Train, K.E. (1998) "Recreational Demand Models with Taste Differences Over People", *Land Economics* **74(2)**, 230-239.

Train, K. (1999) "Halton Sequences for Mixed Logit", working paper, University of California at Berkeley.

Train, K., D. McFadden and M. Ben-Akiva (1987) "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices", *Rand Journal of Economics* **18(1)**, 109-123.

Walker, J.L. (2001) *Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables*, Ph.D. Dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.