Estimability in the Multinomial Probit Model

David S. Bunch

February 1991
Reprint No. 71

The University of California
Transportation Center

University of California
Berkeley, CA 94720

The University of California Transportation Center

The University of California Transportation Center (UCTC) is one of ten regional units mandated by Congress and established in Fall 1988 to support research, education, and training in surface transportation. The UC Center serves federal Region IX and is supported by matching grants from the U.S. Department of Transportation, the California State Department of Transportation (Caltrans), and the University.

Based on the Berkeley Campus, UCTC draws upon existing capabilities and resources of the Institutes of Transportation Studies at Berkeley, Davis, and Irvine; the Institute of Urban and Regional Development at Berkeley; the Graduate School of Architecture and Urban Planning at Los Angeles; and several academic departments at the Berkeley, Davis, Irvine, and Los Angeles campuses. Faculty and students on other University of California campuses may participate in Center activities. Researchers at other universities within the region also have opportunities to collaborate on selected studies. Currently faculty at California State University, Long Beach, and at Arizona State University, Tempe, are active participants.

UCTC's educational and research programs are focused on strategic planning for improving metropolitan accessibility, with emphasis on the special conditions in Region IX. Particular attention is directed to strategies for using transportation as an instrument of economic development, while also accommodating to the region's persistent expansion and while maintaining and enhancing the quality of life there.

The Center distributes reports on its research in working papers, monographs, and in reprints of published articles. For a list of publications in print, write to the address below.

University of California
Transportation Center

108 Naval Architecture Building
Berkeley, California 94720
Tel: 415/643-7378
FAX: 415/643-5456

Authors of papers reporting on UCTC-sponsored research are solely responsible for their content. This research was supported by the U.S. Department of Transportation and the California State Department of Transportation, neither of which assumes liability for its content or use.

Estimability in the Multinomial Probit Model

David S. Bunch

Graduate School of Management
University of California at Davis

Reprint No. 71

Transportation Research
Vol. 25B, No. 1 (February 1991)

The University of California Transportation Center
University of California at Berkeley

Transpn. Res.-B, Vol. 25B, No. 1, pp. 1-12, 1991
Printed in Great Britain.

0191-2615/91 $3.00+.00

ESTIMABILITY IN THE MULTINOMIAL PROBIT MODEL

DAVID S. BUNCH
Graduate School of Management, University of California, Davis, Davis, CA 95616, U.S.A.

(*Received* 11 *July* 1989; *in revised form* 15 *February* 1990)

Abstract — Random utility models often involve terms which represent alternative-specific errors, and the main attractive feature of the multinomial probit (MNP) model is that it allows a rather general covariance structure for these errors. However, since observed choices only reveal information regarding utility differences, and since scale cannot be determined, not all parameters in an arbitrary MNP specification may be identified. This paper examines identification restrictions that arise in the linear-in-parameters multinomial probit framework, and provides discussion and recommendations for estimation and analysis of probit normalizations.

1. INTRODUCTION

Recent advances in computational methods for estimating multinomial probit models have stimulated renewed interest on this topic: for example, McFadden (1989) and Pakes and Pollard (1989) suggest simulation methods which may lead to practical probit estimation codes for more than four alternatives. In addition, Kamakura (1989) demonstrates, via a simulation study, that the Mendell-Elston approximation to the multivariate normal CDF is more accurate than the Clark method, and may provide an alternative solution to the problem. Bunch and Kitamura (1989) corroborate Kamakura's results in a study using empirical data, and discuss improved algorithms for maximum likelihood estimation.

This paper considers another important practical issue, the problem of formulating multinomial probit (MNP) model specifications for which the parameters are estimable. A discussion of parameter estimability issues and some illustrative examples appear in section 3.1 of Daganzo (1979); however, the topic is difficult and no general comprehensive theory is offered by Daganzo, nor do we believe one is likely to be offered in the near future. The discussion and conclusions presented here focus on the linear-in-parameters MNP model with taste variation and correlated random errors. This framework is quite flexible and is consistent with many specifications discussed in the literature.

The identification difficulties which arise are primarily due to the random errors, which are usually associated with the effect of unobserved attributes on a choice object's utility. A general model specification includes the possibility of correlations among the utilities of objects which might share unobserved attributes. Unfortunately, information about the underlying utilities is available only through observation of discrete choices, which depend on *differences* of utilities. This, plus the issue of scaling, lead to restrictions on the number of estimable parameters.

The material presented here overlaps that of Dansie (1985) and Albright, Lerman, and Manski (1977), and is perhaps well-known to some (certainly not all) workers in the field. However, the extent to which this is true is unclear. Bunch and Kitamura (1989) give a brief review of empirical applications, and a significant proportion of them were found to contain model specification errors. This is important since misspecified models potentially compound the already troublesome computational difficulties inherent in computing MNP estimates, and could be a contributing factor to the relative dearth of successful empirical applications of MNP in the published literature.

2. LINEAR-IN-PARAMETERS MULTINOMIAL PROBIT MODELS

In this paper we will consider a standard discrete choice modeling situation in which an individual drawn at random from a population makes a choice from a set of J mutually exclusive alternatives. A common example in the transportation literature is the choice of mode for the work commute, where $J = 3$ or 4 and the alternatives consist of car, train, bus, shared ride, etc. Following McFadden (1981, 1989), assume that the utility of alternative j for individual n, U_{nj}, is given by the general form

$$U_{nj} = X_{nj}^T \alpha_n, \; j = 1, \ldots, J, \tag{1}$$

where X_{nj} *is a* K-vector of explanatory variables which may be a function of the attributes of alternative j and individual n. The K-vector α_n contains the taste weights for individual n, and may be rewritten as $\alpha_n = \theta + \delta_n$, where θ is the mean taste weight for the population and δ_n is the (unobserved) random deviation from the mean for individual n. (Ideally, θ and the distribution of δ_n would vary as an explicit function of the observed characteristics of the individual, but this consideration is usually suppressed for simplicity.)

In theory, utility should be a function of generic (or "real") attributes, and should not depend on nominal attributes such as the labels "car," "bus," etc. In practical applications it is also desirable for the model to include only generic variables, since the model may then be used more effectively in forecasting, especially for testing the effect of adding new choice alternatives. However, it has generally been observed that including alternative-specific dummy variables significantly improves the fit of discrete choice models. One interpretation is that unobserved attributes are often empirically correlated with the nominal labels of the alternatives (McFadden *et al.*, 1977). For example, the attributes "lack of flexibility," or "lack of comfort," might be correlated with the nominal label "bus."

Equation (1) readily accommodates the inclusion of dummy variables, which is the approach taken by Albright, Lerman, and Manski (1977). Alternatively, we may assume that X_{nj} in eqn (1) includes only generic variables, and add the additional terms μ_j and ϵ_{nj}, where μ_j is the mean of the alternative-specific errors, and ϵ_{nj} represents a random deviation from the mean. The ϵ_{nj} term may be regarded to include both the effects of unobserved attributes and any other sources of observation-specific random error. This gives the following model, expressed now in vector notation:

$$U_n = X_n^T(\theta + \delta_n) + \mu + \epsilon_n, \tag{2}$$

where $U_n, \mu, \epsilon_n \in \Re^J$, $\theta, \delta_n \in \Re^K$, and $X_n \in \Re^{K \times J}$. To get a multinomial probit model, one adds the theoretically appealing assumption that the random terms have multivariate normal distributions:

$$\delta_n \sim MVN(0, \Sigma_\delta),\ \Sigma_\delta \in \Re^{K \times K} \text{ and } \epsilon_n \sim MVN(0, \Sigma_\epsilon),\ \Sigma_\epsilon \in \Re^{J \times J}. \tag{3}$$

Note that in this formulation the δ and ϵ terms are assumed to be independent, which is slightly more restrictive that the model implied by eqn (1). This framework is consistent with Hausman and Wise (1978).

Now, the probability that individual n selects alternative j is given by the MNP model:

$$P(j \mid V_U(\theta, \mu, X_n), \Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X_n)) = \text{Prob}[U_{nj} > U_{ni} \text{ for all } i \neq j] \tag{4a}$$

$$= \int_{u_j=-\infty}^{\infty} \int_{u_1=-\infty}^{u_j} \cdots \int_{u_{j-1}=-\infty}^{u_j} \int_{u_{j+1}=-\infty}^{u_j} \cdots \int_{u_J=-\infty}^{u_j} \phi(u \mid V_U, \Sigma_U) du_1 \cdots du_J \tag{4b}$$

where

$$V_U(\theta, \mu, X_n) = X_n^T\theta + \mu,$$

$$\Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X_n) = X_n^T \Sigma_\delta X_n + \Sigma_\epsilon,$$

and $\phi(x \mid m, S)$ is the multivariate normal density function with mean m and covariance S. If we assume that Σ_δ and Σ_ϵ are positive definite then it is straightforward to show that

$\Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X_n)$ is also positive definite, which is desirable for establishing regularity conditions: see Daganzo (1979), or, hereafter, "Daganzo."

The purpose of this paper is to discuss the issues and problems of specifying the linear-in-parameters MNP model so that all the parameters are estimable. A primary concern is specification of Σ_ϵ, which is especially important since most published applications of MNP focus exclusively on estimating the effects of observation-specific errors (i.e. they assume (2) with $\delta_n = 0$).

3. REDUCTION OF DIMENSION

As has been often noted, one of the difficulties with MNP is that it requires evaluation of the multivariate integral (4b), which does not have a closed form solution. The usual first step is to reduce the dimension of the integral from J to J-1 using the transformation discussed by Daganzo (1979, pp. 43–44 and pp. 94–95). Assume that we wish to compute $P(j \mid V_U(\theta, \mu, X), \Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X))$ and define $\Delta_j \in \Re^{J-1\times J}$ by

$$\Delta_j = \begin{array}{c|cccccccccc|}
 & 1 & 2 & 3 & \cdots & j-1 & j & j+1 & \cdots & J-1 & J \\
1 & 1 & 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 0 \\
2 & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 0 \\
3 & 0 & 0 & 1 & & 0 & -1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
j-1 & 0 & 0 & 0 & & 1 & -1 & 0 & & 0 & 0 \\
j & 0 & 0 & 0 & & 0 & -1 & 1 & & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
J-2 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 1 & 0 \\
J-1 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 1
\end{array} \; . \tag{5}$$

The transformation $Z = \Delta_j U$ applied to (4) gives

$$\begin{aligned} P(j \mid V_U(\theta, \mu, X), \Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X)) &= \text{Prob}\,[U_i - U_j < 0, i \neq j] \\ &= \Phi(0 \mid V_Z, \Sigma_Z) \end{aligned} \tag{6}$$

where

$$V_Z = \Delta_j V_U = \Delta_j(X^T\theta + \mu) = \Delta_j X^T\theta + \Delta_j\mu, \tag{7a}$$

$$\Sigma_Z = \Delta_j \Sigma_U \Delta_j^T = \Delta_j(X^T\Sigma_\delta X + \Sigma_\epsilon)\Delta_j^T = \Delta_j X^T\Sigma_\delta X\Delta_j^T + \Delta_j\Sigma_\epsilon\Delta_j^T; \tag{7b}$$

Z, $V_Z \in \Re^{J-1}$, $\Sigma_Z \in \Re^{(J-1)\times(J-1)}$, and Φ is the cumulative distribution function for the (J − 1)-dimensional multivariate normal distribution.

Define $m_j \in \Re^{J-1}$ by $m_j = \Delta_j\mu$ and $C_j \in \Re^{(J-1)\times(J-1)}$ by $C_j = \Delta_j\Sigma_\epsilon\Delta_j^T$, where C_j is symmetric. Consider the matrix $M_j \in \Re^{(J-1)\times(J-1)}$ which is obtained by taking Δ_j and deleting the Jth column; M_j is of full rank (i.e. rank $J - 1$). Next, without loss of generality choose alternative J as the "reference alternative." It is straightforward to show that $\Delta_j = M_j\Delta_J$ for all j, and hence $m_j = M_j m_J$ and $C_j = M_j C_J M_j^T$ for all j. It follows that (7) may be rewritten as

$$V_Z = \Delta_j X^T\theta + M_j m_J, \tag{8a}$$

$$\Sigma_Z = \Delta_j X^T\Sigma_\delta X\Delta_j^T + M_j C_J M_j^T, \tag{8b}$$

and thus choice probabilities evaluated via (6) may *always* by expressed in terms of m_J and C_J, which together contain $(J - 1) + J(J - 1)/2$ parameters. It follows that μ and Σ_ϵ together only have $(J - 1) + J(J - 1)/2$ identifiable parameters. As an illustration, consider the case $J = 3$ and $j = 2$, which gives (taking into account the symmetry of Σ_ϵ):

$$m_2 = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_3 - \mu_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 - \mu_3 \\ \mu_2 - \mu_3 \end{bmatrix} = M_2 m_3,$$

and

$$\begin{aligned} C_2 &= \begin{bmatrix} \sigma_{11} - 2\sigma_{21} + \sigma_{22} & \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{22} \\ \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{22} & \sigma_{33} - 2\sigma_{31} + \sigma_{22} \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} - 2\sigma_{31} + \sigma_{33} & \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} \\ \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} & \sigma_{22} - 2\sigma_{32} + \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & -1 \end{bmatrix} \\ &= M_2 C_3 M_2^T, \end{aligned}$$

where the σ_{ij}'s denote the elements of Σ_ϵ.

In addition to these considerations, it turns out that one can also rescale the problem so as to eliminate another parameter. Specifically, evaluation of $\Phi(0 \mid kV_Z, k^2\Sigma_Z)$ gives the same result as evaluating $\Phi(0 \mid V_Z, \Sigma_Z)$, for $k > 0$. Hence, suitable selection of k can eliminate one more parameter. This leads to the assertion that the model (3) has a total of $K + K(K + 1)/2 + J + J(J - 1)/2 - 2$ identifiable parameters. Since θ, μ, Σ_δ, and Σ_ϵ together have $K + K(K + 1)/2 + J + J(J + 1)/2$ parameters, this implies that $J + 2$ parameters are inestimable and must be fixed via some normalization. (The fact that Σ_ϵ has only $J(J - 1)/2$ identifiable parameters was noted by Daganzo, who also recognized that one could sometimes eliminate another parameter by rescaling. Albright, Lerman, and Manski (1977) state the above result for the number of free parameters, referring the reader to technical memoranda.)

Note that for the binary probit model with no taste variation (i.e. $J = 2$ and $\delta = 0$), there are K taste weights, $J - 1$ alternative specific dummy variables, and $J(J - 1)/2 - 1 = 0$ free parameters in Σ_ϵ. Although this is generally known, even experienced investigators publishing in refereed journals can miss this point; for example, in a study of travel mode choice Johnson and Hensher (1982) erroneously assume that Σ_ϵ has one free parameter. In studies involving more complex MNP models, other authors have completely missed the requirements described above, and assume that all the parameters are in principle estimable. See, for example, Currim (1982) and van Lierop (1986). For a brief review of MNP empirical applications which includes further discussion, see Bunch and Kitamura (1989).

These restrictions have implications for practical estimation of MNP models. Recall that Daganzo discusses MNP in a slightly more general framework by considering the MNP function $P(j \mid V(\theta, A), \Sigma(\theta, A))$, where θ is a general parameter vector and A is a matrix of attributes. (Note, however, that many of his examples fit in the linear-in-parameters framework.) In discussing the practical issues of parameter estimation, he recommends that $\Sigma(\theta, A)$ be specified so as to be positive definite over all feasible values of θ, since otherwise it "would not represent a covariance matrix and the program would not return meaningful values." Daganzo recommends two possibilities:

(i) express $\Sigma(\theta, A)$ as a product of a matrix and its transpose (i.e. $\Sigma(\theta, A) = C(\theta, A)C(\theta, A)^T$),

(ii) express $\Sigma(\theta, A)$ as a function of θ and A directly, placing simple bounds on θ so as to ensure positive definiteness.

We may regard $C(\theta, A)$ to be lower triangular with positive diagonal elements (i.e. the Cholesky factorization of $\Sigma(\theta, A)$). For the case we are considering here, either of these approaches could be difficult when choosing a specification for Σ_U. Directly writing $\Sigma_U = CC^T$ is not practical if taste variation is included in the model. One could express each of Σ_δ and Σ_ϵ in terms of Cholesky factorizations, but this could be tricky for Σ_ϵ (or, Σ_U in the case of fixed tastes) since we have at most $J(J - 1)/2$ free parameters to work with in Σ_ϵ. For most of the following discussion, we will assume that we wish to choose a normalization which utilizes all available parameters.

Albright, Lerman, and Manski (1977), which we denote "ALM" in the sequel, deal with the issue by choosing one normalization from "among the alternative formally equivalent normalizations." First, they choose to express both Σ_δ and Σ_ϵ in terms of Cholesky factorizations. Next, they choose to fix J parameters by setting the last row (and column) of Σ_ϵ, and correspondingly, the last row of its Cholesky factorization, to zero. Finally, they choose to fix the scale of the specification by constraining the diagonal elements of Σ_ϵ so that (trace Σ_ϵ)/J equals the variance of the standard Weibull distribution. (This was done to facilitate comparisons between MNP and multinomial logit estimates.) Note that this is inconsistent with Daganzo's recommendation in the no-taste-variation case: this specification of Σ_ϵ, while a valid covariance matrix, is only positive semi-definite and hence violates Daganzo's regularity conditions.

In fact, ALM are doing slightly more than simply choosing an "arbitrary normalization": they are choosing to work directly in (J − 1)-space, estimating C_J so that C_J is constrained to be positive definite (which implies that C_j for all $j \neq J$ must also be positive definite). This is a more general approach than choosing to perform the estimation using arbitrarily constrained formulations in the original J-dimensional space.

As an example, consider the case $J = 3$. For simplicity, we suppress the scaling issue at this stage: one can assume that scale has been fixed via elimination of a parameter component from either θ, μ, or Σ_δ. For $J = 3$ Σ_ϵ has $3(3 - 1)/2 = 3$ free parameters. The ALM normalization (with no rescaling) is given by:

$$\Sigma_\epsilon^{\mathrm{ALM}} = \begin{bmatrix} \sigma_{11} & \sigma_{21} & 0 \\ \sigma_{21} & \sigma_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \qquad (9)$$

Another possible normalization, which assumes independence of the random errors, is given by a diagonal covariance matrix:

$$\Sigma_\epsilon^{\mathrm{D}} = \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}, \qquad (10)$$

with σ_{11}, σ_{22}, $\sigma_{33} > 0$ to ensure positive definiteness. Now, suppose that a maximum likelihood estimation routine produces an estimate for C_J and returns the following result:

$$C_J = \begin{bmatrix} 1.5 & -0.4 \\ -0.4 & 1.2 \end{bmatrix} \qquad (11)$$

Then the ALM normalization is simply

$$\Sigma_\epsilon^{\mathrm{ALM}} = \begin{bmatrix} 1.5 & -0.4 & 0 \\ -0.4 & 1.2 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \qquad (12)$$

However, the corresponding Σ_ϵ^D is computed to be:

$$\Sigma_\epsilon^{\mathrm{D}} = \begin{bmatrix} 1.9 & 0 & 0 \\ 0 & 1.6 & 0 \\ 0 & 0 & -0.4 \end{bmatrix}, \qquad (13)$$

which is clearly not a valid covariance matrix. Problems could arise if the estimation routine were attempting to directly estimate the diagonal normalization: a boundary solution would probably be the result.

It seems clear that the best course to follow is to first estimate C_J, and to then explore the various possible normalizations by performing transformations between C_J and Σ_ϵ. If any particular specification is inappropriate, it will show up as an invalid covariance matrix without jeopardizing the performance of the estimation routine. In addition, the possibility of accidentally choosing an inestimable normalization will be minimized, since the process of analyzing the transformation will reveal the mistake. This is discussed next.

4. TRANSFORMATIONS

As above, assume for now that fixing the scale is not an issue. In addition, assume that the specifications we are considering are fairly simple, (i.e. parameters do not appear simultaneously across θ, μ, Σ_δ, Σ_ϵ). (Although this restriction is not necessary, it greatly simplifies the analysis and discussion.) In what follows, it will often be convenient to regard symmetric matrices as vectors in packed storage form (e.g. when $J = 3$ then $\Sigma_\epsilon \in \Re^{J(J+1)/2}$ with $\Sigma_\epsilon = (\sigma_{11}, \sigma_{21}, \sigma_{22}, \sigma_{31}, \sigma_{32}, \sigma_{33})^T$, and $C_J \in \Re^{(J-1)/2}$ with $C_J = (c_{11}, c_{21}, c_{22})^T$).

Suppose that $\beta \in \Re^{(J-1)J/2}$ denotes the vector of "identified" parameters. Then candidate normalizations may be represented as mappings from $\Re^{(J-1)J/2}$ to $\Re^{J(J+1)/2}$. For example, consider the ($J = 3$) ALM normalization in eqn (9) which in matrix form can be written as

$$\Sigma_\epsilon^{\text{ALM}} = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_3 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{14}$$

In packed form this particular normalization may be represented by the (linear) mapping

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ 0 \\ 0 \\ 0 \end{bmatrix} = h_1(\beta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = H_1\beta. \tag{15}$$

Recall that the "observable" parameters reside in $C_J \in \Re^{(J-1)J/2}$ (in packed form), which is obtained through the mapping $g(\cdot)$ from $\Re^{J(J+1)/2}$ to $\Re^{(J-1)J/2}$, defined below. Define the function $f(\beta) = g(h(\beta))$ which maps $\Re^{(J-1)J/2}$ to itself. Then the identification problem reduces to verifying that $f(\cdot)$ represents a unique invertible transformation between β and C_J. The relevant requirement from the Inverse Function Theorem—see Apostol (1974)—is that the Jacobian determinant of f be different from zero.

To illustrate, consider the simple class of $J = 3$) normalizations which are restricted so that either $\sigma_{ij} = 0$ or $\sigma_{ij} = \sigma_{kl}$. This includes both of the previous examples:

$$\Sigma_1 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow h_1(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \beta = H_1\beta \tag{16}$$

and

$$\Sigma_2 = \begin{bmatrix} \beta_1 & 0 & 0 \\ 0 & \beta_2 & 0 \\ 0 & 0 & \beta_3 \end{bmatrix} \Rightarrow h_2(\beta) = \begin{bmatrix} \beta_1 \\ 0 \\ \beta_2 \\ 0 \\ 0 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \beta = H_2\beta. \qquad (17)$$

Some additional possibilities are:

$$\Sigma_3 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_1 & \beta_3 \\ 0 & \beta_3 & \beta_1 \end{bmatrix} \Rightarrow h_3(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_1 \\ 0 \\ \beta_3 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \beta = H_3\beta, \qquad (18)$$

$$\Sigma_4 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_3 & 0 \\ 0 & 0 & \beta_3 \end{bmatrix} \Rightarrow h_4(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ 0 \\ 0 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \beta = H_4\beta, \qquad (19)$$

$$\Sigma_5 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_1 & 0 \\ 0 & 0 & \beta_3 \end{bmatrix} \Rightarrow h_5(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_1 \\ 0 \\ 0 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \beta = H_5\beta, \qquad (20)$$

$$\Sigma_6 = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \\ \beta_2 & \beta_1 & \beta_3 \\ \beta_3 & \beta_3 & \beta_1 \end{bmatrix} \Rightarrow h_6(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_1 \\ \beta_3 \\ \beta_3 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \beta = H_6\beta. \qquad (21)$$

The function $g(\,\cdot\,)$ which maps Σ_ϵ to C_j may be defined by the linear transformation G given by

$$g(\Sigma_\epsilon) = \begin{bmatrix} 1 & 0 & 0 & -2 & 0 & 1 \\ 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 1 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{22} \\ \sigma_{31} \\ \sigma_{32} \\ \sigma_{33} \end{bmatrix} = G\Sigma_\epsilon, \qquad (22)$$

and hence for this class of normalizations $f(\beta) = g(h(\beta)) = GH\beta$ is the composite of two linear transformations. Application of the chain rule gives the Jacobian of f, $Df = Dg \circ Dh = GH$. It follows that

$$Df_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad Df_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$Df_3 = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 0 & -2 \\ 1 & 1 & -1 \end{bmatrix}, \qquad Df_4 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix},$$

$$Df_5 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \qquad Df_6 = \begin{bmatrix} 2 & 0 & -2 \\ 1 & 1 & -2 \\ 2 & 0 & -2 \end{bmatrix}.$$

The Jacobians Df_1 through Df_4 are of full rank, and hence the β's are identifiable for the corresponding specifications. In contrast, Df_5 and Df_6 are not of full rank, and the β's are not identifiable in Σ_5 and Σ_6. Thus, one may not arbitrarily choose any normalization which appears to have the correct number of free parameters. It is also essential to verify that the transformation implied by the normalization is invertible, a point which has been overlooked until recently.

In particular, consider Σ_5. If the scaling of this specification is fixed by dividing by β_1, the resulting normalization is

$$\Sigma_5' = \begin{bmatrix} 1 & \beta_2' & 0 \\ \beta_2' & 1 & 0 \\ 0 & 0 & \beta_3' \end{bmatrix}. \tag{23}$$

This unidentified normalization appears in Daganzo (1979), is quoted by Currim (1982), and is used in a simulation study by Horowitz *et al.* (1982). The fact that Σ_5' is unidentified was pointed out by Dansie (1985).

The above approach is general and is easily extended to $J > 3$. For example, consider a mode choice problem with $J = 4$ where the choice set is {drive alone, carpool, train, bus}. There are $4(4 - 1)/2 = 6$ identifiable parameters in Σ_ϵ, and two possible candidate normalizations are:

$$\Sigma_6 = \begin{bmatrix} \beta_1 & \beta_2 & 0 & 0 \\ \beta_2 & \beta_3 & 0 & 0 \\ 0 & 0 & \beta_4 & \beta_5 \\ 0 & 0 & \beta_5 & \beta_6 \end{bmatrix}, \tag{24}$$

and

$$\Sigma_7 = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_3 \\ \beta_2 & \beta_1 & \beta_4 & \beta_4 \\ \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_3 & \beta_4 & \beta_6 & \beta_5 \end{bmatrix}. \tag{25}$$

For Σ_6, the assumption is that each of the alternatives has an error term with a different variance, and that there is correlation of unobserved attributes between the pairs {drive alone, car pool} and {train, bus}, respectively. Error terms for {drive, carpool} are assumed to be uncorrelated with error terms for {train, bus}. In contrast, for Σ_7 the error terms for {drive, carpool} are assumed to be correlated with the error terms for {train, bus}, and each of these pairs is assumed to have the same variance. A straightforward application of the procedure developed above reveals that Σ_6 is identified, where as Σ_7 is not.

5. SCALING

Although one can generally consider fixing the scale of (2) by dividing through by a coefficient parameter, the standard practice in published MNP applications is to incorporate scaling in the specification of Σ_ϵ. This seems more intuitive, and is based on what are probably more acceptable assumptions. The identification results for the examples of the previous section still hold if one chooses to include scaling, as a consequence of the simple structure that was assumed.

If thinking about the specification in J-space, one could divide Σ_ϵ by a constant multiple of one of the (nonzero) variance components, leaving one or more constant terms on the diagonal. (If one is keeping track, a corresponding adjustment would take place for θ and μ, multiplying them by the positive square root of the same expression.) Alternatively, one could consider the "observable" parameters to live in $(J - 1)$-space, and divide all the elements of C_J by c_{11} $(=\sigma_{11} - 2\sigma_{1J} + \sigma_{JJ})$ so that the estimated parameters reside in a symmetric $(J - 1)$-dimensional matrix with a "1" in the upper left-hand corner. The matrix C_J may be constrained to be positive definite through a convenient Cholesky factorization B_J with $J(J - 1)/2 - 1$ free parameters (i.e. $C_J = B_J B_J^T$) where:

$$B_J = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ b_{21} & b_{22} & 0 & \cdots & 0 \\ b_{31} & b_{32} & b_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{J1} & b_{J2} & b_{J3} & \cdots & b_{JJ} \end{bmatrix}. \tag{26}$$

Transformations which directly incorporate scaling, defined using the spaces $\Re^{J(J+1)/2-1}$ and $\Re^{(J-1)J/2-1}$, may be studied using the ideas of the previous section, although the exercise is more tedious and a bit less intuitive. It should be clear that one can readily move from one normalization to another via C_J, and rescale when it is convenient. This is illustrated in the next section.

6. NUMERICAL EXAMPLES

A major application of choice models in transportation is the estimation of commuter mode choice. Most applications of this type have used the logit model, but two major examples using linear-in-parameters MNP are Hausman and Wise (1978) and the report by Albright, Lerman, and Manski. Both use a data set collected in Washington, D.C. in which the choices are {car, shared-ride, bus}.

Hausman and Wise (HW) use a subset containing 100 observations, and restrict themselves to models for which the most general specification assumes uncorrelated random errors (i.e. that Σ_δ and Σ_ϵ are diagonal). The generic attributes are trip cost divided by personal income, in-vehicle travel time, and out-of-vehicle travel time. The choice probabilities are evaluated using numerical integration.

In contrast, ALM use 1353 observations, and estimate a quite general specification which allows for correlated random errors. They use essentially the same generic attributes as HW, but include two mode-specific variables on available autos per licensed driver. They evaluate choice probabilities via Clark's approximation – see Daganzo (1979).

An interesting feature of these reported results is that both obtain estimates for a fully-specified Σ_ϵ matrix. In fact, the two specifications are essentially those given by eqns (9) and (10) above, but with the following adjustments for scaling purposes:

$$\Sigma_\epsilon^{\text{ALM}} = \begin{bmatrix} \sigma_{11} & \sigma_{21} & 0 \\ \sigma_{21} & w - \sigma_{11} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\Sigma_\epsilon^{\text{HW}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}.$$

Based on the discussion in the previous two sections, one can transform the HW and ALM results into various identifiable normalizations for comparison purposes. Table 1 includes such a comparison using the (identified) normalizations discussed above. There are no striking differences between the two sets of results, even though they were obtained with different sample sizes and different assumptions regarding taste variation. It is clear from this example that, depending upon the modeling assumptions one is willing to make, different interpretations involving the effect of unobserved variables are possible.

However, recall from the example in eqns (11) through (13) that in some situations candidate normalizations could be rejected if they produce invalid covariance matrices. This is illustrated in Bunch and Kitamura (1989), in which trinomial probit models of car ownership are estimated using the approach recommended here. A comparison of various normalizations is made, with some being rejected as invalid.

7. MORE PARSIMONIOUS SPECIFICATIONS

In the section on transformations we limited ourselves to normalizations which utilize all available parameters; however, this is not necessary and one might wish to consider more parsimonious specifications involving fewer parameters. In this case the relationships between C_j and Σ_ϵ may still be examined, but the researcher must give careful thought to what is being assumed.

For example, the specification Σ_5 (or, equivalently, Σ_5') is unidentified. Dansie (1985) shows that the three covariance matrices

$$\Sigma_A = \begin{bmatrix} 1 & \sigma_{21} & 0 \\ \sigma_{21} & 1 & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}$$

$$\Sigma_B = \begin{bmatrix} 1 & \sigma_{21} & 0 \\ \sigma_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Sigma_C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}$$

are equivalent, but that only the second two are identified, with one free parameter. As a consequence, he recommends estimating the model using a one-parameter C_j matrix given by

Table 1. Alternative Σ_ϵ normalization for MNP mode choice results

Albright, Lerman, and Manski	Hausman and Wise
$\begin{bmatrix} 1 & 0.45 & 0 \\ 0.45 & 1.21 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.50 & 0 \\ 0.50 & 1.51 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1.38 & 0 \\ 0 & 0 & 0.82 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2.02 & 0 \\ 0 & 0 & 1.00 \end{bmatrix}$
$\begin{bmatrix} 1 & -0.21 & 0 \\ -0.21 & 1 & -0.34 \\ 0 & -0.34 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & -0.51 & 0 \\ -0.51 & 1 & -0.51 \\ 0 & -0.51 & 1 \end{bmatrix}$
$\begin{bmatrix} 1 & -0.39 & 0 \\ -0.39 & 1.53 & 0 \\ 0 & 0 & 1.53 \end{bmatrix}$	$\begin{bmatrix} 1 & -1.04 & 0 \\ -1.04 & 3.08 & 0 \\ 0 & 0 & 3.08 \end{bmatrix}$

$$C_J = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

This is equivalent to assuming that, whatever the normalization may look like, we expect that the variance of $\epsilon_1-\epsilon_3$ equals the variance of $\epsilon_2-\epsilon_3$. Note that many normalizations may satisfy this requirement. In particular, if the estimation routine produces a result with $\rho < 0$, then neither of the two normalizations Σ_B and Σ_C will be valid covariance matrices; of course others could be found which are.

8. SUMMARY

The linear-in-parameters multinomial probit framework in eqn (2) includes a random error vector which in general has a covariance matrix Σ_ϵ, but unfortunately not all the covariance parameters are identified. Section 4 develops the arguments which give the number of estimable parameters, and recommends that estimation be performed in terms of a matrix C_J. Multiple normalizations may correspond to the same estimated C_J, but some may not produce valid covariance matrices. Choosing a particular (valid) normalization is essentially a *modeling decision*, as illustrated via the examples in section 4. This feature of MNP is often overlooked in the sweeping laudatory descriptions of the model, which extol the generality of the approach versus more highly restrictive models such as multinomial logit. In fact, users of probit must in the final analysis make modeling assumptions which are analogous to choosing among various alternative tree structures in the nested multinomial logit or tree extreme value models.

To perform this exercise, however, knowledge of the number of estimable parameters is still not enough. Arbitrarily selected specifications for Σ_ϵ which contain the prescribed number of estimable parameters may not be identified, since the transformation between the parameters in the specification and the "observable" C_J parameters may not be unique and invertible. Examples are given which demonstrate that an analysis is essential for ensuring valid results. Specification errors are common in the literature and these issues appear to not be generally understood. The results presented here provide useful guidelines for those practitioners seeking to apply probit models.

Acknowledgements—This work was supported in part by the USDOT through the Region Nine Transportation Center at University of California, Berkeley. The author thanks Joel Horowitz for providing a copy of the Albright, Lerman, and Manski (1977) report, and Ryuichi Kitamura, David M. Rocke, and an anonymous referee for helpful discussions and comments on an earlier draft of this paper.

REFERENCES

Albright R. L., Lerman S. R., and Manski C. F. (1977) Report on the development of an estimation program for the multinomial probit model. Preliminary report prepared for the Federal Highway Administration, October, p. 54.

Apostol T. M. (1974) *Mathematical Analysis*, (Second edition). Addison-Wesley, Reading, MA.

Bunch D. S., and Kitamura R. (1989) Multinomial probit estimation revisited: Testing of new algorithms and evaluation of alternative model specifications for trinomial models of household car ownership. Transportation Research Group Research Report UCD-TRG-RR-4, University of California, Davis, CA.

Currim I. S. (1982) Predictive testing of consumer choice models not subject to independence of irrelevant alternatives. *J. Marketing Res.*, **19**, 208-222.

Daganzo C. (1979) *Multinomial Probit: The Theory and Its Application to Demand Forecasting*. Academic Press, New York.

Dansie B. R. (1985) Parameter estimability in the multinomial probit model. *Transpn. Res. B*, **19B**, 526-528.

Hausman J. A., and Wise D. A. (1978) A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica*, **46**, 403-426.

Horowitz J. L., Sparmann J. M., and Daganzo C. F. (1982) An investigation of the accuracy of the Clark approximation for the multinomial probit model. *Transpn. Sci.*, **16**, 3, 382-401.

Johnson L. and Hensher D. (1982) Application of multinomial probit to a two-period panel data set. *Transpn. Res. A*, **16A**, 5/6, 457-464.

Kamakura W. (1989) The estimation of multinomial probit models: A new calibration algorithm, *Transpn. Sci*, **23**(4), 253-265.

van Lierop W. (1986) *Spatial Interaction Modeling and Residential Choice Analysis*. Gower Publishing, Aldershot, England.

McFadden D. (1981) Econometric models of probabilistic choice. In C. F. Manski and D. McFadden, (Eds.) *Structural Analysis of Discrete Data with Econometric Applications*, pp. 198-272. MIT Press, Cambridge, MA.

McFadden D. (1989) A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica,* **57**(5), 995-1026.

McFadden D., Tye W., and Train K. (1977) An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model. *Trnspn. Res. Rec.*, **637**, 39-46.

Pakes A. and Pollard D. (1989) Simulation and the asymptotics of optimization estimators. *Econometrica,* **57**(5), 1027-1057.