A New Approach to Specify and Estimate Non-Normally Mixed Multinomial Probit Models

Chandra R. Bhat*
The University of Texas at Austin
Dept of Civil, Architectural and Environmental Engineering
1 University Station C1761, Austin TX 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
E-mail: bhat@mail.utexas.edu

and

Raghuprasad Sidharthan
The University of Texas at Austin
Dept of Civil, Architectural and Environmental Engineering
1 University Station C1761, Austin TX 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
E-mail: raghu@mail.utexas.edu

*corresponding author

Original version: July 4, 2011
Revised version: February 11, 2012

ABSTRACT

The current paper proposes the use of the multivariate skew-normal distribution function to accommodate non-normal mixing in cross-sectional and panel multinomial probit (MNP) models. The combination of skew-normal mixing and the MNP kernel lends itself nicely to estimation using Bhat's (2011) maximum approximate composite marginal likelihood (MACML) approach. Simulation results for the cross-sectional case show that our proposed approach does well in recovering the underlying parameters, and also highlights the pitfalls of ignoring non-normality of the continuous mixing distribution when such non-normality is present. At the same time, the proposed model obviates the need to assume a pre-specified parametric distribution for the mixing, and allows the estimation of a very flexible, but still parsimonious, mixing distribution form.

Keywords: multinomial probit, mixed models, maximum approximate composite marginal likelihood, maximum simulated likelihood, multivariate skew-normal distribution

1. INTRODUCTION

Econometric discrete choice analysis is an essential component of studying individual choice behavior and is used in many diverse fields to model consumer demand for commodities and services. The decision principle used in almost all discrete choice models corresponds to utility maximization, which is based on the Lancastrian (1971) notion of the assignment of a composite utility to each alternative in the choice set (based on alternative and individual attributes) followed by the choice of the alternative with the highest utility. Further, since the analyst does not observe all individual and context-related factors that contribute to choice decisions, one or more stochastic elements (or random error terms) are introduced in the utility of alternatives. Different ways of introducing the stochastic elements lead to different discrete choice model structures. Thus, consider a cross-sectional choice situation with a single choice occasion per individual, and assume independence among the choice behaviors of individuals.[1] Then, the simplest model form, corresponding to the multinomial logit (MNL) model introduced by Luce and Suppes (1965) and McFadden (1974), assumes a *single composite* independently and identically distributed or IID (across alternatives) random utility error term with a Gumbel (or Type I extreme-value) distribution. This leads to the simple and elegant MNL model form, but also leaves the model form saddled with the familiar independence from irrelevant alternatives (IIA) property. Maintaining a single composite Gumbel error term in utilities, while relaxing the independence assumption (across alternatives), moves the model form from the multinomial logit to the generalized extreme-value (GEV) class of models proposed by McFadden (1978). On the other hand, relaxing the identically distributed assumption (across alternatives) with the Gumbel distribution assumption leads to the Heteroscedastic Extreme Value (HEV) model form proposed by Bhat (1995). Finally, still maintaining a single composite error term but now with a normal distribution, when combined with relaxation of the independence and/or identical distribution assumptions, generates the multinomial probit (MNP) model form originally proposed by Hausman and Wise (1978) and Daganzo (1979). Of these model forms, the MNP form allows the most flexible error covariance structures (up to certain limits of identifiability; see Train, 2009, Chapter 5), though it also entails more estimation effort since it requires the evaluation of a

[1] The use of a cross-sectional choice situation with independence across individual decision-maker choices is simply for exposition convenience in this introduction section.

multidimensional normal orthant probability function with an $(I-1)$ dimensional integral in the general case (where I is the number of alternatives).

A substantial amount of the early theoretical developments in discrete choice modeling was focused on a single composite error term. Over the past decade and a half, attention has shifted more toward the use of multiple error terms through the introduction of a mixing random distribution structure in the utility function of alternatives that is independent of the kernel error term. Essentially, the mixing structure superimposes additional stochastic terms over the "kernel" error term discussed in the previous paragraph. There are several reasons for this shift toward mixing structures. First, in a cross-sectional context, it is very plausible that there are unobserved variations across individuals in the sensitivity to relevant exogenous attributes (such as differential sensitivity due to unobserved factors to travel time and travel cost in a travel mode choice model). Ignoring these variable-specific stochasticity effects and instead using a single composite error term in the utility function will, in general, lead to inconsistent coefficient estimates and trade-off estimates, as well as incorrect substitution patterns across alternatives (see Bhat, 1997a).[2] A second reason for the increasing use of mixing structures is that they provide the ability to introduce heteroscedasticity across utilities in the closed-form GEV models through an error-components specification, as discussed in Train (2009). It also provides the ability to generate correlation across alternatives through an error-components specification. The use of a mixing structure over the closed-form GEV kernel-based model can then essentially achieve any desired covariance pattern. At the same time, and especially when the number of alternatives far exceeds the number of mixing random terms needed to capture the "true" covariance pattern, the maximum simulated likelihood (MSL) estimation of the mixed GEV model is generally much easier and faster than a non-mixed MNP model (see Bhat *et al*., 2008 and Train, 2009 for detailed discussions). A third reason for using mixing structures is that, when using GEV-based kernels, mixing structures enable the introduction of error dependencies across

[2] There are a few exceptions to this rule, one of which is when an MNP kernel error term is mixed with normally distributed random coefficients. Assuming the usual linear-in-parameters utility functional form, the net effect is that the combination of variable-specific random terms and the kernel error term can be recast back into an MNP utility form with a single composite error term (due to the closure property of the normal distribution under affine transformations -- a linear transformation followed by a translation). That is, the marginal distribution of utility obtained by integrating out the normal mixing distribution puts the utility back into a normal distribution form. In fact, this was the genesis of Hausman and Wise's MNP model formulation, in which the "composite" error terms of the alternatives have a covariance matrix that is parameterized based on the mixing structure. However, this kind of affine closure is not achieved with GEV or HEV kernel models. Further, closure is also not generally achieved with a non-normal mixing distribution with the MNP "kernel", except in a special case which is exploited in this paper.

the choice occasions of the same decision-maker in panel or repeated choice contexts (see Li *et al*., 2010). Even when using an MNP kernel, the mixing structure can provide substantial econometric and computational efficiency to capture panel effects. Further, the mixing approach is almost identical when dealing with cross-sectional choice data or panel data, and poses no conceptual and likelihood estimation coding differences.

There is yet another reason to consider a mixing approach in discrete choice modeling. This has to do with explicitly specifying the random mixing distribution on variables in a way that is consistent with theoretical notions. In fact, the ability to do so is critical to the observation made by McFadden and Train (2000) that the mixed multinomial logit model is capable of approximating any random utility maximization model. Thus, for example, one may want to consider bounded distributions (such as a log-normal distribution or a Rayleigh distribution) for cost and time coefficients in a travel mode choice model, so that the coefficients on these variables are bounded at the upper end. On the other hand, the coefficients on some other variables may be appropriately considered as being unbounded. Further, there are several types of continuous distributions that may be used to capture the profile of population sensitivity to variables.[3] In the context of continuous mixing distributions, the normal distribution has been used quite extensively in the past. However, several studies (see, for example, Amador *et al*., 2005, Train and Sonnier, 2005, Hensher *et al.*, 2005, Fosgerau, 2005, Greene *et al.*, 2006, Balcombe *et al.*, 2009, and Torres *et al.*, 2011) have underscored the potentially serious mis-specification consequences (in terms of theoretical considerations, data fit, as well as trade-off evaluations) of using the normal distribution. In particular, the symmetric nature of the normal distribution, when combined with mean values that may not be too far away from zero, implies that a significant fraction of individuals may have an unexpected sign on variables (such as a

[3] Note here that discrete distributions may also be used for the mixing. If the mixing vector is assumed to take M possible value states with state-specific probabilities, this leads to the familiar latent class model used in marketing (see Kamakura and Russell, 1989, Chintagunta *et al.*, 1991) and transportation (see Bhat, 1997b, Greene and Hensher, 2003, Hess *et al*., 2007, and Train, 2008). On the other hand, if a discrete distribution is considered separately for each individual random coefficient, this is essentially a non-parametric distribution (see Bastin *et al.*, 2010, Cherchi *et al.*, 2009, Fosgerau, 2006). However, the use of a continuous distribution dominates the literature, at least in part because it offers efficiency in the number of mixing distribution parameters to be estimated. Further several studies that have compared discrete distribution methods with continuous distributions have not found a clear pattern of which of the two approaches is superior (see, for instance, Greene and Hensher, 2003, Birol *et al.*, 2006, and Hynes *et al.*, 2008). Some recent studies have also considered a combination of discrete and continuous distributions for the mixture in the form of a mixture of normal distributions (see Campbell *et al.*, 2010), though such mixtures of normal distributions have some of the same problems as the simple normal distribution (as discussed subsequently).

positive coefficient on cost or time). For instance, Train and Sonnier (2005), in their analysis of vehicle choice, found that 22% of the population preferred vehicles with a higher purchase price, and 37% of the population preferred vehicles with a higher operating cost, when they used a normal distribution for the cost coefficients. On the other hand, when Train and Sonnier used a log-normal distribution and a bounded Johnson's SB distribution for the cost coefficients, such results were avoided and they also obtained better data fits. Finally, another issue with using normally distributed cost and other coefficients is that this leads to a breakdown of the WTP calculation because the moments of the ratio of two normally distributed random terms do not exist (see Cedilnik *et al*., 2006, Daly *et al*., 2011).

As indicated already, there have been several earlier studies that have successfully estimated non-normal distributions for the mixing distribution. All of these studies use a multinomial logit model kernel over which mixing is specified. However, the general experience has been that, even when successful, such estimations take a longer time for convergence (relative to normal distributions). This is particularly so for asymmetric distributions with long tails, such as the log-normal distribution. Further, in some cases, the maximum simulated likelihood (MSL) of models with non-normal mixing fails due to numeric/computational problems. It is not uncommon to see researchers consider non-normal distributions only to eventually revert to the use of a normal distribution (see, for example, Bartels *et al*., 2006 and Small *et al*., 2005). In addition to these problems specific to the use of non-normal distributions, MSL inference techniques can have other limitations, including a rapid degradation in accuracy as the number of dimensions of mixing increases, and problems with the accuracy (or lack thereof) of the covariance matrix of the estimator. These issues may be traced back to the use of a simulation approach to evaluate the log-likelihood function, which leads to a highly nonlinear and non-smooth second derivatives surface of the log-simulated likelihood function.

Recently, Bhat (2011) proposed an alternative maximum approximate composite marginal likelihood (MACML) inference approach to estimate the multinomial probit (MNP) model. His basis for preferring an MNP kernel rather than a multinomial logit or GEV kernel originates from several considerations. First, in cases such as a spatial analysis where the utility of spatial alternatives are correlated based on proximity, or in situations where the utility of individuals for alternatives have a spatial dependency component based on the usual spatial error/lag formulations used in spatial econometrics (see Anselin, 1988), the resulting parametric

covariance structure across alternatives or across decision-makers is simply infeasible or extremely inefficient to incorporate with a mixing approach over a restrictive Gumbel kernel covariance surface. Second, when a normal mixing distribution is used, the resulting "mixed MNP" model collapses back to an MNP model due to the closure property of the normal distribution under affine transformations. This, along with the MACML inference procedure, implies the need only to evaluate univariate and bivariate cumulative normal distribution function evaluations, regardless of the number of alternatives or the number of choice occasions per individual or the nature of social/spatial dependence structures. Further, the MACML procedure uses an *analytic approximation* method rather than a *simulation evaluation* method to evaluate the multivariate normal cumulative distribution function, which improves the ability to accurately and precisely recover the parameters and their covariance matrix estimates (because of the smooth nature of the first and second derivatives of the approximated analytic log-likelihood function). The net result is that the MNP kernel with the MACML inference approach leads to substantial computational gains compared to the MSL estimation of normally-mixed MNL and GEV models, as well as enables estimation in cases where the MSL estimation of mixed MNL and GEV approaches are simply infeasible.

One problem, however, with Bhat's MACML approach as it stands is that it is only applicable to the normally-mixed case. However, as discussed earlier, a normal mixing distribution may not be appropriate in several cases. What is needed then is a model that is able to include both a general covariance kernel structure as well as non-normal mixing, while also still being able to be estimated using the MACML approach. This is the objective of the current paper. Specifically, we introduce the use of a multivariate skew-normal distribution function for mixing with an MNP kernel model. The skew-normal distribution, considered by O'Hagan and Leonard (1976) and formalized by Azzalini (1985) for the univariate case, has been extended to the multivariate case by Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (1999). Since these initial contributions, more research on different types of multivariate generations of the skew-normal distribution and their properties have been undertaken (see Gonzalez-Farias *et al.*, 2004, Arellano-Valle and Genton, 2005, Gupta *et al.*, 2004, Arellano-Valle and Azzalini 2006, 2008, Azzalini, 2011). As discussed later, the multivariate skew normal (MSN) distribution retains several attractive properties of the multivariate normal distribution, and an MNP kernel model mixed with this distribution also lends itself nicely to estimation using the

MACML approach. At the same time, the MSN distribution is tractable, parsimonious in parameters that regulate the distribution and its skewness, and includes the normal distribution as a special interior point case. It also is a very flexible unimodal density structure that allows a "seamless" and "continuous" variation from normality to non-normality, and can replicate a variety of smooth unimodal density shapes with tails to the left or right as well as with a high modal value (sharp peaking) or low modal value (flat plateau). The asymmetry accruing from the skewness of the distribution also can allow the density to be pretty much confined to the positive (or negative) half-line. In this sense, it includes a likeness of the log-normal density function as a special case, but with tails that are thin as in the normal density function (which makes estimation easier than in the log-normal case). Despite these desirable properties, there has been little explicit consideration of the skew normal distribution for random terms even in the linear regression field with continuous observations (but see Jara *et al.,* 2008, Meintanis and Hlavka, 2010, and Molenaar *et al*., 2010), and there has been no consideration whatsoever of this distribution in the discrete choice field.[4]

The rest of this paper is structured as follows. The next section discusses the fundamental structure and properties of the univariate and multivariate skew normal distributions. The third section presents the model framework and estimation procedure for the proposed skew-normally mixed MNL model. Section 4 undertakes a simulation exercise to assess the ability of the proposed model to recover underlying parameters. Finally, Section 5 summarizes the key findings of the paper.

2. THE SKEW-NORMAL DISTRIBUTION

The literature on the skew-normal distribution is quite vast, but also scattered. In this section, we compile and present all the most relevant properties of the distribution in the context of application for mixed MNP models. The section begins with a characterization of the univariate skew-normal distribution and then proceeds to the more relevant case of the multivariate skew-normal distribution.

[4] However, it should be noted that the skew normal distribution has appeared implicitly in the context of such models as the stochastic frontier model (see Aigner *et al*., 1977) and in other studies involving the study of truncated normal variables (for example, Birnbaum, 1950 and Weinstein, 1964). This is because one of the stochastic representations of a skew-normally distributed variable happens to be as the convolution of a normal variable and a half-normal variable. However, the explicit use of the skew-normal as a distributional assumption for one or more random terms, as in the current paper, has seen little consideration in the econometric field.

2.1. The Univariate Skew-Normal Distribution

A random variable *Y* is labeled as being skew-normally distributed with a location parameter ξ ($\xi \in \Re$), a scale parameter ω ($\omega > 0$), and a shape parameter α ($\alpha \in \Re$) if its probability density function is as follows:

$$f(y;\xi,\omega^2,\alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left\{\alpha\left(\frac{y-\xi}{\omega}\right)\right\}, \tag{1}$$

where $\phi(.)$ and $\Phi(.)$ represent the standard normal density and cumulative distribution function, respectively. When $\alpha = 0$, the density collapses to that of a normal distribution with mean and variance parameters of ξ and ω^2, respectively. Setting Y = $\xi + \omega Z$, we obtain a standardized version of the probability density function of the skew-normal distribution (corresponding to the density function of Z that has a location parameter of 0 and scale parameter of 1) given by $\widetilde{\phi}(z;\alpha) = 2\phi(z)\Phi(\alpha z)$. The density function for Y in Equation (1) may be written in terms of the standard density function as $\omega^{-1}\widetilde{\phi}(z;\alpha)$, where $z = \omega^{-1}(y-\xi)$. Appendix A.1 presents the moment generating function and the moments of the standardized skew-normal distribution (SSN).

An important stochastic representation for *Z* that is useful for random generation from the SSN distribution is obtained using a conditioning mechanism. Specifically, consider two bivariate normally distributed variables M_1 and M_2:

$$\begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{2}$$

Then, $Z = M_2 \,|\, (M_1 > 0)$ has the SSN density function $\widetilde{\phi}(z;\alpha)$, where the relationship between ρ and α is as follows: $\rho = \frac{\alpha}{\sqrt{1+\alpha^2}}$ (see Appendix A.2 for a derivation). Using this conditioning mechanism, the cumulative distribution function for Z may be obtained as follows:

$$\begin{aligned} P(Z < z) = \widetilde{\Phi}(z;\,\alpha) &= \frac{P(M_1 > 0, M_2 < z)}{P(M_1 > 0)} \\ &= 2\,P(-M_1 < 0, M_2 < z) = 2\Phi_2(0, z, -\rho); \rho = \frac{\alpha}{\sqrt{1+\alpha^2}}. \end{aligned} \tag{3}$$

Thus, the cumulative SSN distribution function may be written in terms of a bivariate cumulative standard normal distribution function, and the cumulative distribution function for the non-standardized skew-normally distributed variable *Y* may be obtained as:

$$P(Y < y) = \widetilde{\Phi}\left(\frac{y-\xi}{\omega};\ \alpha\right) = 2\Phi_2\left(0, \frac{y-\xi}{\omega}, -\rho\right); \rho = \frac{\alpha}{\sqrt{1+\alpha^2}}. \quad (4)$$

For the extension to the multivariate skew-distribution, and especially for use with the multinomial probit model, an alternate parameterization of *Z* (referred to by Arellano-Valle and Azzalini, 2006 as the unified skew-normal variable) will be helpful. This is based on the conditioning mechanism discussed above. In this alternate parameterization, the univariate SSN density function is written as $\widetilde{\phi}(z;\rho)$ and the univariate cumulative distribution function is written as $\widetilde{\Phi}(z;\rho)$, with ρ replacing α.

Figure 1 shows the shapes of the normal density function (solid line) and the SSN density functions for three positive values of ρ (the plots are mirrored across the y-axis for negative values of ρ). As the value of the shape parameter ρ increases, the skewness of the distribution increases and the density shows sharper peaking. As $\rho \to 1$, the SSN density tends toward a half-normal density function. Note also that, as the shape parameter increases, the right skewness increases not because the extreme right tail gets longer but because the left tail becomes shorter and shorter (relative to the normal distribution). This is a desirable property in the likelihood convergence of mixed models, and is unlike the log-normal distribution whose right tail gets very long rapidly as the variance of the distribution increases.

2.2. The Multivariate Skew-Normal Distribution Function

There are several multivariate versions of the skew-normal distribution in the literature (see Arellano-Valle and Azzalini, 2006 for a discussion of these many variants, and a unified treatment of these). All of these share several properties similar to the multivariate normal distribution. In this paper, we select the multivariate skew distribution version originally proposed by Azzalini and Dalla Valle (1996) for a number of reasons. This version is efficient in the number of additional parameters to be estimated, allows independence between skew-normally distributed and normally-distributed elements in a multivariate vector (useful in selectively imposing skew-normality only on certain coefficients), is closed under any affine

transformation of the skew-normally distributed vector (is the key to the MACML estimation of the MNP model), and is closed under the sum of independent skew-normally distributed and normally distributed vectors of the same dimensions (is the key to non-normally mixing distributions superimposed on an MNP kernel). As importantly, the cumulative distribution function of a *D*-variate skew normally distributed variable of the Azzalini and Dalla Valle type requires only the evaluation of a $(D+1)$-dimensional multivariate cumulative normal distribution function.

Consider a multivariate skew-normally (MVSN) distributed random variable vector $\mathbf{Y} = (Y_1, Y_2, Y_3, \ldots, Y_D)'$ with a $(D \times 1)$-location parameter vector ξ $(\xi \in \Re^D)$, and a $(D \times D)$-symmetric positive-definite covariance matrix $\mathbf{\Omega}$. Let the correlation matrix corresponding to $\mathbf{\Omega}$ be $\mathbf{\Omega}^*$, and let $\boldsymbol{\omega}$ be a $(D \times D)$-diagonal matrix formed by the standard deviations of $\mathbf{\Omega}$ (ω_j is the *j*th diagonal element of the matrix $\boldsymbol{\omega}$). Then, we may write: $\mathbf{\Omega}^* = \boldsymbol{\omega}^{-1}\mathbf{\Omega}\boldsymbol{\omega}^{-1}$. Setting $\mathbf{Y} = \boldsymbol{\xi} + \boldsymbol{\omega}\mathbf{Z}$, we obtain a standardized version of the multivariate probability density function of the skew-normal distribution (corresponding to the density function of $\mathbf{Z}$ that has a location parameter of $\mathbf{0}$ and a correlation matrix $\mathbf{\Omega}^*$). As in the univariate case, it can be shown that the random variable $\mathbf{Z}$ is obtained through a latent conditioning mechanism on a $(D+1)$-variate normally distributed vector $\widetilde{\mathbf{M}} = (\widetilde{M}_1, \widetilde{\mathbf{M}}_2')'$, where $\widetilde{M}_1$ is a latent (1×1)-vector and $\widetilde{\mathbf{M}}_2$ is a $(D \times 1)$-vector:

$$\begin{pmatrix} \widetilde{M}_1 \\ \widetilde{\mathbf{M}}_2 \end{pmatrix} \sim N_{d+1}\left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \mathbf{\Omega}_+^*\right), \text{ where } \mathbf{\Omega}_+^* = \begin{pmatrix} 1 & \boldsymbol{\rho}' \\ \boldsymbol{\rho} & \mathbf{\Omega}^* \end{pmatrix}. \tag{5}$$

$\boldsymbol{\rho}$ is a $(D \times 1)$-vector, each of whose elements must lie between –1 and +1. The matrix $\mathbf{\Omega}_+^*$ is also a positive-definite correlation matrix. Then, $\mathbf{Z} = \widetilde{\mathbf{M}}_2 \mid (\widetilde{M}_1 > 0)$ has the standard multivariate skew-normal (SMVSN) density function shown below:

$$\widetilde{\phi}_D(\mathbf{z}; \mathbf{\Omega}_+^*) = 2\phi_D(\mathbf{z}; \mathbf{\Omega}^*)\,\Phi(\boldsymbol{\alpha}'\mathbf{z}), \text{ where } \boldsymbol{\alpha} = \frac{(\mathbf{\Omega}^*)^{-1}\boldsymbol{\rho}}{\left(1 - \boldsymbol{\rho}'(\mathbf{\Omega}^*)^{-1}\boldsymbol{\rho}\right)^{1/2}}. \tag{6}$$

where $\phi_D(.)$ and $\Phi(.)$ represent the standard multivariate normal density function of *D* dimensions and the standard univariate cumulative distribution function, respectively. We write

$\mathbf{Z} \sim \text{SMVSN}(\boldsymbol{\Omega}_+^*)$. The probability density function of the random variable $\mathbf{Y}$ $[\mathbf{Y} \sim \text{MVSN}(\boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\Omega}_+^*)]$ may be written in terms of the SMVSN density function above as:

$$f_D(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\Omega}_+^*) = \left(\prod_{j=1}^{D} \omega_j \right)^{-1} \widetilde{\phi}_D(\mathbf{z}; \boldsymbol{\Omega}_+^*), \text{ where } \mathbf{z} = \boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\xi}). \tag{7}$$

The moment generating function of $\mathbf{Z}$ and its first three moments are presented in Appendix A.3. The cumulative distribution function for $\mathbf{Z}$ may be obtained as:

$$P(\mathbf{Z} < \mathbf{z}) = \widetilde{\Phi}_D(\mathbf{z}; \boldsymbol{\Omega}_+^*) = 2\Phi_{D+1}(\mathbf{0}, \mathbf{z}, \boldsymbol{\Omega}_-^*); \ \boldsymbol{\Omega}_-^* = \begin{pmatrix} 1 & -\boldsymbol{\rho}' \\ -\boldsymbol{\rho} & \boldsymbol{\Omega}^* \end{pmatrix}. \tag{8}$$

The corresponding cumulative distribution function for $\mathbf{Y}$ is:

$$P(\mathbf{Y} < \mathbf{y}) = \widetilde{\Phi}_D\left(\boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\xi}); \boldsymbol{\Omega}_+^*\right) = 2\Phi_{D+1}\left(0, \boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\xi}), \boldsymbol{\Omega}_-^*\right). \tag{9}$$

The close correspondence with the normal distribution leads to several desirable properties of the multivariate skew-normal (MVSN) distribution. The ones that are key to the proposal in this paper to use the MSN distribution for mixing in MNP models are listed and discussed below.

Property 1:

The sum of a MVSN distributed vector $\mathbf{Y}$ (of dimension $D \times 1$) $[\mathbf{Y} \sim \text{MVSN}(\boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\Omega}_+^*)]$ and an independently distributed multivariate normally (MVN) distributed vector $\mathbf{W}$ (also of dimension $D \times 1$) $[\mathbf{W} \sim \mathbf{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})]$ is still MVSN distributed:

$$\mathbf{Y} + \mathbf{W} \sim \text{MVSN}(\boldsymbol{\xi} + \boldsymbol{\mu}, \widetilde{\boldsymbol{\omega}}, \widetilde{\boldsymbol{\Omega}}_+^*), \qquad \text{where} \qquad \widetilde{\boldsymbol{\Omega}}_+^* = \begin{pmatrix} 1 & \widetilde{\boldsymbol{\rho}}' \\ \widetilde{\boldsymbol{\rho}} & \widetilde{\boldsymbol{\Omega}}^* \end{pmatrix}, \ \widetilde{\boldsymbol{\Omega}}^* = (\widetilde{\boldsymbol{\omega}})^{-1} \widetilde{\boldsymbol{\Omega}} (\widetilde{\boldsymbol{\omega}})^{-1}, \widetilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega} + \boldsymbol{\Sigma},$$

$\widetilde{\boldsymbol{\rho}} = (\widetilde{\boldsymbol{\omega}})^{-1} \boldsymbol{\omega} \boldsymbol{\rho}$, and $\widetilde{\boldsymbol{\omega}}$ is the diagonal matrix of standard deviations of $\widetilde{\boldsymbol{\Omega}}$.

Proof: There are several ways to prove this property, but perhaps the easiest is to use the moment generating functions of $\mathbf{Y}$ and $\mathbf{W}$. Specifically, we have (from Appendix A.3):

$$M_{Y+W}(\mathbf{t}) = M_Y(\mathbf{t}) \times M_W(\mathbf{t}) = \left[2\exp\left(\boldsymbol{\xi}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Omega}\mathbf{t} \right) \Phi(\boldsymbol{\rho}'\boldsymbol{\omega}\,\mathbf{t}) \right] \times \left[\exp\left(\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} \right) \right]$$
$$= \left[2\exp\left((\boldsymbol{\xi}' + \boldsymbol{\mu}')\mathbf{t} + \frac{1}{2}\mathbf{t}'(\boldsymbol{\Omega} + \boldsymbol{\Sigma})\mathbf{t} \right) \Phi(\tilde{\boldsymbol{\rho}}'\tilde{\boldsymbol{\omega}}\,\mathbf{t}) \right], \text{ where } \tilde{\boldsymbol{\rho}} = (\tilde{\boldsymbol{\omega}})^{-1}\boldsymbol{\omega}\boldsymbol{\rho}. \tag{10}$$

The above expression is once again in the MVSN moment generating form in Appendix (A.3), from which the property is proved.

Property 2:

The affine transformation of the MVSN distributed vector $\mathbf{Y}$ (of dimension $d \times 1$) $[\mathbf{Y} \sim \text{MVSN}(\boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\Omega}^*_+)]$ as $\mathbf{a} + \mathbf{B}\mathbf{Y}$, where $\mathbf{B}$ is a $(h \times d)$ matrix is also a MVSN distributed vector of dimension $h \times 1$:

$$\mathbf{a} + \mathbf{B}\mathbf{Y} \sim \text{MVSN}(\mathbf{a} + \mathbf{B}\boldsymbol{\xi}, \tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\Omega}}^*_+)], \text{ where } \tilde{\boldsymbol{\Omega}}^*_+ = \begin{pmatrix} \mathbf{1} & \tilde{\boldsymbol{\rho}}' \\ \tilde{\boldsymbol{\rho}} & \tilde{\boldsymbol{\Omega}}^* \end{pmatrix}, \tilde{\boldsymbol{\Omega}}^* = (\tilde{\boldsymbol{\omega}})^{-1}\tilde{\boldsymbol{\Omega}}(\tilde{\boldsymbol{\omega}})^{-1}, \tilde{\boldsymbol{\Omega}} = \mathbf{B}\boldsymbol{\Omega}\mathbf{B}',$$

$\tilde{\boldsymbol{\rho}} = (\tilde{\boldsymbol{\omega}})^{-1}\mathbf{B}\boldsymbol{\omega}\,\boldsymbol{\rho}$, and $\tilde{\boldsymbol{\omega}}$ is the diagonal matrix of standard deviations of $\tilde{\boldsymbol{\Omega}}$.

Proof: The moment generating function of $\mathbf{a} + \mathbf{B}\mathbf{Y}$ may be written as:

$$M_{\mathbf{a}+\mathbf{B}\mathbf{Y}}(\mathbf{t}) = M_{\mathbf{Y}}(\mathbf{a} + \mathbf{B}'\mathbf{t}) = \left[2\exp\left((\mathbf{a} + \mathbf{B}\boldsymbol{\xi})'\mathbf{t} + \frac{1}{2}\mathbf{t}'(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}')\mathbf{t} \right) \Phi(\boldsymbol{\rho}'\boldsymbol{\omega}\mathbf{B}'\mathbf{t}) \right]$$
$$= \left[2\exp\left((\mathbf{a} + \mathbf{B}\boldsymbol{\xi})'\mathbf{t} + \frac{1}{2}\mathbf{t}'(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}')\mathbf{t} \right) \Phi(\tilde{\boldsymbol{\rho}}'\tilde{\boldsymbol{\omega}}\,\mathbf{t}) \right], \text{ where } \tilde{\boldsymbol{\rho}} = (\tilde{\boldsymbol{\omega}})^{-1}\mathbf{B}\boldsymbol{\omega}\,\boldsymbol{\rho} \tag{11}$$

This proves the result. The two properties above provide the marginal distribution of the utilities under a MNP kernel mixed with skew normally distributed and normally distributed random coefficients, which is critical to the MACML estimation of the resulting model, as discussed next.

3. THE MODEL FRAMEWORK

We develop the model framework first in the context of a cross-sectional MNP model and then discuss the panel formulation. However, the skew-normal mixing can also be imposed on any other form of the MNP model, including settings with spatial dependencies and social

dependencies across decision units, and combinations of temporal, spatial, and social dependencies.

3.1. Cross-Sectional MNP Formulation and Estimation

Consider a random-coefficients formulation in which the utility that an individual q $(q = 1, 2, ..., Q)$ associates with alternative i $(i = 1, 2, ..., I)$ is written as:

$$U_{qi} = \boldsymbol{\beta}_q' \mathbf{x}_{qi} + \boldsymbol{\gamma}_q' \mathbf{s}_{qi} + \tilde{\varepsilon}_{qi},\ \boldsymbol{\beta}_q \sim \text{MVSN}(\mathbf{b}, \boldsymbol{\omega}, \boldsymbol{\Omega}_+^*),$$
$$\boldsymbol{\Omega}_+^* = \begin{pmatrix} \mathbf{1} & \boldsymbol{\rho}' \\ \boldsymbol{\rho} & \boldsymbol{\Omega}^* \end{pmatrix}, \boldsymbol{\Omega}^* = (\boldsymbol{\omega})^{-1} \boldsymbol{\Omega} (\boldsymbol{\omega})^{-1}, \boldsymbol{\gamma}_q \sim \text{MVN}(\mathbf{c}, \boldsymbol{\Sigma}), \tilde{\varepsilon}_{qi} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Psi}) \quad (12)$$

where $\mathbf{x}_{qi}$ is a $(D \times 1)$-column vector of exogenous attributes, $\mathbf{s}_{qi}$ is another $(K \times 1)$-column vector of exogenous attributes (including dummy variables for constants, except in one of the I alternative utilities), $\boldsymbol{\beta}_q$ is an individual-specific $(D \times 1)$-column vector of MVSN-distributed coefficients that varies across individuals based on unobserved individual attributes, $\boldsymbol{\gamma}_q$ is another individual-specific $(D \times 1)$-column vector of MVN-distributed coefficients that varies across individuals based on unobserved individual attributes (but with the coefficients on the dummy variables for the constants maintained as fixed coefficients in the vector $\boldsymbol{\gamma}_q$), and $\tilde{\boldsymbol{\varepsilon}}_q = (\tilde{\varepsilon}_{q1}, \tilde{\varepsilon}_{q2}, \tilde{\varepsilon}_{q3}, \ldots, \tilde{\varepsilon}_{qI})'$ is assumed to have a general covariance structure subject to identifiability considerations (let $\tilde{\boldsymbol{\varepsilon}}_\mathbf{q} \sim \text{MVN}_I(\mathbf{0}, \tilde{\boldsymbol{\Psi}})$). In many situations, such as in a path choice model (see Yai *et al*., 1997) or a model with spatial location alternatives (see Bhat and Guo, 2007), a specific parametric structure, based on theoretical considerations appropriate to the context, can be placed on $\boldsymbol{\Psi}$. Similarly, in a pure random coefficients specification (as in Hausman and Wise, 1978), one may consider $\boldsymbol{\Psi}$ to be an identity matrix (or an identity matrix scaled by 0.5 or any other constant). Such specifications help in econometric identification as well as econometric efficiency. If a general covariance structure is adopted, there are many ways to ensure identification. An appealing approach is to take the differences of the error terms with respect to the first error term. Let $\varepsilon_{qi1} = (\tilde{\varepsilon}_{qi} - \tilde{\varepsilon}_{q1})$, and let $\boldsymbol{\varepsilon}_{\mathbf{q1}} = (\varepsilon_{q21}, \varepsilon_{q31}, ..., \varepsilon_{qI1})$. Then, up to a scaling factor, the covariance matrix of $\boldsymbol{\varepsilon}_{q1}$ (say $\boldsymbol{\Psi}_\mathbf{1}$) is identifiable. Next, scale the top left diagonal element of this error-differenced covariance matrix to 1. Thus, there are

$[(I-1)\times(I/2)]-1$ free covariance terms in the $(I-1)\times(I-1)$ matrix $\boldsymbol{\Psi}_1$. Finally, to ensure that whenever differences are taken with respect to the chosen alternative during the maximum approximate composite marginal likelihood (MACML) estimation, these differences are consistent with the same error covariance matrix $\widetilde{\boldsymbol{\Psi}}$ for the undifferenced error term vector $\widetilde{\boldsymbol{\varepsilon}}_q$, $\widetilde{\boldsymbol{\Psi}}$ is constructed from $\boldsymbol{\Psi}_1$ by adding a top row of zeros and a first column of zeros (see Train, 2003; page 134). During the MACML estimation, then, we can obtain the $(I-1)\times(I-1)$ covariance matrix of the error differences taken with respect to the *m*th alternative as $\boldsymbol{\Psi}_{\mathbf{m}} = \boldsymbol{\Gamma}_{\mathbf{m}}\widetilde{\boldsymbol{\Psi}}\boldsymbol{\Gamma}'_{\mathbf{m}}$, where $\boldsymbol{\Gamma}_{\mathbf{m}}$ is a $(I-1)\times I$ matrix which corresponds to the identity matrix of size $(I-1)$ with an extra column of –1's added as the *m*th column.

In Equation (12), we will assume that the random vectors $\boldsymbol{\beta}_q$, $\boldsymbol{\gamma}_q$, and $\widetilde{\boldsymbol{\varepsilon}}_q$ are independent of each other for each individual, as well as that these vectors are independent of the corresponding coefficients of other individuals (this latter assumption can be relaxed within our modeling framework, as will be needed for accommodating spatial or social dependency effects). From the earlier definitions, we can write $\boldsymbol{\beta}_q = \mathbf{b} + \breve{\boldsymbol{\beta}}_q$ with $\breve{\boldsymbol{\beta}}_q \sim \text{MVSN}(\mathbf{0}, \boldsymbol{\omega}, \boldsymbol{\Omega}^*_+)$, and $\boldsymbol{\gamma}_q = \mathbf{c} + \breve{\boldsymbol{\gamma}}_q$ with $\breve{\boldsymbol{\gamma}}_q \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$. Also let $\mathbf{U}_q = (U_{q1}, U_{q2}, ..., U_{qI})'$ ($I\times 1$ vector), $\mathbf{x}_q = (\mathbf{x}_{q1}, \mathbf{x}_{q2}, ..., \mathbf{x}_{qI})'$ ($I \times D$ matrix), and $\mathbf{s}_q = (\mathbf{s}_{q1}, \mathbf{s}_{q2}, ..., \mathbf{s}_{qI})'$ ($I \times K$ matrix). Then, we can write:

$$\mathbf{U}_q = \left[\mathbf{x}_q\mathbf{b} + \mathbf{s}_q\mathbf{c}\right] + \left[\mathbf{x}_q\breve{\boldsymbol{\beta}}_q + \mathbf{s}_q\breve{\boldsymbol{\gamma}}_q + \widetilde{\boldsymbol{\varepsilon}}_q\right], \tag{13}$$

Let $[.]_e$ indicate the e^{th} element of the column vector [.]. Equation (12) can equivalently be written using Equation (13) as:

$$U_{qi} = \left[\mathbf{x}_q\mathbf{b} + \mathbf{s}_q\mathbf{c}\right]_i + \left[\mathbf{x}_q\breve{\boldsymbol{\beta}}_q + \mathbf{s}_q\breve{\boldsymbol{\gamma}}_q + \widetilde{\boldsymbol{\varepsilon}}_q\right]_i, \tag{14}$$

Define $V_{qi} = [\mathbf{x}_q\mathbf{b} + \mathbf{s}_q\mathbf{c}]_i$ and $\boldsymbol{\varepsilon}_{qi} = [\mathbf{x}_q\breve{\boldsymbol{\beta}}_q + \mathbf{s}_q\breve{\boldsymbol{\gamma}}_q + \widetilde{\boldsymbol{\varepsilon}}_q]_i$. Also, assume that individual *q* chooses alternative m_q. In the utility differential form, we may write Equation (14) as:

$$u^*_{qim_q} = U_{qi} - U_{qm_q} = H_{qim_q} + \xi_{qim_q};\ H_{qim_q} = V_{qi} - V_{qm_q} \text{ and } \xi_{qim_q} = \varepsilon_{qi} - \varepsilon_{qm_q};\ i \neq m_q \tag{15}$$

Then stack the utility differentials $u^*_{qim_q}$ $(=U_{qi}-U_{qm_q}, i\neq m_q)$ in the following order: $\mathbf{u}^*_q=(u^*_{q1m_q},u^*_{q2m_q},...,u^*_{qIm_q})'$, an $(I-1)\times 1$ vector. Correspondingly, let $\mathbf{H_q}=(H_{q1m_q},H_{q2m_q},...,H_{qIm_q})'$, an $(I-1)\times 1$ vector, and define $\overline{\mathbf{\Omega}}_q=\mathbf{x}_q\mathbf{\Omega}\mathbf{x}'_q$ ($I\times I$ matrix), $\overline{\mathbf{\Sigma}}_q=\mathbf{s}_q\mathbf{\Sigma}\mathbf{s}'_q$ ($I\times I$ matrix) and $\mathbf{F}_q=\left[\overline{\mathbf{\Omega}}_q+\overline{\mathbf{\Sigma}}_q+\widetilde{\mathbf{\Psi}}\right]$. Based on properties 1 and 2 earlier in the paper, we can derive the location and other parameters of the vector $\mathbf{u}^*_q$, which is also skew-normally distributed. Specifically, by successive applications of property 2 and then property 1, we obtain the following important result:

$$\mathbf{u}^*_q\sim \text{MVSN}(\mathbf{H}_q,\breve{\boldsymbol{\omega}}_q,\breve{\mathbf{\Omega}}_{q_+}{}^*)], \tag{16}$$

$$\breve{\mathbf{\Omega}}_{q_+}{}^*=\begin{pmatrix}1 & \breve{\boldsymbol{\rho}}'_q\\ \breve{\boldsymbol{\rho}}_q & \breve{\mathbf{\Omega}}_q{}^*\end{pmatrix},\ \breve{\mathbf{\Omega}}_q{}^*=(\breve{\boldsymbol{\omega}}_q)^{-1}\breve{\mathbf{\Omega}}_q(\breve{\boldsymbol{\omega}}_q)^{-1},\breve{\mathbf{\Omega}}_q=\mathbf{\Gamma}_{m_q}\mathbf{F}_q\mathbf{\Gamma}'_{m_q},\ \breve{\boldsymbol{\rho}}_q=(\breve{\boldsymbol{\omega}}_q)^{-1}(\mathbf{\Gamma}_{m_q}\mathbf{x}_q)\boldsymbol{\omega}\boldsymbol{\rho}. \tag{17}$$

$\breve{\boldsymbol{\omega}}_q$ is the diagonal matrix of standard deviations of $\breve{\mathbf{\Omega}}_q$. The parameters to be estimated include the **b** and **c** vectors, the elements of the covariance matrices $\mathbf{\Omega}$, $\mathbf{\Sigma}$, and $\widetilde{\mathbf{\Psi}}$, and the $\boldsymbol{\rho}$ parameter vector. Collect all these elements into a single vector $\boldsymbol{\theta}$. Then, one can use the result above to obtain the likelihood contribution of individual *q* choosing alternative *m,* which takes the *I*-dimensional integral form below:

$$L_q(\boldsymbol{\theta})=P(\mathbf{u}^*_q<0)=\widetilde{\Phi}_{I-1}\left((\breve{\boldsymbol{\omega}}_q)^{-1}(-\mathbf{H}_q);\breve{\mathbf{\Omega}}_{q_+}{}^*\right)=2\Phi_I\left(0,(\breve{\boldsymbol{\omega}}_q)^{-1}(-\mathbf{H}_q),\mathbf{\Omega}^*_{q-}\right). \tag{18}$$

It is straightforward to see that if all the elements of $\boldsymbol{\rho}$ are zero, then the likelihood function above collapses to that of an MNP model. If not, the likelihood corresponds to a skew-normally mixed MNP model.

The *I*-dimensional integral in the likelihood contribution of each individual corresponds to the multivariate normal cumulative distribution function. The evaluation of such a function cannot be pursued using quadrature techniques due to the curse of dimensionality when the dimension of integration exceeds two (see Bhat, 2003). Consequently, the probability expression is typically approximated using Geweke-Hajivassiliou-Keane (GHK) simulator-based or the Genz-Bretz (GB) simulator-based techniques in the classical maximum simulated likelihood (MSL) inference approach (see Bhat *et al*., 2010 for a detailed description of these simulators) or

using Markov Chain Monte Carlo (MCMC) techniques in the Bayesian inference approach (see Albert and Chib, 1993, McCulloch and Rossi, 2000, and Train, 2009). However, these MSL and Bayesian techniques can require extensive simulation, can be time-consuming, are not always very straightforward to implement, and can create convergence assessment problems as the number of dimensions of integration increases. On the other hand, the maximum approximate composite marginal likelihood (MACML) approach for estimation of MNP models, in which the MVNCD function is evaluated using an *analytic approximation* method, is quite accurate and very fast.

There is, however, one very important issue that still needs to be dealt with. This concerns the positive definiteness of several matrices in Equation (12). Specifically, for the estimation to work, we need to ensure the positive definiteness of the following matrices: $\boldsymbol{\Omega}_+^*$, $\boldsymbol{\Sigma}$, and $\widetilde{\boldsymbol{\Psi}}$ (note that the positive definiteness of $\boldsymbol{\Omega}_+^*$ ensures the positive definiteness of $\boldsymbol{\Omega}^*$ and therefore $\boldsymbol{\Omega}$; this holds because of the property that any principal square sub-matrix of a positive definite matrix is also positive definite). Of these, one can guarantee the positive-definiteness of $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Psi}}$ in a straightforward fashion using a Cholesky decomposition approach (by parameterizing the likelihood function in terms of the Cholesky-decomposed parameters). To guarantee the positive definiteness of the correlation matrix $\boldsymbol{\Omega}_+^*$, we use the approach of Bhat and Srinivasan (2005). Specifically, let $\mathbf{L}$ be the Cholesky decomposition matrix for $\boldsymbol{\Omega}_+^*$. We need to guarantee that the parameters embedded within $\mathbf{L}$ are such that $\boldsymbol{\Omega}_+^*$ is a correlation matrix. This is done by parameterizing the diagonal terms of $\mathbf{L}$ as follows:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & \sqrt{1-l_{21}^2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{D+1,1} & l_{D+1,2} & l_{D+1,3} & \cdots & \sqrt{1-l_{D+1,1}^2-l_{D+1,2}^2\cdots-l_{D+1,D}^2} \end{bmatrix} \tag{19}$$

In the estimation, the Cholesky elements in the matrix $\mathbf{L}$ are estimated, guaranteeing that $\boldsymbol{\Omega}_+^*$ is indeed a correlation matrix.

3.2. Panel (or Repeated-Choice) MNP Formulation and Estimation

For the panel formulation, we introduce the index '*t*' for choice occasion. For ease in presentation, we will use the same number of choice occasions for each individual. Extension to the case of varying number of choice occasions per individual is straightforward.

Consider the random-coefficients formulation in which the utility that an individual q $(q = 1, 2, ..., Q)$ associates at time period t $(t = 1, 2, ..., T)$ with alternative i $(i = 1, 2, ..., I)$ is written as:

$$U_{qti} = \boldsymbol{\beta}_{\boldsymbol{q}}' \mathbf{x}_{\boldsymbol{qti}} + \boldsymbol{\gamma}_q' \mathbf{s}_{qti} + \tilde{\boldsymbol{\alpha}}_{qi} + \tilde{\boldsymbol{\varepsilon}}_{qti},$$
$$\boldsymbol{\beta}_{\boldsymbol{q}} \sim \text{MVSN}(\mathbf{b}, \boldsymbol{\omega}, \boldsymbol{\Omega}_+^*), \boldsymbol{\Omega}_+^* = \begin{pmatrix} 1 & \boldsymbol{\rho}' \\ \boldsymbol{\rho} & \boldsymbol{\Omega}^* \end{pmatrix}, \boldsymbol{\Omega}^* = (\boldsymbol{\omega})^{-1} \boldsymbol{\Omega} (\boldsymbol{\omega})^{-1}, \ \boldsymbol{\gamma}_q \sim \text{MVN}(\mathbf{c}, \boldsymbol{\Sigma}). \tag{20}$$

where all notations are as earlier except for the introduction of the index '*t*'. However, note that the vector $\mathbf{s}_{\boldsymbol{qti}}$ is now a $(K \times 1)$-column vector of exogenous attributes without including a constant. $\tilde{\alpha}_{\boldsymbol{qi}}$ is a normal random-effect term capturing time-stationary preference effects of individual q for alternative i. Also, as earlier, consider the $(I \times 1)$-vector $\tilde{\boldsymbol{\varepsilon}}_{qt} = (\tilde{\varepsilon}_{qt1}, \tilde{\varepsilon}_{qt2}, \tilde{\varepsilon}_{qt3}, \ldots, \tilde{\varepsilon}_{qtI})'$, and assume that $\tilde{\boldsymbol{\varepsilon}}_{qt} \sim \text{MVN}(0, \tilde{\boldsymbol{\Psi}})$, with the same normalizations on $\tilde{\boldsymbol{\Psi}}$ as in the cross-sectional case (note that the $\tilde{\boldsymbol{\varepsilon}}_{qt}$ error terms are considered independent across individuals and choice occasions, and $\tilde{\boldsymbol{\varepsilon}}_{qt}$, $\boldsymbol{\beta}_{\boldsymbol{q}}$, and $\boldsymbol{\gamma}_{\boldsymbol{q}}$ are also assumed independent for each individual q; $\boldsymbol{\beta}_{\boldsymbol{q}}$ and $\boldsymbol{\gamma}_{\boldsymbol{q}}$ are also independent across individuals). Next, stack the error terms $\tilde{\alpha}_{\boldsymbol{qi}}$ into an $(I \times 1)$-vector $\tilde{\boldsymbol{\alpha}}_q = (\tilde{\alpha}_{q1}, \tilde{\alpha}_{q2}, \tilde{\alpha}_{q3}, \ldots, \tilde{\alpha}_{qI})'$ and let $\tilde{\boldsymbol{\alpha}}_q \sim \text{MVN}_I(\tilde{\mathbf{a}}, \tilde{\boldsymbol{\Lambda}})$. However, since only utility differentials matter, take the differentials of these random effects with respect to the first alternative $\alpha_{qi1} = \tilde{\alpha}_{qi} - \tilde{\alpha}_{q1}$. Then, only the mean vector $\mathbf{a} = [(\tilde{a}_2 - \tilde{a}_1), (\tilde{a}_3 - \tilde{a}_1, ..., (\tilde{a}_I - \tilde{a}_1)]$ and covariance matrix $\tilde{\boldsymbol{\Lambda}}_1$ of $\boldsymbol{\alpha}_{q1} = (\alpha_{q21}, \alpha_{q31}, ..., \alpha_{qI1})$ are identified. At the same time, whenever utility differences are taken with respect to the chosen alternative during the MACML estimation, these utility differences should be consistent with the same mean vector $\tilde{\mathbf{a}}$ and error covariance matrix $\tilde{\boldsymbol{\Lambda}}$ for the undifferenced error term vector $\tilde{\boldsymbol{\alpha}}_q$. To achieve this, we set $\tilde{a}_1 = 0$ (that is, the first element of the vector $\tilde{\mathbf{a}}$ is set to zero), and construct $\tilde{\boldsymbol{\Lambda}}$ from $\tilde{\boldsymbol{\Lambda}}_1$ by adding a first row of zeros and a first column of zeros.

We now set out some additional notation. Write $\widetilde{\alpha}_{qi} = \widetilde{a}_i + \breve{\alpha}_{qi}$, $\mathbf{A} = (\widetilde{a}_1, \widetilde{a}_2, ..., \widetilde{a}_I)'$ ($I \times 1$ vector), $\breve{\boldsymbol{\alpha}}_q = (\breve{\alpha}_{q1}, \breve{\alpha}_{q2}, ..., \breve{\alpha}_{qI})'$ ($I \times 1$ vector) so that $\breve{\boldsymbol{\alpha}}_q \sim MVN_I(0, \widetilde{\boldsymbol{\Lambda}})$. Define $\mathbf{U}_{qt} = (U_{qt1}, U_{qt2}, ..., U_{qtI})'$ ($I \times 1$ vector), $\mathbf{U}_q = (\mathbf{U}'_{q1}, \mathbf{U}'_{q2}, ..., \mathbf{U}'_{qT})'$ ($TI \times 1$ vector), $\widetilde{\boldsymbol{\varepsilon}}_{qt} = (\widetilde{\varepsilon}_{1t1}, \widetilde{\varepsilon}_{qt2}, ..., \widetilde{\varepsilon}_{qtI})'$ ($I \times 1$ vector), $\widetilde{\boldsymbol{\varepsilon}}_q = (\widetilde{\boldsymbol{\varepsilon}}'_{q1}, \widetilde{\boldsymbol{\varepsilon}}'_{q2}, ..., \widetilde{\boldsymbol{\varepsilon}}'_{qT})'$ ($TI \times 1$ vector), $\breve{\mathbf{A}}_q = \mathbf{1}_T \otimes \breve{\boldsymbol{\alpha}}_q$ ($TI \times 1$ vector), $\mathbf{x}_{qt} = (\mathbf{x}_{qt1}, \mathbf{x}_{qt2}, ..., \mathbf{x}_{qtI})'$ ($I \times D$ matrix), $\mathbf{x}_q = (\mathbf{x}'_{q1}, \mathbf{x}'_{q2}, ..., \mathbf{x}'_{qT})'$ ($TI \times D$ matrix), $\mathbf{s}_{qt} = (\mathbf{s}_{qt1}, \mathbf{s}_{qt2}, ..., \mathbf{s}_{qtI})'$ ($I \times K$ matrix), $\mathbf{s}_q = (\mathbf{s}'_{q1}, \mathbf{s}'_{q2}, ..., \mathbf{s}'_{qT})'$ ($TI \times K$ matrix). Let $\mathbf{1}_T$ be a column vector of ones of dimension T, and let $\mathbf{1}_{TT}$ be a matrix of ones of dimension $T \times T$. Then, we can write:

$$\mathbf{U}_q = \left[\mathbf{x}_q \mathbf{b} + \mathbf{s}_q \mathbf{c} + (\mathbf{1}_T \otimes \mathbf{A})\right] + \left[\mathbf{x}_q \breve{\boldsymbol{\beta}}_q + \mathbf{s}_q \breve{\boldsymbol{\gamma}}_q + \breve{\mathbf{A}}_q + \widetilde{\boldsymbol{\varepsilon}}_q\right]. \tag{21}$$

Let $[.]_e$ indicate the e^{th} element of the column vector [.], and let $d_{ti} = (t-1)I + i$. Equation (20) can be equivalently written using Equation (21) as:

$$U_{qti} = \left[\mathbf{x}_q \mathbf{b} + \mathbf{s}_q \mathbf{c} + (\mathbf{1}_T \otimes \mathbf{A})\right]_{d_{ti}} + \left[\mathbf{x}_q \breve{\boldsymbol{\beta}}_q + \mathbf{s}_q \breve{\boldsymbol{\gamma}}_q + \breve{\mathbf{A}}_q + \widetilde{\boldsymbol{\varepsilon}}_q\right]_{d_{ti}}. \tag{22}$$

Define $V_{qti} = \left[\mathbf{x}_q \mathbf{b} + \mathbf{s}_q \mathbf{c} + (\mathbf{1}_T \otimes \mathbf{A})\right]_{d_{ti}}$ and $\boldsymbol{\varepsilon}_{qti} = \left[\mathbf{x}_q \breve{\boldsymbol{\beta}}_q + \mathbf{s}_q \breve{\boldsymbol{\gamma}}_q + \breve{\mathbf{A}}_q + \widetilde{\boldsymbol{\varepsilon}}_q\right]_{d_{ti}}$. Also, assume that individual q chooses alternative m_{qt} at the t^{th} choice instance. In the utility differential form, we may write Equation (22) as:

$$u^*_{qtim_{qt}} = U_{qti} - U_{qtm_{qt}} = H_{qtim_{qt}} + \xi_{qtim_{qt}};\ H_{qtim_{qt}} = V_{qti} - V_{qtm_{qt}} \text{ and } \xi_{qtim_{qt}} = \varepsilon_{qti} - \varepsilon_{qtm_{qt}};\ i \neq m_{qt} \tag{23}$$

Then stack the utility differentials $u^*_{qtim_{qt}} (= U_{qti} - U_{qtm_{qt}}, i \neq m_{qt})$ in the following order: $\mathbf{u}^*_{qt} = (u^*_{qt1m_{qt}}, u^*_{qt2m_{qt}}, ..., u^*_{qtIm_{qt}})'$, an $(I-1) \times 1$ vector, and $\mathbf{u}^*_{\boldsymbol{q}} = \left[\left(\mathbf{u}^*_{\boldsymbol{q1}}\right)', \left(\mathbf{u}^*_{\boldsymbol{q2}}\right)', ..., \left(\mathbf{u}^*_{\boldsymbol{qT}}\right)'\right]'$, an $\left[(I-1) \times T\right] \times 1$ vector. Correspondingly, let $\mathbf{H}_{\boldsymbol{qt}} = (H_{qt1m_{qt}}, H_{qt2m_{qt}}, ..., H_{qtIm_{qt}})'$, an $(I-1) \times 1$ vector; $\mathbf{H}_{\boldsymbol{q}} = (\mathbf{H}'_{\boldsymbol{q1}}, \mathbf{H}'_{\boldsymbol{q2}}, ..., \mathbf{H}'_{\boldsymbol{qT}})'$, an $\left[(I-1) \times T\right] \times 1$ vector. It is easy to see that $\mathbf{u}^*_{\boldsymbol{q}}$ has a mean vector $\mathbf{H}_{\boldsymbol{q}}$. To determine the covariance matrix of $\mathbf{u}^*_{\boldsymbol{q}}$, a few additional matrix definitions are needed. Define $\overline{\boldsymbol{\Omega}}_q = \mathbf{x}_q \boldsymbol{\Omega} \mathbf{x}'_q$ ($TI \times TI$ matrix), $\overline{\boldsymbol{\Sigma}}_q = \mathbf{s}_q \boldsymbol{\Sigma} \mathbf{s}'_q$ ($TI \times TI$ matrix), $\overline{\boldsymbol{\Lambda}} = \left(\mathbf{1}_{TT} \otimes \widetilde{\boldsymbol{\Lambda}}\right)$ ($TI \times TI$ matrix), and $\overline{\boldsymbol{\Psi}} = \mathbf{IDEN}_T \otimes \widetilde{\boldsymbol{\Psi}}$ ($TI \times TI$ matrix). Let

$\mathbf{F}_q = \left[\overline{\mathbf{\Omega}}_q + \overline{\mathbf{\Sigma}}_q + \overline{\mathbf{\Lambda}} + \overline{\mathbf{\Psi}}\right]$, and define $\boldsymbol{M_q}$ as an $[(I-1)\times T]\times[TI]$ block diagonal matrix, with each block diagonal having $(I-1)$ rows and I columns corresponding to the q^{th} individual's t^{th} choice instance. This $(I-1)\times I$ matrix for individual q and observation time period t corresponds to an $(I-1)$ identity matrix with an extra column of -1's added as the $m_{qt}{}^{\text{th}}$ column. For instance, consider the case of $T = 2$, and $I = 4$. Let the qth individual be observed to choose alternative 2 in time period 1 and alternative 1 in time period 2. Then $\mathbf{M}_q$ takes the form below.

$$\mathbf{M}_q = \left[\begin{array}{cccc|cccc} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{array}\right]. \tag{24}$$

Finally, we obtain the results below:

$$\mathbf{u}_q^* \sim \text{MVSN}(\mathbf{H}_q, \breve{\boldsymbol{\omega}}_q, \breve{\mathbf{\Omega}}_{q_+}^*)], \tag{25}$$

$$\breve{\mathbf{\Omega}}_{q_+}^* = \begin{pmatrix} 1 & \breve{\boldsymbol{\rho}}_q' \\ \breve{\boldsymbol{\rho}}_q & \breve{\mathbf{\Omega}}_q^* \end{pmatrix}, \breve{\mathbf{\Omega}}_q^* = (\breve{\boldsymbol{\omega}}_q)^{-1}\breve{\mathbf{\Omega}}_q(\breve{\boldsymbol{\omega}}_q)^{-1}, \breve{\mathbf{\Omega}}_q = \mathbf{M}_q\mathbf{F}_q\mathbf{M}_q', \ \breve{\boldsymbol{\rho}}_q = (\breve{\boldsymbol{\omega}}_q)^{-1}(\mathbf{M}_q\mathbf{x}_q)\boldsymbol{\omega}\boldsymbol{\rho}. \tag{26}$$

$\breve{\boldsymbol{\omega}}_q$ is the diagonal matrix of standard deviations of $\breve{\mathbf{\Omega}}_q$. The parameters to be estimated include the **A**, **b**, and **c** vectors, the elements of the covariance matrices $\mathbf{\Omega}$, $\mathbf{\Sigma}$, $\widetilde{\mathbf{\Lambda}}$, and $\widetilde{\mathbf{\Psi}}$, and the $\boldsymbol{\rho}$ parameter vector. Collect all these elements into a single vector $\boldsymbol{\theta}$. Then, one can use the result above to obtain the likelihood contribution of individual q choosing alternative m, which takes the $[T(I-1)+1]$-dimensional integral form below:

$$L_q(\boldsymbol{\theta}) = P(\mathbf{u}_q^* < 0) = \widetilde{\Phi}_{T(I-1)}\left((\breve{\boldsymbol{\omega}}_q)^{-1}(-\mathbf{H}_q); \breve{\mathbf{\Omega}}_{q_+}^*\right) = 2\Phi_{T(I-1)+1}\left(\mathbf{0}, (\breve{\boldsymbol{\omega}}_q)^{-1}(-\mathbf{H}_q), \breve{\mathbf{\Omega}}_{q-}^*\right). \tag{27}$$

In this panel setting, the parameter vector $\boldsymbol{\theta}$ is estimated by defining "events" in the MACML procedure as the pairs of choice observations across the choice occasions of the individual. Letting the individual's choice at time t be denoted by the index C_{qt}, the CML function for individual q is:

$$
\begin{aligned}
L_{CML,q}(\boldsymbol{\theta}) &= \prod_{t=1}^{T-1}\prod_{w=t+1}^{T} \text{Prob}\,(C_{qt}=m_{qt}, C_{qw}=m_{qw}) \\
&= \prod_{t=1}^{T-1}\prod_{w=t+1}^{T} \text{Prob}\,[\mathbf{u}^{*}_{qt}<0 \text{ and } \mathbf{u}^{*}_{qw}<0] \\
&= \prod_{t=1}^{T-1}\prod_{w=t+1}^{T} \text{Prob}\,[\ddot{\mathbf{u}}^{*}_{qtw}<0]
\end{aligned}
\tag{28}
$$

where $\ddot{\mathbf{u}}^{*}_{qtw}=\left[\left(\mathbf{u}^{*}_{qt}\right)',\left(\mathbf{u}^{*}_{qw}\right)'\right]'$. The computational effort is reduced in the CML above because only pairwise marginal multivariate probabilities are being considered across choice occasions. However, each multivariate orthant probability above still has a dimension equal to $[(I-1)\times 2]+1$:

$$
P(\ddot{\mathbf{u}}^{*}_{qtw}<0)=\widetilde{\Phi}_{2\times(I-1)}\left((\ddot{\boldsymbol{\omega}}_{qtw})^{-1}(-\ddot{\mathbf{H}}_{qtw});\ddot{\boldsymbol{\Omega}}^{*}_{qtw+}\right)=2\Phi_{2\times(I-1)+1}\left(0,(\ddot{\boldsymbol{\omega}}_{qtw})^{-1}(-\ddot{\mathbf{H}}_{qtw}),\ddot{\boldsymbol{\Omega}}^{*}_{qtw-}\right), \tag{29}
$$

where $\ddot{\mathbf{H}}_{qtw}=(\mathbf{H}'_{qt},\mathbf{H}'_{qw})'$, $\ddot{\boldsymbol{\Omega}}^{*}_{qtw+}$ and $\ddot{\boldsymbol{\Omega}}^{*}_{qtw-}$ are appropriate sub-matrices of $\breve{\boldsymbol{\Omega}}^{*}_{q+}$ and $\breve{\boldsymbol{\Omega}}^{*}_{q-}$, respectively (that is, they include elements corresponding to the t^{th} and w^{th} choice occasions of the individual). But such an orthant probability is conveniently computed using the approximation part of the MACML, leading to solely bivariate and univariate cumulative normals.

4. SIMULATION ANALYSIS

In this section, we undertake a simulation experiment with two objectives in mind. The first objective is to examine the ability of the MACML estimation method to recover parameters in the MNP model with skew-normally distributed coefficients. The second objective is to illustrate the problems that may arise from ignoring the skewness in the random coefficient distribution, which is equivalent to assuming that the distribution is normally distributed when it actually is not.

4.1. Experimental Set-Up

A cross-sectional formulation is used for the simulation experiments. Two cases are considered: (1) a three alternative case with three exogenous variables and (2) a five alternative case with five exogenous variables. In both the cases, the values of each of the exogenous variables for the

alternatives are drawn from a standard univariate normal distribution. In particular, a sample of 5000 realizations of the exogenous variables is generated corresponding to 5000 individuals. The first case specifies a skew-normally distributed random coefficient vector $\boldsymbol{\beta}_q$ on all the three exogenous variables, and the second case specifies a skew-normally distributed random coefficient vector $\boldsymbol{\beta}_q$ on the first three exogenous variables and a normally distributed random coefficient vector $\boldsymbol{\gamma}_q$ on the remaining two exogenous variables. For the five-dimensional simulation case, the coefficient vector $\boldsymbol{\gamma}_q$ is assumed to be a realization from $\boldsymbol{\gamma}_q \sim \text{MVN}(\mathbf{c}, \boldsymbol{\Sigma})$, with:

$$\mathbf{c} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{30}$$

In the simulation experiments, the coefficient vector $\boldsymbol{\beta}_q$ is assumed to be a realization from $\boldsymbol{\beta}_q \sim \text{MVSN}(\mathbf{b}, \boldsymbol{\omega}, \boldsymbol{\Omega}^*_+)$, $\boldsymbol{\Omega}^*_+ = \begin{pmatrix} 1 & \boldsymbol{\rho}' \\ \boldsymbol{\rho} & \boldsymbol{\Omega}^* \end{pmatrix}$, $\boldsymbol{\Omega}^* = (\boldsymbol{\omega})^{-1}\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1}$, where

$$\mathbf{b} = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}, \ \boldsymbol{\omega} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1.25 \end{pmatrix}, \text{ and } \boldsymbol{\Omega}^*_+ = \begin{pmatrix} 1 & -0.7 & -0.7 & -0.7 \\ -0.7 & 1 & 0.49 & 0.49 \\ -0.7 & 0.49 & 1 & 0.49 \\ -0.7 & 0.49 & 0.49 & 1 \end{pmatrix}. \tag{31}$$

The correlation matrix $\boldsymbol{\Omega}^*_+$ above is constructed in a specific manner so that the off-diagonal elements of the corresponding Cholesky matrix are all zero, except for the first column which now contains the skew parameters (= –0.7) as its elements.[5] Essentially, this way of constructing the correlation matrix assumes that all the correlations in the augmented four-dimensional correlation matrix (corresponding to the three-dimensional skew-normally distributed random coefficient vector) originates in the skew distribution of the coefficients, with no residual correlation beyond that generated by the skew. Such a specification is parsimonious, and can be

[5] The Cholesky matrix of $\boldsymbol{\Omega}^*_+$ is $\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -0.7 & 0.7141 & 0 & 0 \\ -0.7 & 0 & 0.7141 & 0 \\ -0.7 & 0 & 0 & 0.7141 \end{pmatrix}$

used to reduce the number of parameters to be estimated in the skew-normal MNP model. For instance, in the MNP with three skew-normal coefficients, there is a reduction from nine correlation parameters to just three. More generally, in a model with D skew-normal coefficients, there is a reduction from $\frac{D(D+1)}{2}+D$ to D parameters in the augmented correlation matrix. Clearly, this can be an effective way to allow a large number of skew-normally distributed coefficients without an explosion in the number of model parameters to be estimated. The other benefit of such a specification is that the skew parameter vector $\boldsymbol{\rho}$ is directly estimated because it "sits" as the first column of the Cholesky matrix (minus the first row element).

Another point to note about our skew specification for the $\boldsymbol{\beta}_q$ vector is that the negative values for $\mathbf{b}$ and $\boldsymbol{\rho}$ provide a negative location parameter and leftward skew for the marginal distributions of each of the $\boldsymbol{\beta}_q$ coefficients that is similar to a (negative) log-normal distribution. Such a specification may be considered for cost and other coefficients. Of course, in reality, the skew-normal distribution can be used for all parameters to allow a range of "seamless" and "continuous" marginal distribution possibilities that ranges from normality to non-normality.

The method to generate realizations from the MVSN distribution for $\boldsymbol{\beta}_q$ is based on first drawing a multivariate standard normal vector with correlation matrix $\boldsymbol{\Omega}_+^*$ in the usual way. This constitutes a draw for the latent underlying $(D+1)$-variate normally distributed vector $\widetilde{\mathbf{M}}=(\widetilde{M}_1,\widetilde{\mathbf{M}}_2')'$, where $\widetilde{M}_1$ is a latent (1×1)-vector and $\widetilde{\mathbf{M}}_2$ is a $(D\times 1)$-vector (see Equation (5); D = 3 in the current case). From this multivariate standard normal draw, a D-variate vector from the multivariate standard skew normal distribution is generated as follows:

$$\mathbf{Z}=\begin{cases}\widetilde{\mathbf{M}}_2 & \text{if } \widetilde{M}_1>0\\ -\widetilde{\mathbf{M}}_2 & \text{if } \widetilde{M}_1\le 0.\end{cases} \tag{32}$$

Finally, the error term vector $\boldsymbol{\varepsilon}_q=(\varepsilon_{q1},\varepsilon_{q2},\varepsilon_{q3},\ldots,\varepsilon_{qI})'$ is drawn from $\boldsymbol{\varepsilon}_q\sim \text{MVN}(\mathbf{0},0.5\times\mathbf{IDEN}_I)$, where $\mathbf{IDEN}_I$ is the identity matrix of dimension I (in the notation of Equation (19), $\boldsymbol{\Psi}=0.5\times \text{IDEN}_I$). Thus, we assume and maintain the IID normal assumption for $\boldsymbol{\varepsilon}_q$ in the current simulation experiment. The alternative with the highest utility for each individual q is then identified as the chosen alternative.

The above data generation process is undertaken 40 times with different realizations of the $\boldsymbol{\beta}_q, \boldsymbol{\gamma}_q$, and $\boldsymbol{\varepsilon}_q$ vectors $(q = 1, 2, ..., Q)$ to generate 40 different data sets. The MACML estimator is applied ten times to each dataset, with different sets of permutations (across the ten runs on the same dataset) to decompose the multivariate normal cumulative distribution or MVNCD function into a product of marginal and conditional probabilities (see Bhat, 2011). In each of the ten runs on the same dataset, ten different random permutations are generated and used for each individual (the random permutation varies across individuals) to approximate the MVNCD function for that individual. The approximation error for each parameter (due to using the analytic approximation to the MVNCD function) is obtained by computing the standard deviation of estimated parameters among the 10 different parameter estimates on the same data set.

A number of performance measures are identified to assess the performance of the MACML approach in being able to recover the underlying "true" parameters (which is the first objective of our simulation exercise). The performance measures, and the various steps to compute these measures, are described below:

(1) Estimate the MACML parameters for each data set *s* and for each of 10 independent sets of permutations for computing the MVNCD function.

(2) For each data set *s*, estimate the standard errors (s.e.) (using the sandwich covariance matrix estimator; see McFadden and Train, 2000).

(3) For each data set *s*, compute the mean estimate for each model parameter across the 10 random permutations used. Label this as MED, and then take the mean of the MED values across the data sets to obtain **a mean estimate**. Compute the **absolute percentage bias** (APB) as:

$$APB = \left| \frac{\text{mean estimate - true value}}{\text{true value}} \right| \times 100 .$$

(4) For each data set *s*, compute the median s.e. for each model parameter across the 10 draws. Call this MSED, and then take the mean of the MSED values across the 40 data sets and label this as **the asymptotic standard error.**

(5) Next, compute the standard deviation of the MED values across the 40 data sets to obtain the finite sample standard error for each parameter, and label this as the **empirical standard error.** Note that the asymptotic standard error is essentially an approximation to

this empirical standard error, and the consistency of the estimator for the asymptotic standard error implies that the asymptotic and empirical standard error estimates should be close to one another.

(6) Next, for each data set *s*, compute the approximation standard deviation for each parameter as the standard deviation in the estimated parameter values across the 10 independent permutations (about the MED value). Call this standard deviation as APPMED. For each parameter, take the mean of APPMED across the different data sets. Label this as the **approximation standard error** for each parameter.

(7) For each parameter, compute an **approximation adjusted asymptotic standard error** as follows: $\sqrt{(\text{asymptotic standard error})^2 + (\text{approximation standard error})^2}$. Similarly, compute an **approximation adjusted empirical standard error** as follows: $\sqrt{(\text{empirical standard error})^2 + (\text{approximation standard error})^2}$.

The second objective of examining the implications of ignoring skewness when actually present is achieved by generating data exactly as discussed above. Once generated, we estimate a simple normally-mixed MNP model on the data, assuming (incorrectly) that $\boldsymbol{\rho} = 0$ (using ten random permutation per individual in the computation of the MVNCD function, exactly as earlier). We will refer to this model as the MNP-normal (or MNP-N) model. We compare this MNP-N model with the skew normally-mixed (or MNP-SN) model. For this comparison, we ignore approximation error issues and undertake a single MNP-N estimation on each of the 40 datasets generated. We then randomly pick one of the MNP-SN model estimates for each of the 40 datasets (as already estimated earlier), and use that to compare with the MNP-N model. The performances of the two models are evaluated by (1) comparing the mean APB values across parameters and (2) undertaking a likelihood ratio test (LRT) for each of the 40 datasets. For the mean APB computation, the APB in the skewness parameters is not included in the MNP-SN model because the MNP-N implicitly assumes that $\boldsymbol{\rho} = 0$ (this allows an "apples to apples" comparison between the MNP-N and MNP-SN models). For the likelihood ratio test, we compare the test statistic for each data set with the table chi-squared distribution value with three degrees of freedom (corresponding to each of the three skew parameters in the $\boldsymbol{\rho}$ vector being zero). The number of times out of the 40 data sets that the MNP-SN model rejects the MNP-N model is then obtained, along with the mean value of the LRT statistic across the 40 data sets.

4.2. Simulation Results

4.2.1 Ability of MACML to Recover Model Parameters

The results for the first objective of evaluating the ability to recover model parameters are summarized in the Table 1 for the three alternative case with three exogenous variables, and in Table 2 for the five alternative case with five exogenous variables.

4.2.1.1 The Three Alternative Case with Three Exogenous Variables

The results in Table 1 for the three alternative case indicate that the MACML method does reasonably well in recovering the true parameters. The absolute percentage bias (APB) ranges from 7.1% to 11.2% across the parameters, with a mean value of 9.2% (see the last row of the table under the "absolute percentage bias" column). The APB values are generally somewhat smaller and more stable (across parameters) for the location parameters of the distributions of the $\boldsymbol{\beta}_q$ parameter vector (*i.e.*, the b parameter estimates in the table) than for the skew parameter estimates (*i.e.*, the ρ values) or the scale parameter estimates of the distribution of the $\boldsymbol{\beta}_q$ parameter vector (*i.e.*, the ω parameters in the table). This is not surprising, because the b parameters enter more linearly in the likelihood function of Equation (18) (through the mean of the MVNCD function) than do the skew and scale parameters (that enter more non-linearly and in a complex manner through the covariance matrix of the MVNCD function). One can also observe that all the parameters associated with the third variable are recovered better than the first two variables, perhaps because of the higher standard deviation of this coefficient (=1.25) relative to the other two coefficients. When there is higher variation in a coefficient, it provides more information in the data to pin down the moments of its distribution.

The asymptotic and empirical standard error values (reflecting sampling standard error) are quite close to one another, reflecting the consistency of the MACML estimator of the asymptotic covariance matrix. These sampling standard error estimates of the parameters indicate good efficiency of the MACML estimator, with the standard errors being between 8%-15% of the mean values of the estimator. Also, the approximation standard error estimates are smaller than the sampling standard errors. On average, the approximation standard error is about 60% of the corresponding asymptotic and empirical standard error values. On the other hand, in a similar simulation setting, the approximation standard error of the MACML estimator with just one permutation per individual (as opposed to ten used here) was found to be only of the order of

13% of the sampling standard errors when the MACML approach was applied to a strictly normally-distributed coefficients model (see Bhat and Sidharthan, 2011). Clearly, even though the skew-normally distributed coefficients can be viewed as originating from an augmented and truncated multivariate normal distribution, and the cumulative distribution function of the skew-normal distribution may be written as that of a normal distribution function with an added dimension, the introduction of asymmetry does appear to introduce more approximation error in the MACML approach. This is an issue that needs further examination in the future. Nonetheless, this should not detract from the fact that the MACML estimator still does very well. In fact, the final column provides the approximation-adjusted asymptotic and empirical standard errors for the MACML estimator, which are only 13-25% higher than the corresponding unadjusted standard errors. Also, the approximation-adjusted standard errors are still only 10-17% of the corresponding mean values of the estimators, indicating that the approximation standard errors introduced by the MACML approach are small in the larger inference context.

4.2.1.2. Five Alternative Case with Five Variables

The results for the five alternative case with five variables are summarized in Table 2. The APB is of the same order as that in the case with three skew-normal coefficients, and ranges from 3% to 18.5% with a mean of 9.4%. As in the previous section, the APB values are smaller and more stable for the *b* parameter estimates than for the ρ and ω parameter estimates. Further, there is a clear increase in the APB values for the ρ and ω parameter estimates compared to the case with three coefficients. However, the APB for the parameters characterizing the normally distributed coefficients (see the *c* and the σ parameters in the fourth and fifth row panels of Table 2, respectively) are estimated very well, with the APBs ranging from 3-6.5% (mean of the APBs for these parameters is 4.5%, which is less than half of the overall mean APB of 9.4%).

The sampling (asymptotic and empirical) standard error values of the parameters continue to indicate good efficiency of the MACML estimator, with the sampling standard errors ranging between 5%-14% of the mean values of the estimator. Also, the approximation standard error estimates continue to be smaller than the sampling standard error estimates. On average, the approximation standard errors are about 45% of the corresponding asymptotic standard error estimates and 40% of the corresponding empirical standard error values, which is even better than the three-dimensional case. While the approximation errors are close to the sampling

standard errors for the skewness elements ρ, this is because the standard errors are extremely small for these elements in the first place. At the end, the approximation-adjusted asymptotic and empirical standard errors are only 5-16% of the mean values of the estimator, which is about the same range as the unadjusted standard errors as a percentage of the mean values.

To summarize, the MACML inference approach does very well in recovering the parameters in a skew-normally mixed MNP model (with or without normally mixed coefficients). However, there is also evidence that there is some kind of a relative degradation of performance when skew-normally distributed coefficients are introduced (relative to the case when there are only normally-distributed coefficients, in which case the MACML approach does extremely well). Some of this degradation is surely attributable to the more difficult asymmetric shapes that need to be characterized with skew-normal distributions. More explorations are needed to examine such behavior. However, despite the relative degradation, the MACML model is able to recover all parameters well, with the approximation errors being quite inconsequential in the larger sampling inference context.

4.2.2 Effects of Ignoring Skewness in the Coefficient Distribution

This section focuses on the implications of ignoring skewness when actually present. The results are presented in Table 3 for both the three dimensional case (three alternatives-three variable case) and the five dimensional case (five alternatives-five variable case). The results clearly show the poor performance of the MNP-N model (which assumes away any skewness) relative to the MNP-SN model (which explicitly accommodates skewness). The mean APB value across the location parameters is of the order of 60% in the MNP-N model compared to the corresponding mean APB value of 6-8% from the MNP-SN model. The scale parameters also have a larger mean APB in the MNP-N model compared to the MNP-SN model. Overall, the use of a normal distribution when there is skew in the random parameters can lead to seriously mis-estimated distributions for the random parameters. This, in turn, will then lead to mis-estimated willingness to pay and welfare measures. An interesting observation from the five-dimensional analysis, though, is that if there are truly normally distributed coefficients in the model, these do not appear to be substantially affected by mis-specifications on the other coefficients (as can be noticed from the similar mean APB values for the mean elements of the γ_q vector and the covariance elements of the γ_q vector).

The log-likelihood values at convergence from the MNP-SN model is always better than from the MNP-N model in all the 40 generated data sets. The mean value of the log-likelihood ratio statistic across all the 40 data sets for each of the three-dimensional and five-dimensional cases is provided in Table 3. Also, for each and every data set, the log-likelihood ratio statistic is higher than the corresponding chi-squared table value (see the last row of Table 3).

Overall, the results clearly highlight the bias in characterizing the distribution of random coefficients if skewness effects in the coefficients are ignored when actually present.

5. CONCLUSION

In the current paper, we propose the use of the multivariate skew-normal distribution function to accommodate non-normal mixing in MNP models. The multivariate skew normal (MSN) distribution retains several attractive properties of the multivariate normal distribution. It is tractable, parsimonious in parameters that regulate the distribution and its skewness, and includes the normal distribution as a special interior point case. It also is a very flexible unimodal density structure that allows a "seamless" and "continuous" variation from normality to non-normality, and can replicate a variety of smooth unimodal density shapes. At the same time, we propose the use of an MNP kernel because the combination of skew-normal mixing over the MNP kernel lends itself perfectly to estimation using the maximum approximate composite marginal likelihood (MACML) approach. This is because of two properties of the skew distribution. The first is that it is closed under any affine transformation of the skew-normally distributed vector, and the second is that it is closed under the sum of a skew-normally distributed vector and a normally distributed vector of the same dimensions. As importantly, the cumulative distribution function of the D-variate skew normally distributed variable requires only the evaluation of a $(D+1)$-dimensional multivariate cumulative normal distribution function. All of these properties are gainfully exploited in the paper to formulate an MNP model with non-normal mixing, while also being able to estimate the model in a simple and computationally efficient MACML approach. To our knowledge, this is the first paper to propose and formulate a skew-normally mixed MNP model.

A simulation exercise is undertaken to evaluate the ability of the proposed approach to recover parameters in the skew-normally mixed MNP model. Two cases are considered: (1) a three alternative case with three exogenous variables and (2) a five alternative case with five

exogenous variables. The first case considers a three-variate skew normal distribution for the coefficients on the three exogenous variables, while the second case considers a three-variate skew normal distribution for three variables and a bivariate normal for two variables. The results show that our proposed approach does very well in recovering the parameters in a skew-normally mixed MNP model. In addition, the simulation results clearly highlight the bias in characterizing the distribution of random coefficients as well as the poor data fit if skewness, when actually present, is ignored away. Ongoing efforts are focused on additional simulation experiments to examine the effectiveness of the approach in settings with spatial dependencies and social dependencies across decision units, and combinations of temporal, spatial, and social dependencies.

ACKNOWLEDGEMENTS

The authors would like to acknowledge support from the Sustainable Cities Doctoral Research Initiative at the Center for Sustainable Development at The University of Texas at Austin. The authors are grateful to Lisa Macias for her help in formatting this document, and to an anonymous referee and Fred Mannering for helpful comments on an earlier version of this document.

REFERENCES

Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function model. *Journal of Econometrics* 6(1), 21-37.

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88(422), 669-679.

Amador, F.J., Gonzales, R., Ortuzar, J., 2005. Preference heterogeneity and willingness to pay for travel time savings. *Transportation* 32(6), 627-647.

Anselin, L., 1988. *Spatial econometrics: Methods and models*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Arellano-Valle, R.B., Azzalini, A., 2006. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* 33(3), 561-574.

Arellano-Valle, R.B., Azzalini, A. 2008. The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis* 99(7), 1362-1382.

Arellano-Valle, R.B., Genton, M.G.. 2005. On fundamental skew distributions. *Journal of Multivariate Analysis* 96(1), 93-116.

Azzalini, A., 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12(2), 171-178.

Azzalini, A., 2011. Selection models under generalized symmetry settings. *Annals of the Institute of Statistical Mathematics*, forthcoming.

Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B* 61(3) 579-602.

Azzalini, A., Dalla Valle, A., 1996. The multivariate skew-normal distribution. *Biometrika* 83(4), 715-726.

Balcombe, K., Chalak, A., Fraser, I.M., 2009. Model selection for the mixed logit with bayesian estimation. *Journal of Environmental Economics and Management* 57(2), 226–237.

Bartels, R., Fiebig, D.G., van Soest, A., 2006. Consumers and experts: an econometric analysis of the demand for water heaters. *Empirical Economics* 31(2), 369-391.

Bastin, F., Cirillo, C., Toint, P.L., 2010. Estimating non-parametric random utility models, with an application to the value of time in heterogeneous populations. *Transportation Science* 44(4) 537-549.

Bhat, C.R., 1995. A heteroscedastic extreme-value model of intercity mode choice. *Transportation Research Part B* 29(6), 471-483.

Bhat, C.R., 1997a. Work travel mode choice and number of nonwork commute stops. *Transportation Research Part B* 31(1), 41-54.

Bhat, C.R., 1997b. An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science* 31(1), 34-48.

Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B* 37(9), 837-855.

Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B* 45(7), 923-939.

Bhat, C.R., Guo, J.Y., 2007. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B* 41(5), 506-526

Bhat, C.R., Sidharthan, R., 2011. A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. *Transportation Research Part B* 45(7), 940-953.

Bhat, C.R., Srinivasan, S., 2005. A multidimensional mixed ordered-response model for analyzing weekend activity participation. *Transportation Research Part B* 39(3), 255-278.

Bhat, C.R., Eluru N., Copperman, R.B., 2008. Flexible model structures for discrete choice analysis. In *Handbook of Transport Modelling, 2nd edition*, Hensher, D.A., Button, K.J. (eds.), Ch. 5, pp. 75-104, Elsevier Science.

Bhat, C.R., Varin, C., Ferdous, N., 2010. A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered response model. In *Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications*, Greene W., Hill R.C. (eds.), Vol. 26, pp. 65-106, Emerald Group Publishing Limited.

Birnbaum, Z.W., 1950. Effect of linear truncation on a multinormal population. *Annals of Mathematical Statistics* 21(2), 272-279.

Birol, E., Karousakis, K., Koundouri, P., 2006. Using economic valuation techniques to inform water resources management: A survey and critical appraisal of available techniques and an application. *Science of the Total Environment* 365(1-3), 105-122.

Campbell, D., Doherty, E., Hynes, S., Van Rensburg, T., 2010. Combining discrete and continuous mixing approaches to accommodate heterogeneity in price sensitivities in environmental choice analysis. *84th Agricultural Economics Society Annual Conference*, March 29-31, Edinburgh, Scotland.

Cedilnik. A., Kosmelj, K., Blejec, A., 2006. Ratio of two random variables: a note on the existence of its moments. *Metodološki Zvezki - Advances in Methodology and Statistics* 3(1), 1-7.

Cherchi, E., Cirillo, C., Polak, J., 2009. User benefit assessment in presence of random taste heterogeneity: comparison between parametric and nonparametric models. *Transportation Research Record* 2132, 78-86

Chintagunta, P.K., Jain, D.C., Vilcassim, N.J., 1991. Investigating heterogeneity in brand preferences in logit models for panel data. *Journal of Marketing Research* 28(4), 417-428.

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

Daly, A., Hess, S., Train, K., 2011. Assuring finite moments for willingness to pay in random coefficient models. *Transportation* 39(1), 19-31.

Ellison, B.E., 1964. Two theorems for inferences about the normal distribution with applications in acceptance sampling. *Journal of the American Statistical Association* 59(305), 89-95.

Fosgerau, M., 2005. Unit income elasticity of the value of travel time savings. Presented at 8th NECTAR Conference, Las Palmas G.C., June 2-4.

Fosgerau, M., 2006. Investigating the distribution of the value of travel time savings. *Transportation Research Part B* 40(8), 688-707.

González-Farías, G., Domínguez-Molina, A., Gupta, A.K., 2004. Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference* 126(2), 521-534.

Greene, W.H., Hensher, D.A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B* 37(8), 681-698.

Greene W.H, Hensher, D.A., Rose, J.M., 2006. Accounting for heterogeneity in the variance of the unobserved effects in mixed logit models (NW transport study data). *Transportation Research Part B* 40(1), 75-92

Gupta, A.K., González-Farías, G., Domínguez-Molina, A., 2004. A multivariate skew normal distribution. *Journal of Multivariate Analysis* 89(1), 181-190.

Hausman, J.A., Wise, D.A., 1978. A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46(2), 403-426.

Hess, S., Bierlaire, M., Polak, J.W., 2007. A systematic comparison of continuous and discrete mixture models. *European Transport* 37, 35-61

Hensher, D.A., Rose, J.M., Greene, W.H., 2005. *Applied Choice Analysis: A Primer*. Cambridge University Press, Cambridge, U.K.

Hynes, S., Hanley, N., Scarpa, R., 2008. Effects on welfare measures of alternative means of accounting for preference heterogeneity in recreational demand models. *American Journal of Agricultural Economics* 90(4), 1011-1027.

Jara, A., Quintana, F., San Martín, E., 2008. Linear mixed models with skew-elliptical distributions: a Bayesian approach. *Computational Statistics and Data Analysis* 52(11), 5033-5045.

Kamakura, W.A., Russell, G.J., 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* 26(4) 379-390.

Lancaster, K., 1971. *Consumer Demand: A New Approach*. Columbia University Press, New York

Li, Z., Hensher, D.A., Rose, J.M., 2010. Willingness to pay for reliability in passenger transport: a review and some new empirical evidence. *Transportation Research Part E* 46(3), 384-403.

Luce, R., Suppes, P., 1965. Preference, utility, and subjective probability. In *Handbook of Mathematical Psychology, Volume III*, Luce, R., Bush, R., Galanter E. (eds.), John Wiley & Sons, New York.

McCulloch, R.E., Rossi, P.E., 2000. Bayesian analysis of the multinomial probit model. In *Simulation-Based Inference in Econometrics*, Mariano, R., Schuermann, T., Weeks, M.J., (eds.), 158-178, Cambridge University Press, New York.

McFadden, D., 1974. The measurement of urban travel demand. *Journal of Public Economics* 3(4), 303-328.

McFadden, D., 1978. Modeling the choice of residential location. *Transportation Research Record* 672, 72-77.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15(5), 447-470.

Meintanis, S.G., Hlávka, Z., 2010. Goodness-of-fit tests for bivariate and multivariate skew-normal distributions. *Scandinavian Journal of Statistics* 37(4), 701-714.

Molenaar, D., Dolan, C.V., Verhelst, N.D., 2010. Testing and modeling non-normality with the one-factor model. *British Journal of Mathematical and Statistical Psychology* 63(2), 293-317.

O'Hagan, A., Leonard, T., 1976. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* 63(1), 201-203.

Small, K.A., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica* 73(4), 1367-1382.

Torres, C., Hanley, N., Riera, A., 2011. How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments. *Journal of Environmental Economics and Management* 62(1), 111-121.

Train, K., 2003. *Discrete Choice Methods with Simulation*, 1st ed. Cambridge University Press, Cambridge.

Train, K.E., 2008. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling* 1(1), 40-69

Train, K., 2009. *Discrete Choice Methods with Simulation*, 2nd ed. Cambridge University Press, Cambridge.

Train, K., Sonnier, G., 2005. Mixed logit with bounded distributions of correlated partworths. In *Applications of Simulation Methods in Environmental and Resource Economics*, Scarpa, R., Alberini, A., (eds.), Ch. 7, pp. 117-134, Springer, Dordrecht, The Netherlands.

Weinstein, M.A., 1964. The sum of values from a normal and a truncated normal distribution. *Technometrics* 6(1), 104-105.

Yai, T., Iwakura, S., Morichi, S., 1997. Multinomial probit with structured covariance for route choice behavior. *Transportation Research Part B* 31(3), 195-207.

Appendix A.1

The moments of the SSN distribution are most easily obtained from the moment generating function of Z, which is given by:

$$M(t) = E[\exp(tz)] = \int_{z=-\infty}^{\infty} \exp(tz)\,\phi(z)\,\Phi(\alpha z)\,dz$$

$$= 2\int_{z=-\infty}^{\infty} \exp(tz)\,\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-z^2}{2}\right)\Phi(\alpha z)\,dz$$

$$= 2\int_{z=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left\{z^2 - 2tz + t^2\right\} + \frac{t^2}{2}\right)\Phi(\alpha z)\,dz$$

$$= 2\int_{z=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(\frac{t^2}{2}\right)\exp\left(-\frac{1}{2}(z-t)^2\right)\Phi(\alpha z)\,dz$$

$$= 2\exp\left(\frac{t^2}{2}\right)\int_{z=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(z-t)^2\right)\Phi(\alpha z)\,dz$$

$$= 2\exp\left(\frac{t^2}{2}\right)\int_{u=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)\Phi(\alpha(u+t))\,du, \text{ where } u = z\text{-}t$$

$$= 2\exp\left(\frac{t^2}{2}\right)\int_{u=-\infty}^{\infty} \phi(\mathrm{u})\,\Phi(\alpha(u+t))\,du$$

$$= 2\exp\left(\frac{t^2}{2}\right)\mathrm{E}\left(\Phi(\alpha u + \alpha t)\right)$$

$$= 2\exp\left(\frac{t^2}{2}\right)\Phi\left(\frac{\alpha t}{\sqrt{1+\alpha^2}}\right) \text{ using } \mathrm{E}\left\{\Phi(\alpha u + \alpha t)\right\} = \Phi\left(\frac{\alpha t}{\sqrt{1+\alpha^2}}\right) \text{ (from Ellison, 1964)}$$

In the above expression, $\rho = \dfrac{\alpha}{\sqrt{1+\alpha^2}}$. From above, the first three moments of the distribution may be written as follows with $b = \sqrt{2/\pi}$:

$$E(Z) = \mu_Z = \left.\frac{dM(t)}{dt}\right|_{t=0} = b\rho; \quad \mathrm{Var}(Z) = \sigma_Z^2 = \left.\frac{d^2M(t)}{dt^2}\right|_{t=0} = 1 - b^2\rho^2, \text{ and}$$

$$\mathrm{Skew}(Z) = \gamma_Z = \left.\frac{d^3M(t)}{dt^3}\right|_{t=0} = \left(\frac{4-\pi}{2}\right)^2\left(\frac{\mu_Z^2}{\sigma_Z^2}\right)^{3/2},$$

where γ_Z is the Pearson index of skewness that is a measure of asymmetry. When $\alpha = 0$, $\gamma_Z = 0$ as should be the case for the normal distribution. The moments for the variable $Y = \xi + \omega Z$, which is non-standard skew-normally distributed, may be obtained as $\mu_Y = \xi + \omega\mu_Z$, $\sigma_Y^2 = \omega^2(1 - \sigma_Z^2)$, and $\gamma_Y = \gamma_Z$.

Appendix A.2

From Equation (5),

$$\begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Then,

$$f(M_1, M_2 \mid M_1 > 0) = \frac{\phi_2(M_1, M_2)}{\text{Prob}[M_1 > 0]} = 2\,\phi_2(M_1, M_2),$$

where ϕ_2 is the standard bivariate normal density function.

$$\begin{aligned} f(M_2 \mid M_1 > 0) &= \int_{M_1=0}^{\infty} 2\,\phi_2(M_1, M_2)\, dM_1 \\ &= 2 \int_{M_1=0}^{\infty} \phi(M_2) \frac{1}{\sqrt{1-\rho^2}} \phi\left[\frac{M_1 - \rho M_2}{\sqrt{1-\rho^2}}\right] dM_1 \\ &= 2\,\phi(M_2)\left[1 - \Phi(-\alpha M_2)\right], \quad \text{where } \alpha = \frac{\rho}{\sqrt{1-\rho^2}}, \end{aligned}$$

or $f(z) = \tilde{\phi}(z;\alpha) = 2\,\phi(z)\,\Phi(\alpha z)$

Appendix A.3

The moment generating function of $\mathbf{Z}$ is:

$$M_Z(\mathbf{t}) = 2\exp\left(\frac{1}{2}\mathbf{t}'\boldsymbol{\Omega}^*\mathbf{t}\right)\Phi(\boldsymbol{\rho}'\mathbf{t}).$$

The first three moments of the distribution may subsequently be obtained from the function above in a straightforward fashion with $b = \sqrt{2/\pi}$:

$E(\mathbf{Z}) = \boldsymbol{\mu}_Z = b\boldsymbol{\rho}$; $\text{Var}(\mathbf{Z}) = \boldsymbol{\Omega}^* - \boldsymbol{\mu}_\mathbf{Z}\boldsymbol{\mu}_\mathbf{Z}'$, and

$$\text{Skew}(\mathbf{Z}) = \boldsymbol{\gamma}_Z = \left(\frac{4-\pi}{2}\right)^2 \left(\frac{\boldsymbol{\mu}_\mathbf{Z}'(\boldsymbol{\Omega}^*)^{-1}\boldsymbol{\mu}_\mathbf{Z}}{\mathbf{1} - \boldsymbol{\mu}_\mathbf{Z}'(\boldsymbol{\Omega}^*)^{-1}\boldsymbol{\mu}_\mathbf{Z}}\right)^3,$$

The moments for the variable $\mathbf{Y} = \boldsymbol{\xi} + \boldsymbol{\omega}\mathbf{Z}$, which is non-standard skew-normally distributed, may be obtained as $\boldsymbol{\mu}_\mathbf{Y} = \boldsymbol{\xi} + \boldsymbol{\omega}\boldsymbol{\mu}_\mathbf{Z}$, $\text{Var}(\mathbf{Y}) = \boldsymbol{\omega}\mathbf{Var(Z)}\boldsymbol{\omega}$, and $\boldsymbol{\gamma}_Y = \boldsymbol{\gamma}_Z$. For future reference, we will also write the moment generating function of $\mathbf{Y}$ (obtained from Equation (11)) as follows:

$$M_Y(\mathbf{t}) = E[\exp(\mathbf{t}'\mathbf{Y})] = E(\exp[\mathbf{t}'(\boldsymbol{\xi} + \boldsymbol{\omega}\mathbf{Z})]) = \exp(\mathbf{t}'\boldsymbol{\xi})E(\mathbf{t}'\boldsymbol{\omega}\mathbf{Z}) = 2\exp\left(\boldsymbol{\xi}'\mathbf{t} + \frac{\mathbf{1}}{\mathbf{2}}\mathbf{t}'\boldsymbol{\Omega}\mathbf{t}\right)\Phi(\boldsymbol{\rho}'\boldsymbol{\omega}\,\mathbf{t}).$$

LIST OF FIGURES

LIST OF TABLES

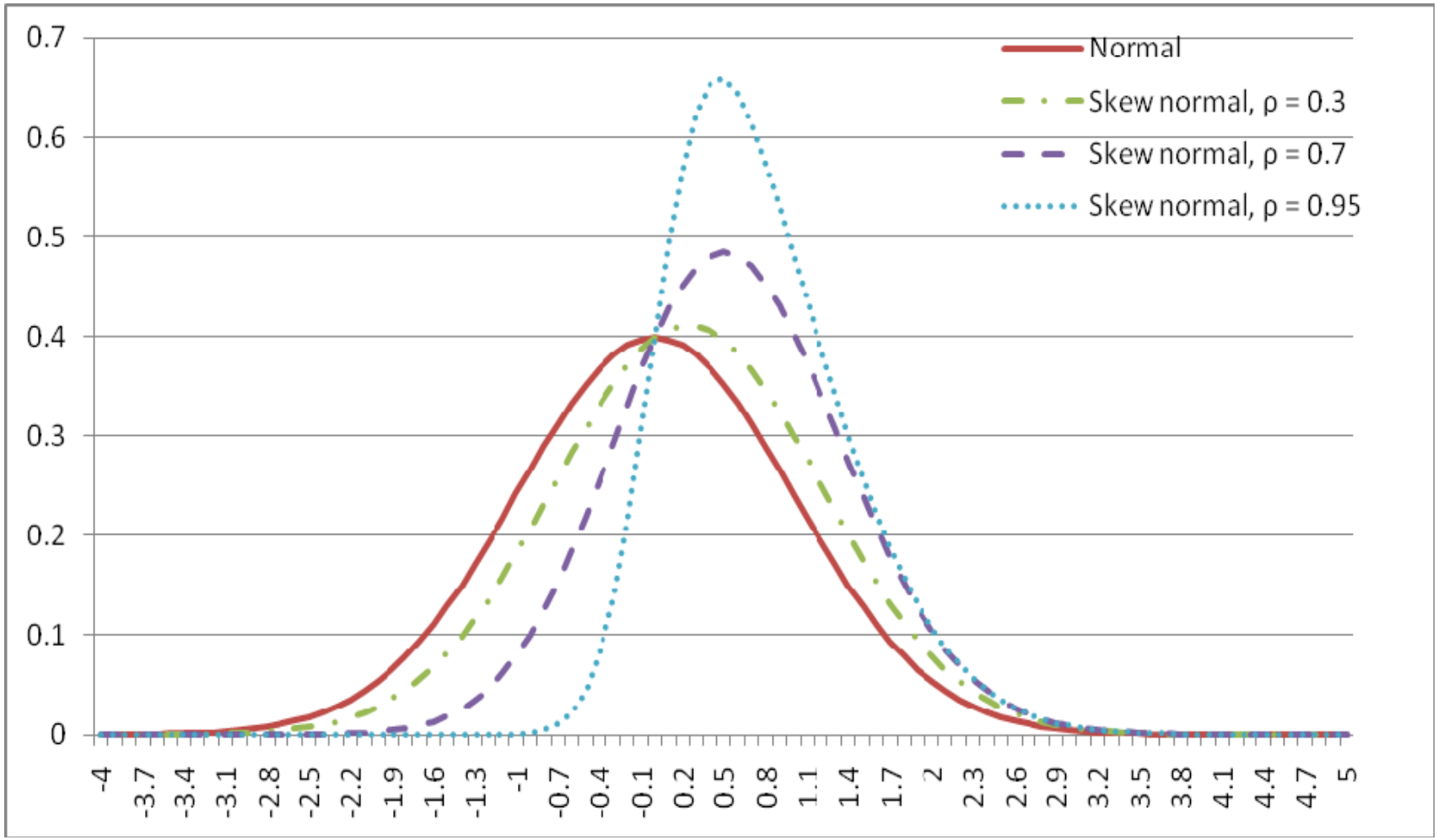

Figure 1. Shape of the SSN density function for a number of positive values of ρ

Table 1. Simulation Results for the Three Alternative-Three Variable Case

Parameter	True Value	Parameter Estimates		Standard Error (SE) Estimates				
		Mean Estimate	Absolute Percentage Bias	Asymptotic SE	Empirical SE	Approximation SE	Approximation Adjusted Asymptotic SE	Approximation Adjusted Empirical SE
Location parameters of the β_q vector								
$b1$	-1.000	-0.906	9.4%	0.116	0.134	0.073	0.137	0.153
$b2$	-1.000	-0.917	8.3%	0.114	0.125	0.072	0.135	0.144
$b3$	-1.000	-0.932	6.8%	0.122	0.127	0.076	0.144	0.149
Skewness parameters of the β_q vector								
$\rho 1$	-0.700	-0.770	10.1%	0.065	0.081	0.048	0.081	0.094
$\rho 2$	-0.700	-0.778	11.2%	0.062	0.064	0.047	0.078	0.079
$\rho 3$	-0.700	-0.750	7.1%	0.061	0.070	0.044	0.076	0.083
Scale parameters of the β_q vector								
$\omega 1$	1.000	1.112	11.2%	0.135	0.144	0.073	0.154	0.162
$\omega 2$	1.000	1.111	11.1%	0.134	0.122	0.068	0.150	0.140
$\omega 3$	1.250	1.344	7.5%	0.150	0.135	0.080	0.170	0.157
Overall Mean Value Across Parameters			9.2%	0.107	0.111	0.065	0.125	0.129

Table 2. Simulation Results for the Five Alternative-Five Variable Case

Parameter	True Value	Parameter Estimates		Standard Error (SE) Estimates				
		Mean Estimate	Absolute Percentage Bias	Asymptotic SE	Empirical SE	Approximation SE	Approximation Adjusted Asymptotic SE	Approximation Adjusted Empirical SE
Location parameters of the β_q vector								
$b1$	-1.000	-0.914	8.6%	0.107	0.120	0.053	0.119	0.132
$b2$	-1.000	-0.917	8.3%	0.106	0.137	0.053	0.119	0.147
$b3$	-1.000	-0.990	1.0%	0.116	0.135	0.058	0.130	0.147
Skewness parameters of the β_q vector								
$\rho 1$	-0.700	-0.825	17.9%	0.036	0.042	0.030	0.047	0.051
$\rho 2$	-0.700	-0.824	17.7%	0.036	0.044	0.028	0.046	0.052
$\rho 3$	-0.700	-0.769	9.9%	0.034	0.036	0.030	0.046	0.046
Scale parameters of the β_q vector								
$\omega 1$	1.000	1.184	18.4%	0.144	0.167	0.067	0.159	0.180
$\omega 2$	1.000	1.168	16.8%	0.143	0.152	0.066	0.157	0.166
$\omega 3$	1.250	1.381	10.5%	0.158	0.162	0.067	0.172	0.175
Mean values of the γ_q vector								
$c1$	1.000	1.041	4.1%	0.107	0.107	0.038	0.114	0.114
$c2$	1.000	1.039	3.9%	0.107	0.112	0.038	0.113	0.118
Covariance elements of the γ_q vector								
$\sigma 1$	1.000	1.065	6.5%	0.126	0.144	0.044	0.134	0.151
$\Sigma 12$	0.500	0.516	3.2%	0.067	0.059	0.022	0.071	0.063
$\sigma 2$	1.000	1.051	5.1%	0.124	0.142	0.045	0.132	0.149
Overall Mean Value Across Parameters			9.4%	0.101	0.111	0.046	0.111	0.121

Table 3. Effects of Ignoring Skewness in the Mixing Distribution (when present)

Evaluation Metric	Three Dimensional Case		Five Dimensional Case	
	Skew Normal	Normal	Skew Normal	Normal
Mean APB				
Location parameters of the β_q vector	7.7%	58.8%	5.8%	60.4%
Scale parameters of the β_q vector	8.8%	18.3%	15.4%	18.3%
Mean values of the γ_q vector	-	-	4.1%	3.4%
Covariance elements of the γ_q vector	-	-	5.1%	4.3%
Across all parameters β_q and γ_q vector	8.3%	38.6%	7.9%	23.3%
Mean log- likelihood value at convergence	-2056.6	-2095.0	-4132.3	-4219.7
Mean value of the log-likelihood ratio statistic across datasets	76.9		174.8	
Number of times the likelihood ratio test (LRT) favors the skew normal model	Every Time when compared to $\chi^2_3 = 11.34$		Every Time when compared to $\chi^2_3 = 11.34$	