Analytic Approximations for Computing Probit Choice Probabilities

Authors: Dr Richard D. Connors [corresponding author]
Institute for Transport Studies, University of Leeds, Leeds, LS2 9JT, United Kingdom
Tel: +44 113 3431799
Fax: +44 113 3435334
Email: R.D.Connors@its.leeds.ac.uk

Dr Stephane Hess
Institute for Transport Studies, University of Leeds, Leeds, LS2 9JT, United Kingdom
Tel: +44 113 3436611
Fax: +44 113 3435334
Email: s.hess@its.leeds.ac.uk

Prof Andrew Daly
RAND Europe, Westbrook Centre, Milton Road, Cambridge, CB4 1YG, United Kingdom
Tel: +44 1223 353 329
Fax: +44 1223 358 845
Mobile +44 7815 871 203
Email: daly@rand.org
Abstract

The multinomial probit model has long been used in transport applications; as the basis for mode- and route-choice in computing network flows, and in other choice contexts when estimating preference parameters. Over recent years, the use of probit in the choice modelling community has declined, with growing focus on mixed logit models. The theoretical appeal of the error structure of the probit model however remains, and recent work on composite marginal likelihood estimation (Bhat, 2010) has renewed interest in the probit model in the specific context of preference estimation. Whether using the probit model for estimating parameters or for evaluating choice proportions, computation of the choice probabilities presents a computational burden, given that they are based on multivariate normal integrals. In this paper we compare two analytical approximation methods for computing both the probit choice probabilities and their derivatives in terms of accuracy and computation efficiency. Here, we show that the “Mendell-Elston” approach outperforms the recently used “Solow-Joe” approach across a range of settings. Wider use of the “Mendell-Elston” in probit work, and implementation within a composite marginal likelihood estimator are thus promising areas for new developments. While the primary motivation for this paper is to help with the estimation and application of discrete choice models, there may be applications in other areas where the computation of multivariate normal integrals is needed.

Keywords
Multinomial Probit, Multivariate Normal Integral, Analytical Approximation, Choice Probabilities

1. Introduction

Consider an individual who is asked to choose from a set of $K$ mutually exclusive alternatives. We define a function, $V(x, \beta)$, intended to capture the attractiveness of each alternative. With alternative $j$ described by the $m$-vector of its attributes $x_j$, and individual $n$ described by attributes $\beta_n$, $V_{n,j} = V(x_j, \beta_n)$ is the attractiveness of alternative $j$ to individual $n$. Across the population of individuals and across alternatives there may be errors in the measurement of attribute data, and influential neglected attributes may exist, meaning that $V_{n,j}$ gives an incomplete representation of the utility. To account for these issues we introduce an additive random variable $e_{n,j}$ and model choices by comparing the perceived attractiveness, $U$, of the alternatives

$$U_{n,j} = V(x_j, \beta_n) + e_{n,j} \quad \text{for } j = 1, \ldots, K \text{ and for each } n. \quad (1)$$

The random utility paradigm models choice by identifying the chosen alternative as the one with the highest utility. Given data $x_j, \beta_n$, if the distribution for the random errors is known then the probability $P_{n,j}$ that decision maker $n$ will choose alternative $j$ can, in principle, be calculated.

$$P_{n,j}(x_j, \beta_n) = \Pr[U_{n,j} > U_{n,k}; k = 1, \ldots, K, k \neq j] \quad (2)$$

The probability in (2) is an integral over the distribution of $e$. The structure of the model therefore derives from the specification of the error term’s density function $g(e; \xi)$, with $\xi$ the distribution parameters. This multivariate integral will have a closed form solution only for specific choices of distribution. In particular, if the random error terms in (1) are assumed to be independent and identically distributed according to a type I extreme value (a.k.a. Gumbel) distribution, the choice probabilities can be expressed as

$$P_{n,i}(x, \beta_n) = \frac{e^{V_{n,i}}}{\sum_{j=1}^{K} e^{V_{n,j}}} \quad (3)$$

This assumption gives rise to the celebrated multinomial logit (MNL) model (cf. McFadden, 1974), for which the choice probabilities can be calculated very quickly and accurately. While the MNL model is convenient and easy to use, it has a number of important restrictions. Initial gains in flexibility were obtained through the use of generalised extreme value distributions, giving rise to structures such as Nested Logit (cf. Daly and Zachary, 1978; McFadden, 1978a; Williams, 1977) and Cross-Nested Logit (cf. Vovsha 1997), which maintain a closed form solution for the integrals in Equation (2).

While the standard closed form GEV models are appealing due to their low computational cost in estimation and application, there are limits to the degree of complexity that can be conveniently represented by making...
use only of (generalised) extreme value distributions for $e$. In particular, researchers have sought to represent a number of core phenomena, including unexplained taste heterogeneity across respondents and across choices for the same respondent, correlation across choices for the same respondent, and heteroskedasticity across respondents and/or alternatives.

In aiming to retain the underlying GEV structure, researchers have achieved these gains in flexibility through the incorporation of additional random terms on top of the extreme value errors. This gives rise to Mixed Logit models (with applications going back as far as Boyd and Mellman, 1980, and Cardell and Dunbar, 1980, and discussions of important properties in Dalal & Klein, 1988, McFadden & Train, 2000 and Hensher & Greene, 2003), or in more general terms Mixed GEV models (see e.g. Hess et al., 2005). These models have been used principally for the representation of random taste heterogeneity across respondents, i.e. random coefficients, which is the specification we focus on here. A mathematically equivalent specification can be used to introduce inter-alternative correlation or heteroskedasticity; here, the additional error terms are simply interacted with 0-1 dummy terms as opposed to continuous attributes; a model that is typically referred to as the error components specification of the Mixed Logit model (see Walker, 2001; also Paag et al., 2001, which includes both forms of variation).

Equations (2) and (3) above are conditional on a specific realisation of the vector $\beta_n$, representing the sensitivity of respondent $n$ to the various attributes describing an alternative. In a random coefficients model, we now allow $\beta_n$ to vary across respondents according to some distribution, $\beta \sim f(\beta|\Omega)$, with $\Omega$ parameterising the density function. While the estimation of a closed form discrete choice model entails the estimation of the individual attribute parameters (the $\beta$s) based on choice data, typically using the method of maximum likelihood, we now need to estimate the parameters of the distribution of those coefficients. If $i = i(n)$ is the chosen alternative for individual $n$ and the $\beta$s are assumed to vary across the population, then the likelihood is

$$L(\Omega) = \prod_n \int_\beta P_{n,i}(x_i|\beta)f(\beta|\Omega)d\beta$$

Maximum likelihood estimation requires the repeated testing of different distributional parameters, $\Omega$. For each candidate $\Omega$, the computational burden is in evaluating the integral over $\beta$. Monte Carlo is perhaps the most commonly used approach, though all methods require multiple evaluations of the choice probabilities in the integrand. Crucially, the same overall concept (and hence estimation approach) applies independently of the specific model used inside the integral, i.e. for $P_{n,i}(x_i|\beta)$. Indeed, while this will typically be of MNL form, it could equally well be a more general model belonging to the Logit family, or indeed a non-Logit model. The analyst thus needs to make an assumption not just for the distribution of $\beta$, but also for the underlying distribution of $e$. Depending on this latter assumption, $P_{n,i}(x_i|\beta)$ itself may not be of closed form, and we will have a model in which the choice probabilities conditional on $\beta$ need to be approximated numerically themselves, before dealing with the outside integration over the distribution of $\beta$.

Equation (4) refers to the case where we have a single choice per respondent, i.e. cross-sectional data. This formulation has been extended to the case of ‘panel data’, where each individual makes several choices (labelled by $t = 1, ..., T_n$ for individual $n$). Here, the base assumption is that each individual’s preference parameters stay constant across choice tasks (typically attributed to Revelt & Train, 1997, but see also Stiratelli et al., 1984), but that they vary across different individuals. More recent work (Bhat and Castelar, 2002; Bhat and Sardesai, 2006; Hess and Rose, 2009) has expanded on this by additionally allowing for variations across choice situations for the same respondent. In this general specification, we have $\beta_{n,t} = \alpha_n + \gamma_{n,t}$ where alpha varies over individuals $\alpha \sim f(\alpha|\Omega_\alpha)$ and gamma varies over the choices of an individual $\gamma \sim g(\gamma|\Omega_\gamma)$. Then with the chosen alternative being $i = i(n, t)$ and distribution parameters $\Omega = \{\Omega_\alpha, \Omega_\gamma\}$, we have

$$L(\Omega) = \prod_n \int_\alpha \prod_t \left[ \int_\gamma P_{n,i,t}(x_{i,t}|\beta = \alpha + \gamma)g(\gamma|\Omega_\gamma) dy \right] f(\alpha|\Omega_\alpha)d\alpha$$

(5)
As in the simpler formulation above, the integrals in the likelihood function need to be evaluated at many candidate values of \( \Omega \) to determine the ML estimate. In this formulation the computational burden is compounded by the integral of a product of integrals. These calculations again require very many evaluations of the integrand(s) and hence the choice probabilities. In the case where we do not allow for additional heterogeneity across tasks for the same respondent, the inner integral in Equation (5) drops out. Nevertheless, there remains an integral of a product of probabilities, rather than a product of integrals of individual probabilities, as in the cross-sectional specification in Equation (4). A transformation of (5) suggested by Daganzo and Sheffi (1982) to reduce the nesting of the integrals, at the expense of increasing the dimensionality of each one, is set out in the Appendix, which further extends the possibilities for panel data, including the possibility of including both between and within-individual variation.

The choice modelling community has had a particular focus on models with an underlying extreme value distribution, with gains in flexibility obtained by the incorporation of additional random terms. It should however be clear that a different way of obtaining flexibility is by making alternative assumptions about the underlying error term \( e \), i.e. potentially moving away from an extreme value distribution. The particular case where the error terms have a joint multivariate normal (MVN) distribution is referred to as the (multinomial) probit model (dating back to Thurstone, 1927), which has been used in many published studies stretching back to the 1940s (for example Finney, 1944a, 1944b). In the particular context of choice modelling, interest in the probit model waned with the increasing use mixed logit, which has a number of advantages in terms of distributional assumptions and ease of estimation. However, the error structure of the probit model remains appealing, and interest in it has been reinvigorated by Bhat’s recent work (Bhat, 2010; Bhat and Sidharthan, 2010; Bhat et al., 2010) that highlights the benefit of adopting the probit choice model within a composite marginal likelihood (CML) framework for the estimation of choice preferences based on survey data. This renewed interest in the probit model is a further motivation for attempting to revisit the issue of its high computational cost.

While the interest in mixed logit has been primarily based on the desire for an adequate representation of between-individual heterogeneity (notwithstanding the work on error components), probit is often seen as ideally suited to accommodate complex correlation structures. For this reason probit is particularly attractive for modelling route choice (see Sheffi, 1985). In a transport network, routes across the network are composed of many constituent links. If two alternative routes share a given link, it seems reasonable that the attractiveness of these two routes should be correlated for any particular individual. Route attractiveness (equivalently, cost) is typically defined as the sum of the costs of the constituent links. Introducing normally-distributed random error terms to the attractiveness (costs) at the link level gives rise to MVN distributed route costs with route overlap reflected in the correlation structure of the variance-covariance matrix.

The probit model can be used directly to account for inter-alternative correlation as well as heteroskedasticity across alternatives. Similarly, a treatment of taste heterogeneity (notwithstanding the work on error components) is straightforward, as long as the analyst is willing to rely on normally distributed coefficients. It should be noted that further flexibility can be obtained through a framework as in Equations (4), essentially equating to a mixed probit model, where the error structure of the probit model is used alongside any additional random terms needed for further flexibility, e.g. in the case where normally distributed coefficients are not advisable. This double integration clearly leads to further computational complexity. However, if all the random variations are of the normal form, and the model is linear in the coefficients, then the model can be evaluated as a single probit integral without further complexity. Similarly, complex inter-respondent correlation patterns arising in the case of panel data can be accommodated through additional layers of integration, although it should be noted that it is similarly possible to express a sequence of choices through a single probit integral by recomposing the choice set, albeit at an increased computational cost (see Appendix).

Independently of the treatment of any panel structure of the data, or the presence of any additional layers of integration, the core issue remains the computation of individual probit probabilities, given by multivariate normal integrals. Indeed, it should be noted that the attractiveness of the probit model, arising in particular from the flexibility offered by the MVN error covariance structure, should be balanced against (a) the
difficulty of computing the choice probabilities that do not have a closed form expression, and (b) the number of parameters that may need to be estimated and associated difficulties of identification (Dansie, 1985; Bunch, 1991; Keane, 1992). However, the analytical simplicity of the normal distribution, and in particular its survival under linear transformation, mean that, if all of the variation in the model follows normal distributions, the model can be expressed either through specifying the covariance matrix of the random terms, or, equivalently, by specifying the covariance matrix of the alternative utilities. This equivalence can be exploited to optimise numerical processing.

The purpose of this paper is to address the first issue, computing the choice probabilities which is a critical component of all analyses outlined above. Indeed, in all estimation problems, there is a need to evaluate the choice probabilities efficiently and accurately. Errors in calculating the individual choice probabilities result in a misrepresentation of the integrand and hence error in the computed volume under this hypersurface i.e. the likelihood function (however the integral is evaluated or approximated). This in turn will lead to the wrong likelihood function being optimised, possibly leading to biased parameter estimates. Accurate calculation of the choice probabilities therefore underlies the MLE algorithm. There is also widespread use of probit models for prediction, where the data and distribution of error terms are assumed known. In the analysis of transport networks for example, route choice is a key component. Variations in perception and exogenous factors (such as weather) mean that route travel times may be perceived differently by each driver, and hence travellers will choose between the available routes in some way that can be expressed as a RUM. A standard formulation would include attributes such as route length, travel time and toll, each with a specified distribution of associated preference parameters across the population of travellers. Computing the equilibrium flows requires travel demand to be split between the available routes (there may be tens of routes considered between each origin-destination). For a large network model computation of the equilibrium flows necessitates many repeated evaluations of route choice probabilities. Once again, errors in the calculation of the choice probabilities will have major impacts in this case.

The remainder of this paper is organised as follows. The next section presents multinomial probit methodology and discusses the specific analytical approximation techniques used in the work. This is followed by the empirical work in Section 3. Finally, Section 4 presents the conclusions from our work.

2. Methodology

Consider the issue of calculating the MVN integral necessary to determine choice probabilities for a probit model. The choice probability for a given alternative \( j \) from a choice set containing \( K \) elements \([j = 1,2,...,K]\) is given by

\[
P_j = \Pr[U_j > U_k; k = 1, ..., K, k \neq j]
\]

where the vector of random utilities \([U_j] = U\) is multivariate normal distributed with mean \( V \) and covariance \( \Sigma \).

\[
U \sim \text{MVN}(V, \Sigma)
\]

where utility comprises the deterministic utility given by \( V \), and, without loss of generality, a zero-mean random vector \( e \)

\[
U_j = V_j + e_j \text{ where } e \sim \text{MVN}(0, \Sigma)
\]

Thus

\[
P_j = \Pr[e_j - e_k > V_x - V_j; k \in C; k \neq j]
\]

and the differences \( Y_{jk} = e_j - e_k \) are also normally distributed. Define

\[
\Delta^j = [\delta_{lm}] \text{ where } \delta_{lm} = \begin{cases} -1 & \text{for } m = l \\ 1 & \text{for } l = j \\ 0 & \text{otherwise} \end{cases}, \quad l = 1,2,...,K; \ m = 1,2,...,K - 1
\]

The \( \Delta^j \) matrix needed to give the differences for computing \( P_j \) is generated from a \( K \times K \) zero matrix by putting -1 on the diagonal, setting the \( j \)-th row to 1s and deleting the \( j \)-th column. The \( K - 1 \) vector of differences \( y^j = \Delta^j \top e \) have zero mean and covariance \( \Omega^j = \Delta^j \top \Sigma \Delta^j \). Then with \( y^j = \Delta^j \top V \) we can write the \( j \)-th choice probability

\[
\]
\[ P_j = \Pr [Y^j < y^j] = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_{K-1}} \frac{1}{\sqrt{2\pi}^{K-1} |\Omega_j|} \exp \left[ -\frac{1}{2} y^{T}(\Omega_j)^{-1} y \right] dY_1 dY_2 \cdots dY_{K-1} \]  

(11)

where the upper limits are given by the \( y^j \). For simplicity of notation, we will now begin to drop the index \( j \). Finally we transform to standard (multivariate) normal by setting all variances to unity; writing the diagonal matrix

\[ S = \text{diag} \left( \frac{1}{\sqrt{\Omega_{1,1}}}, \ldots, \frac{1}{\sqrt{\Omega_{K-1,K-1}}} \right) \]  

(12)

then defining

\[ Z = S \cdot Y^j \]  

(13)

The covariance matrix for \( Z \) is \( \Sigma_Z = S \Omega_j S \), which then has ones along its diagonal. With \( z = S \cdot y^j \) the choice probability we wish to compute is now

\[ P_j = \Phi_{\Sigma_Z}(z). \]

with \( \Phi_{\Sigma} \) the standard multivariate normal CDF with covariances given by \( \Sigma \), i.e.

\[ P_j = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_{K-1}} \frac{1}{\sqrt{2\pi}^{K-1} |\Sigma_Z|} \exp \left[ -\frac{1}{2} Z^{T}\Sigma_Z^{-1} Z \right] dZ_1 dZ_2 \cdots dZ_{K-1} \]  

(14)

The choice probability for each alternative is hence equivalent to the evaluation of the CDF for the \( K - 1 \) dimensional standard multivariate normal with mean \( \theta \), unit variances and covariance \( \Sigma_Z \) (comprising correlations \( \rho_{ij} \)). All lower limits of integration are \( -\infty \) and the upper limits are finite but may take positive or negative values.

The choice probability for each of the \( K \) alternatives requires one evaluation of the CDF for the \( K - 1 \) dimensional standard multivariate normal. We therefore restrict attention to the case of an \( n \)-dimensional random variable \( Z \sim \text{MVN}(0, \Sigma) \) with unit variances and dimension \( n \), (for ease of notation set \( n = K - 1 \)). We wish to compute the probability

\[ P = \Pr [Z_1 < z_1, Z_2 < z_2, \ldots, Z_n < z_n] \]  

(15)

Numerical integration methods have been developed to accurately compute such MVN integrals (see Drezner and Wesolowsky, 1990; Genz, 1992, 2004; Drezner, 1994; Genz and Bretz, 1999, 2002); indeed such methods are already implemented in commercial software (e.g. MATLAB) offering off-the-shelf accuracy of \( 10^{-4} \) for each choice probability in choice sets having up to 25 alternatives. However, the execution time for these methods increases very rapidly with both the problem dimension and the accuracy demanded from them. For practical applications that require calculation of very many choice probabilities, these methods are not fast enough.

The standard approach seen in most applications is to evaluate the choice probabilities via crude frequency simulation (Manski and Lerman, 1981): the utilities of the alternatives are drawn from their distribution, the alternative with the highest utility is chosen. This is repeated a number of times, and the choice probability of each alternative is approximated by its choice frequency. This approach faces several difficulties: capturing low probability alternatives, the results depend on the seed used for the random number generator (i.e. ‘noise’ is present), and the frequencies are not continuous. The latter issue was tackled by McFadden (1989) who processed the simulated frequencies via a logit function. Possibly the most sophisticated and now most widely used simulation-based approach is the GHK probability simulator (see Börsch-Supan and Hajivassiliou, 1993). This approach rewrites the MVN integral as a product of marginal conditional probabilities, and evaluates these marginal probabilities in turn by Monte-Carlo simulation. As it is a probability simulator, GHK avoids the problems of estimating low probabilities, and the probabilities generated are continuous in the parameters. There remains the issue of noise common to all simulation approaches.
In this paper we investigate an alternative family of approaches to computing the probit choice probability MVN integrals, that of analytical approximation. Here the MVN integral interest is transformed into an approximating integral that can be evaluated more easily. Several analytic approximations have been proposed in the literature including the approximation of Clark (1961), Mendell and Elston (1974), Solow (1990) and Joe (1995), Langdon’s (1984a, 1984b) separated split procedure, the first order multi-normal approximations (Hohenbichler and Rackwitz, 1983; Tang and Melchers, 1986), and Taylor series approximations (Cox and Wermuth, 1991; Olson and Weissfeld, 1991). These methods have all been compared to one another within the specific context of algorithms for solving probit stochastic user equilibrium (Rosa, 2003, unpublished). This detailed study recommended the “Mendell-Elston” (ME) approach as providing the best compromise between speed and accuracy, and hence has subsequently been used in the context of network equilibrium (Connors et al., 2007). In contrast, Bhat’s recent work used a method due to Solow and Joe (Bhat, 2010), hereafter referred to as SJ. In this paper we will therefore restrict attention to just these two analytic approximation methods that we shall refer to as ME and SJ.

Mendell and Elston (1974) employ a result of Aitken (1935) to extend the bivariate results of Pearson (1903) to the multivariate case. Rewriting the MVN integral (15) in terms of conditional probabilities:

\[
P = Pr[Z_1 < z_1] \cdot Pr[Z_2 < z_2|Z_1 < z_1] \cdot Pr[Z_3 < z_3|[Z_1 < z_1, Z_2 < z_2]] \cdot \ldots \cdot Pr[Z_n < z_n|[Z_1 < z_1, \ldots, Z_{n-1} < z_{n-1}]]
\]

The ME approach is to approximate each univariate conditional distribution (that are not normal) with a normal distribution matching the mean and variance. Kamakura (1989) gives a clear account of the ME approximation for the evaluation of probit choice probabilities 1.

Joe (1995) presents several closely related methods, extending the work of Solow (1990) to the more general problem of computing probabilities for “rectangular areas” i.e.

\[
Pr(w_1 < Z_1 < z_1, w_2 < Z_2 < z_2, \ldots, w_n < Z_n < z_n) \text{ where } w_i, z_i \in \mathbb{R}
\]

We consider only the case of evaluating orthant probabilities, having infinite lower limits. For such cases, the simplest SJ approach is to write

\[
P = Pr[Z_2 < z_2, Z_1 < z_1] \cdot \prod_{i=3}^{n} Pr[Z_i < z_i|[Z_1 < z_1, Z_2 < z_2, \ldots, Z_{i-1} < z_{i-1}]]
\]

The conditional probabilities in the product are then approximated using indicator functions and the following result

If \(Y_1 \sim MVN(\mu_1, \Sigma_1), Y_2 \sim MVN(\mu_2, \Sigma_2)\) then \(E(Y_2|Y_1 = y_1) \approx \mu_2 + \Sigma_{21}(y_1 - \mu_1)\).

Extended versions of this approach were also proposed, founded on the evaluation of an exact trivariate or quadrivariate probability, in place of the bivariate used in (17), with remaining terms similarly collected as a product of conditional probabilities.

In both ME and SJ approaches the order of conditional terms may influence the result. For ME, Kamakura (1989) proposes ordering the terms within the calculation of each choice probability according to their covariances. Joe suggests that (for both ME and SJ approaches) an average is taken over many permutations of the variate order. Since the number of available permutations is \(m! / 2\). Joe suggests 100-10000 should be sufficient for high dimensional problems. Clearly averaging over multiple orderings will increase the computation time proportionately with the number of repetitions comprising the average.

Comparisons of analytic approximation methods have appeared in published articles, though Joe (1995) appears to be the only comparison including both ME and SJ. These tests considered problems of dimension 5 and 9, with constant positive correlation of 0.1 and then 0.4, computing the specific choice probabilities

1 The ME calculation scheme described contains a slight error
\[ P = Pr[Z_1 < z_1, Z_2 < z_2, \ldots, Z_n < z_n] \text{ with } z_1 = z_2 = \cdots = z_n = w \]

For the cases \( w = 0, 0.2, 0.4, 0.6, 0.8 \). Therefore, all tests comparing ME and SJ comprise a total of just 20 individual choice probabilities. The results indicate SJ is more accurate than ME. Computation time is not explicitly compared, though ME is noted to be faster.

In the remainder of this paper we present a wide range of numerical tests to investigate the accuracy and efficiency of both ME and SJ methods. In addition, to help the potential user to make comparisons, computations using Monte Carlo are also included.

3. Empirical work

3.1. Test Methodology

Tests carried out in most published literature consider the direct evaluation of the multivariate normal CDF over a (hyper-) rectangular region, or an orthant. In the tests presented here we consider MVN-distributed utilities whose differences will then be transformed into standard MVN variates (as outlined above), to calculate the choice probabilities for each alternative.

Each test therefore comprises the calculation of choice probabilities for several vectors of MVN-distributed utilities \( U \sim MVN(\mathbf{V}, \Sigma) \). The choice probabilities themselves, and hence the accuracy of each approximation method, will depend on the mean and the variance-covariance matrix. Specifically, it may be that some approximation methods do better/worse when faced with correlations, or large differences in the variances or in the means of the alternatives.

Two sets of diagonal terms for the covariance matrix are considered: the identity matrix \( \Sigma_I \), or variances increasing from 1 to 5 (in increments depending on the dimension) \( \Sigma_D \). Three alternatives for the off-diagonal terms are considered: zero, constant correlation \( \Sigma_C \) and “striped” with an AR(1) specification, \( \Sigma_A \):

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 4 & 0 \\
0 & 0 & 0 & 0 & 5
\end{bmatrix}, \Sigma_C =
\begin{bmatrix}
0 & 0.2 & 0.2 & 0.2 & 0.2 \\
0.2 & 0 & 0.2 & 0.2 & 0.2 \\
0.2 & 0.2 & 0 & 0.2 & 0.2 \\
0.2 & 0.2 & 0.2 & 0 & 0.2 \\
0.2 & 0.2 & 0.2 & 0.2 & 0
\end{bmatrix}, \Sigma_A =
\begin{bmatrix}
0 & 0.3 & 0.09 & 0.027 & 0.0081 \\
0.3 & 0 & 0.3 & 0.09 & 0.027 \\
0.09 & 0.3 & 0 & 0.3 & 0.09 \\
0.027 & 0.09 & 0.3 & 0 & 0.3 \\
0.0081 & 0.027 & 0.09 & 0.3 & 0
\end{bmatrix}
\]

The tests are built from several components

I. The number of alternatives (for each choice) determines the problem dimension \( N \).

II. Six variance-covariance matrices: \( \Sigma_I, \Sigma_I + \Sigma_C, \Sigma_I + \Sigma_A, \Sigma_D, \Sigma_D + \Sigma_C, \Sigma_D + \Sigma_A \).

III. For each covariance matrix, \( R \) mean vectors, \( \mathbf{V} \), will be tested. These are drawn from the uniform distribution \( U([-L, L]^N) \). For a fixed covariance matrix, a large range for the mean increases the occurrence of 0/1 probabilities.

IV. We investigate the change in accuracy resulting from averaging multiple “internal orderings” of the variates. These re-orderings are random permutations of the variate order.

Monte Carlo evaluation of the choice probabilities is by crude frequency simulation, using multivariate normal distributed pseudorandom numbers.
To illustrate the impact of sampling the mean, $V$, from different ranges, the figure above shows two bivariate utility distributions with the same covariance matrix. With the mean sampled from $U([-1,1]^2)$, marked by the dashed square, both alternatives have non-zero choice probabilities (assuming a realistic level of computational precision). With the mean sampled from $U([-10,10]^2)$, many cases will result in choice probabilities that are effectively 0/1. Such cases do not therefore rely on highly accurate estimation of that part of the PDF above or below the diagonal; an accurate calculation of the choice probabilities can be obtained by simply taking all or nothing.

A single computation returns the choice probabilities for a given $U^i \sim MVN(V, \Sigma)$. With $N$ alternatives these are $N$-vectors of choice probabilities: $P_{ml}^i$ from “exact” computation (using the numerical integration of MATLAB), $P_{sj}^i$ from the SJ approximation and $P_{me}^i$ from the ME. We consider two types of error, for the SJ method these are

$$
SJ\text{-Mean} = \text{mean}\{|P_{ml}^i - P_{sj}^i|, i = 1, \ldots, R\}
$$

$$
SJ\text{-Max} = \Sigma_{i=1}^R \max\{|P_{ml}^i - P_{sj}^i|\}
$$

The errors are therefore amalgamated across the $R$ choice situations. Similar notation is used for ME errors.

### 3.2. Role of Internal Averaging

The first set of tests considers the improvement gained by averaging over multiple reordering of the variates (as suggested by Joe 1995). The figures below show the dependency of errors in the choice probabilities on the number of re-orderings included in the average. For each number of internal re-orderings, the distribution of absolute errors is also displayed via boxplots on the right. This comprises a cross section of the data displayed on the left. The numbers of internal re-orderings used are irregularly spaced and the x-axis scale on the boxplots should be inspected carefully.
Many more tests were performed than are displayed here. ME significantly outperforms SJ, in fact the ME method with no averaging (a single internal ordering of terms) is more accurate than the SJ method with all permutations included.

For tests with increased range for $V \sim U([-1,1]^N)$ (equivalently, reducing the variance) the maximum element of $V$ has choice probability close to unity and the others close to zero; the shape and volume of the integrand becomes less important. In such cases the SJ method is found to be significantly more accurate than shown above, and is indistinguishable from the accuracy of the ME method. Note that both methods occasionally show an increase in the maximum error as more re-orderings are included; this unsurprising as the chance of a single very poor estimate increases.

A weakness of the SJ method arises from the need to invert a matrix relating to the covariance. Following the published SJ approach (as described in Joe 1995) can lead to ill conditioning and hence inaccurate results. For example, with $V$ sampled from $U([-10,10]^N)$ and covariance matrix $\Sigma_I + \Sigma_{C2}$ (the identity with off diagonal elements all = 0.5) we find erratic behaviour in the SJ method that is not addressed by increasing the number of internal re-orderings. It may be possible to design a different algorithmic implementation of the SJ approach to guard against this problem, but this is outside the scope of this paper.
3.3. Comparison of Execution Speed With Precision and Dimension

The plots below illustrate the trade off between execution speed and accuracy. Considered in turn are $N = 6,10,15,20$ with each problem dimension resulting in a pair of plots.

For each problem dimension 100 random draws of the deterministic utility are taken $V \sim U(-0.5, 0.5)$. These values are then fixed across all evaluations for this problem dimension.

For each method (ME,SJ,MC) six lines are plotted corresponding to the six choices of variance-covariance matrix: $\Sigma_1, \Sigma_1 + \Sigma_{C2}, \Sigma_1 + \Sigma_{A3}, \Sigma_D, \Sigma_D + \Sigma_{C2}, \Sigma_D + \Sigma_{A3}$ (defined in Section 3.1).

Each line plotted corresponds to a single choice of covariance matrix using one method; each line has three points on it. These three points correspond to increasing accuracy in that method. For ME and SJ this means increasing the number of internal re-orderings, with the three tests being 1,10,20 re-orderings. Earlier tests indicate that 20 internal re-orderings are sufficient to attain most of the accuracy increase from including re-orderings into the SJ method (for problems up to dimension 25). The MC tests use 10 000, 50 000, 100 000 draws in turn, giving three points on each MC line.

For a given method and covariance matrix, the choice probabilities are computed for these 100 $V$-vectors. The absolute errors for each individual choice probability are calculated (100 x number of alternatives choice probabilities). The left hand plots show the mean absolute error, the right hand plots show the maximum absolute error.

Not surprisingly the MC method maintains consistent execution times across all problem instances, whereas the execution for ME and SJ both increase substantially with problem dimension. The mean and maximum errors are significantly worse for SJ than ME in all scenarios. To allow comparison the axes have been scaled to reveal the detail, this means that often the SJ results have disappeared off to the top left. The y-limits of all plots are the same to aid comparison.

The six covariance matrices do influence the results, particularly for the ME method. However it is the contribution of the diagonal terms that dominate, regardless of the size and structure of off-diagonal terms. Therefore the legend and line-style plotted only distinguishes between the diagonal terms all being unity denoted “Var(1)” and the diagonal terms interpolating from 1 to 5 denoted “Var(1-5)”. 

![Graph with plots illustrating the comparison of execution speed with precision and dimension.](image)
3.4. Distribution of Error with Individual Probability Value

To investigate whether there are systematic errors in these approximation methods – are small probabilities, for example, always under or over-estimated?

These tests combine draws from $V \sim U([-1,1]^N)$ and $V \sim U([-10,10]^N)$ with $\Sigma_i$ plus each of the three covariance matrices. In each case 50 draws for the deterministic utility leads to 150 choice situations. For the SJ and ME methods we consider a single variate ordering; with 10 re-orderings an improvement in accuracy is gained corresponding to those illustrated in previous tests. For the MC calculation we use 50 000 draws for each choice situation. For each set of results the scale and range of the y-axis has been fixed for ease of comparison. The distribution of (the ‘exact’) probabilities is displayed for reference.
Increasing to 7 alternatives, we used 50 choice situations and each of the three covariance matrices (hence 1050 data points plotted for each method). With 10 alternatives we consider only 100 choice situations (draws for \( V \)) and hence plot 1000 choice probability data points for each method in each figure above.

### 3.5. Computing Derivatives

As Daganzo (1979, p72) explains (attributed to McFadden, 1978b), the (off diagonal) partial derivatives of the choice probabilities can be expressed exactly as the choice probability from a probit model with one fewer alternative.

\[
\frac{\partial P_i(V, \Sigma)}{\partial V_{i'}} = -\left( \frac{1}{2\pi|\Sigma|} \left| \Sigma(ij) \right|^{\frac{1}{2}} \exp \left( \frac{1}{2} K(ij) \right) \right) P_i(V(ij), \Sigma(ij))
\]

where \( i' = i - 1 \) if \( j < i \), \( i' = i \) if \( j > i \) and \( \Sigma(ij), K(ij), V(ij) \) are simple transformations of \( V \) and \( \Sigma \) with the \( j \)th rows and columns removed. The diagonal terms \( \partial P_i/\partial V_i \) can then be computed immediately from the off-diagonal derivatives. Hence fast and accurate calculation of the derivatives is equivalent to fast and accurate calculation of the choice probabilities as presented above.
For the estimation of choice models, the derivatives of the log-likelihood function are of interest, therefore requiring first derivatives of the choice probabilities with respect to the preference parameters. These are easily constructed from the above result via application of the chain rule to the specification of \( V \) in (1). In traffic networks, sensitivity analysis of the equilibrium flows can be an important tool and this relies on the Jacobian of the choice probabilities exactly as written above (see for example Connors et al., 2007).

For completeness, we also briefly present results showing the accuracy of using numerical differences to approximate these partial derivatives, when the choice probabilities being differenced have been calculated using ME, SJ or Monte Carlo.

We consider two problem dimensions, using 20 draws of \( V \sim U([-1,1])^6 \) and 5 draws of \( V \sim U([-1,1])^{10} \). The covariance matrices are \( \Sigma = \Sigma_I + \Sigma_{C2} + \Sigma_D + \Sigma_{A3} \). For a single \( V \), we construct the difference matrix (shown here for \( N=3 \) to conserve space)

\[
\Delta P = \frac{1}{\delta} \begin{bmatrix}
P_1(V) - P_1(V_1) & P_1(V) - P_1(V_2) & P_1(V) - P_1(V_3) \\
P_2(V) - P_2(V_1) & P_2(V) - P_2(V_2) & P_2(V) - P_2(V_3) \\
P_3(V) - P_3(V_1) & P_3(V) - P_3(V_2) & P_3(V) - P_3(V_3)
\end{bmatrix}
\]

(19)

Where \( V_i = V - \delta \times [0,...,0,1,0,...,0]^T \) with a one in the \( i \)-th place. This approximate Jacobian is calculated using choice probabilities according to each method. The ME and SJ methods use 10 internal repetitions. Calculations using Monte Carlo employ 50000 draws and reuse the same set of pseudo-random numbers for every calculation.

Note that errors in the ‘exact’ method may influence the results when differencing small probabilities. Therefore we record how many choice probabilities are small enough that they are within the computational error of the exact method (nS). The errors computed are mean absolute error, and max absolute error, averaged over all 60 combinations of \( V \) and \( \Sigma \).

<table>
<thead>
<tr>
<th>( \Sigma )</th>
<th>nS</th>
<th>Mendell-Elston</th>
<th>Monte-Carlo</th>
<th>Solow-Joe</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Sigma_I )</td>
<td>566</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AbsE= ( [0.1273,0.0124,0.0012] )</td>
<td>AbsE= ( [0.0370,0.0088,0.0028] )</td>
<td>AbsE= ( [1.7245,0.1658,0.0158] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MaxE= ( [0.4887,0.0458,0.0046] )</td>
<td>MaxE= ( [0.1222,0.0321,0.0117] )</td>
<td>MaxE= ( [4.9116,0.4742,0.0471] )</td>
</tr>
<tr>
<td>( \Sigma_I + \Sigma_{C2} )</td>
<td>551</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AbsE= ( [0.1380,0.0138,0.0014] )</td>
<td>AbsE= ( [0.0408,0.0096,0.0027] )</td>
<td>AbsE= ( [1.6295,0.1583,0.0161] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MaxE= ( [0.4680,0.0464,0.0047] )</td>
<td>MaxE= ( [0.1352,0.0388,0.0117] )</td>
<td>MaxE= ( [4.8932,0.4684,0.0469] )</td>
</tr>
<tr>
<td>( \Sigma_D + \Sigma_{A3} )</td>
<td>671</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AbsE= ( [0.4119,0.0405,0.0043] )</td>
<td>AbsE= ( [0.0333,0.0071,0.0022] )</td>
<td>AbsE= ( [1.2491,0.1253,0.0117] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MaxE= ( [1.4980,0.1459,0.0159] )</td>
<td>MaxE= ( [0.1069,0.0299,0.0093] )</td>
<td>MaxE= ( [3.8890,0.3902,0.0371] )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \Sigma )</th>
<th>nS</th>
<th>Mendell-Elston</th>
<th>Monte-Carlo</th>
<th>Solow-Joe</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Sigma_I )</td>
<td>525</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AbsE= ( [0.1034,0.0098,0.0011] )</td>
<td>AbsE= ( [0.0337,0.0063,0.0017] )</td>
<td>AbsE= ( [2.2385,0.2218,0.0200] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MaxE= ( [0.3233,0.0298,0.0037] )</td>
<td>MaxE= ( [0.1233,0.0323,0.0092] )</td>
<td>MaxE= ( [8.0835,0.7967,0.0684] )</td>
</tr>
<tr>
<td>( \Sigma_I + \Sigma_{C2} )</td>
<td>572</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AbsE= ( [0.1224,0.0117,0.0014] )</td>
<td>AbsE= ( [0.0340,0.0064,0.0019] )</td>
<td>AbsE= ( [2.3042,0.2743,0.0252] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MaxE= ( [0.3866,0.0379,0.0049] )</td>
<td>MaxE= ( [0.1258,0.0364,0.0109] )</td>
<td>MaxE= ( [7.7979,0.9589,0.0885] )</td>
</tr>
<tr>
<td>( \Sigma_D + \Sigma_{A3} )</td>
<td>840</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
<td>( \delta = [0.001,0.01,0.1] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AbsE= ( [0.3154,0.0305,0.0032] )</td>
<td>AbsE= ( [0.0320,0.0050,0.0012] )</td>
<td>AbsE= ( [1.3779,0.1346,0.0131] )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MaxE= ( [1.6156,0.1589,0.0161] )</td>
<td>MaxE= ( [0.1291,0.0259,0.0074] )</td>
<td>MaxE= ( [5.5829,0.5044,0.0510] )</td>
</tr>
</tbody>
</table>

The results depend on the value of \( \delta \) used to compute the differences. All tests show that SJ is the least accurate method. For \( \delta = 0.001 \) Monte Carlo is significantly better than ME, for \( \delta = 0.01 \) these two approaches are similar and at \( \delta = 0.1 \) the ME method is perhaps more accurate, though there is little to choose between them.
4. Conclusion

The multinomial probit model continues to be used widely in a network modelling context, for example as the basis for mode- and route-choice in computing network flows. In addition, there is renewed interest in probit in the choice modelling community as a result of recent work on composite marginal likelihood estimation (Bhat, 2010). Despite the growing focus on Mixed Logit in recent years, the Multinomial Probit model remains a structure with substantial theoretical appeal owing to its flexible error structure. The main factor in its relatively limited use has always been the computational cost. Even in the recent work by Bhat (2010), the computation of the multivariate normal integrals that the model is based on remains a core issue.

In this paper we have compared two analytical approximation methods for computing both the probit choice probabilities and their derivatives in terms of accuracy and computation efficiency. The first of these is the “Solow-Joe” (SJ) approach which was used in the recent work by Bhat (2010). The other is the “Mendell-Elston” (ME) approach. We believe that the results presented in this paper stem from the most thorough and in-depth comparison of the two methods in a multinomial probit context, and maybe also beyond.

The empirical results are based on a series of systematic tests of the two analytical approaches to the approximation of probit probabilities. While these tests are not exhaustive, they cover key aspects of the comparative performance of these procedures. The headline result is clear, in that the ME approach outperforms the SJ approach across a range of settings.

Indeed, the ME method appears to be more accurate than the SJ method; with no scenario found that showed ME to be systematically significantly worse than SJ, although the difference declined when many probabilities were close to 0 or 1. ME allows faster computation than SJ for the same number of internal repetitions and in many cases is more accurate than SJ with many fewer repetitions. Additionally, the results from Section 3.4 show that SJ is more inaccurate in computing small probabilities than ME (or indeed even MC) even ignoring the fact it is also slower, where the computation of such small probabilities is crucial in choice model estimation given the large influence that outlying choices can have on results. The difference in execution speed is primarily due to the fact that the SJ method requires evaluation of the bivariate normal CDF, whereas ME only requires univariate normal CDF evaluations. The “improved” versions of SJ (based on exact trivariate or quadrivariate MVN evaluations) will be very much slower. Both methods are simple to implement in computer code.

Each method is vulnerable. The SJ method can be highly inaccurate in the case of strongly correlated variates, when the variances are small in comparison with the mean. This arises due to the need to invert an ill conditioned matrix, which possibly could be addressed with a more sophisticated algorithm. The ME method appears to be stable in all problem scenarios, though can lose accuracy when faced with variates having very different variances.

There are many parameter dimensions to test in assessing whether one approximation method is more accurate than another. The tests reported here are a small but illustrative selection of those performed in this research, and further tests are of course always advisable. Moreover, we cannot offer bounds for the errors of these methods. Finally, for problems with few alternatives exact evaluation of choice probabilities is computationally feasible using established methods of numerical integration. For higher dimensional problems, ME offers a reasonably accurate and efficient method for computing choice probabilities.

With the strength of results presented here, it seems that the wider use of the “Mendell-Elston” method in probit work, and implementation within a composite marginal likelihood estimator are thus promising areas for new developments. While the primary motivation for this paper is to help with the estimation and application of discrete choice models, there may be applications in other areas where the computation of multivariate normal integrals is needed.

Appendix: Specification of MNP model with error terms correlated across choice tasks
Initially we follow Daganzo and Sheffi (1982) to show how to calculate the probability of a series of $T$ correlated probit probabilities among $J$ alternatives by calculating an exactly equivalent single choice probability from $(1 + T \ (J - 1))$ alternatives. For a given choice occasion $t$, define the matrix $J^{*} \ (J - 1)$

$$\Delta_t = [\Delta_{jkt}^c]$$

where $\Delta_{jkt}^c = 1$ if $j = k$ and $j < c$ or $j = k + 1$ and $j > c$

$$-1 \text{ if } j = c_t^2$$

$$0 \text{ otherwise;}$$

$c_t$ is the alternative chosen on the $t^{th}$ choice occasion.

The effect of post-multiplying by $\Delta_t$ is to convert variables defined over alternatives to differences between the chosen and unchosen alternatives (unchosen minus chosen).

We can extend this definition to cover all choice occasions by defining the $JT * (J-1)T$ block-diagonal matrix $\Delta$, which has $T$ blocks of $\Delta_t$, $t = 1, ..T$. Post-multiplying by $\Delta$ gives us differences over all $T$ choice occasions.

The key result from Daganzo and Sheffi is that the probability of observing the choice sequence

$$\text{Prob(choice sequence)} = \text{Prob}\{U_{c_1} \geq U_{j1} \forall j, \ldots, U_{c_T} \geq U_{jT} \forall j\}$$

$$= \text{Prob}\{U \Delta \leq 0\}$$

is obviously equal to the probability of choosing an alternative with utility 0 from a choice set of size $1 + T \ (J - 1)$ in which the utilities of the other alternatives are given by $U \Delta$.

The discussion above simply repeats the suggestion of Daganzo and Sheffi in the notation of the present paper. However, this idea can be taken further.

The model can conveniently be parameterised by the mean marginal value of the attributes $\beta$ and the covariance matrix of those parameters $\Omega = \Lambda \Lambda'$, where $\Lambda$ is a Cholesky form. The use of the Cholesky form is a convenient parameterisation to ensure the positive definiteness of $\Omega$. Regarding the utility mean and covariance, the following three points can be made.

1. The mean utilities are $\beta X \Lambda$, where $X$ is the observed attribute data.

2. The utility covariance can be seen as comprising two components. First, there is a between-individual component $(X \Lambda) \Omega_b (X \Lambda)$, where $\Omega_b$ is the part of the covariance matrix of the $\beta$s that relates to between-individual variation. This is what is anticipated in Daganzo and Sheffi.

3. Generalising this approach, however, there could also be a within-individual component, which may include both ‘white noise’ and parametrised variation. It appears to be reasonable to assume that the within and between variation is independent. Moreover, within-individual variation is, by definition, independent between the various choices made by the individual. The within-individual covariance matrix then has the form $\sum_t (X \Delta_t) \Omega_w (X \Delta_t)$, which would be added to the between variation.

Daganzo and Sheffi also open the possibility of allowing for ‘state dependence’, in which each choice is dependent on previous choices. However, this extension simply requires the $X$ for each choice to be a function of choices made on previous occasions. No further complication to the calculation of probabilities arises, though in model estimation the issue of endogeneity will require attention.

---

2 Note the error in the paper of Daganzo and Sheffi at this point (footnote 6).
Acknowledgment

This work described in this paper was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the project “Estimation of travel demand models from panel data”, grant EP/G033609/1. The second author also acknowledges the support of the Leverhulme Trust in the form of a Leverhulme Early Career Fellowship.

Bibliography


Finney, D.J., 1944b. The application of probit analysis to the results of mental tests. Psychometrika 9, 31–39.


