A New Mixed MNP Model Accommodating a Variety of Dependent Non-Normal Coefficient Distributions

Chandra R. Bhat (corresponding author)
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Phone: 1-512-471-4535; Fax: 1-512-475-8744
Email: bhat@mail.utexas.edu
and
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Patrícia S. Lavieri
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Phone: 1-512-471-4535; Fax: 1-512-475-8744
Email: lavieri@utexas.edu

ABSTRACT
In this paper, we propose a general copula approach to accommodate non-normal continuous mixing distributions in multinomial probit (MNP) models. In particular, we specify a multivariate mixing distribution that allows different marginal continuous parametric distributions for different coefficients. A new hybrid estimation technique is proposed to estimate the model, which combines the advantageous features of each of the maximum simulated likelihood inference technique and Bhat's maximum approximate composite marginal likelihood (MACML) inference approach. The effectiveness of our formulation and inference approach is demonstrated through simulation exercises and an empirical application.

Keywords: copula, heterogeneity, MACML, multinomial probit, choice modeling.

1. INTRODUCTION

Econometric discrete choice analysis constitutes the underlying framework for analyzing demand for a variety of consumer commodities and services. For many decades, the discrete choice model employed was the multinomial logit (MNL) model (Luce and Suppes, 1965 and McFadden, 1974), which 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. However, over the past two decades, it has become much more common place to acknowledge the presence of unobserved taste sensitivity in response to variables, as well as accommodate non-IID kernel error terms across alternatives. A general approach to do so is to use a multivariate normal kernel mixed with an appropriately distributed random coefficients vector, which we will label as the mixed multinomial probit (or mixed MNP) model.[1]

An important consideration in the random multivariate mixing (random coefficients) distribution is to explicitly specify it 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 model (whether with an extreme value kernel or an MNP kernel) is capable of approximating any random utility maximization model.[2] For example, it is possible that an analyst may want to specify a naturally bounded distribution (such as a log-normal distribution or a Rayleigh distribution) for cost and time coefficients in a travel choice model, so that the coefficients are strictly negative. Indeed, several studies (see, for example, Amador *et al*., 2005, Train and Sonnier, 2005, Hensher *et al*., 2005, Balcombe *et al.*, 2009, and Torres *et al*., 2011) have underscored the potentially serious misspecification consequences (in terms of theoretical considerations, data fit, as well as trade-off evaluations) of using an unbounded distribution (specifically the normal distribution). Besides, another issue with using an unbounded distribution

---

[1]An analogous structure may be obtained by essentially adding an IID Gumbel error term across alternatives to the multivariate normal coefficients, leading to a mixed multinomial logit model; see Bhat, 1997 and Revelt and Train, 1998 for the first multivariate applications of this type of a model. Alternatively, one can add a multivariate extreme value (MEV) error vector kernel to the utility of the alternatives, combined with additional non-identical kernel error terms, to the random coefficients vector (see, for example, Bhat and Guo, 2007). But, as discussed in detail in Bhat (2011), all these structures essentially achieve the same purpose, and the choice is simply a matter of convenience. Besides, the use of an MNP kernel has substantial advantages when combined with recently proposed analytic methods of evaluating a multivariate cumulative normal distribution (MVNCD) function that have been shown to be much more computationally efficient than traditional simulation approaches. Also, when extensions to accommodate correlation across decision makers due to spatial and/or social interactions are considered, the MNP kernel is much easier and more efficient. We will henceforth focus in this paper on the MNP kernel.

[2] Just to clarify a myth. The mixed multinomial logit model is no more general than the mixed MNP model, as long as we allow the mixing distribution with the MNP kernel to be non-normal, as we do so in the current paper.

that straddles the zero value for the cost coefficient is that it leads to a breakdown of the willingness to pay (WTP) calculations (see Cedilnik *et al*., 2006, Daly *et al*., 2011).

Bhat and Sidharthan (2012) developed a mixed MNP model using a multivariate skew-normal (MVSN) mixing distribution (see also Bhat *et al*., 2015). This model is very effective because the mixing of the MVSN random coefficients distribution with an independent MVN kernel distribution puts the composite error term back to an MVSN form. The MVSN 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 multivariate normal distribution as a special interior point case. It also is a very flexible unimodal density structure that 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 skewness to the right or left is generated by moving probability mass to the left or right of the mean of the normal distribution but keeping the tails thin as in the normal density function, which helps substantially in estimation. In particular, a left-skew is generated by keeping the left tail similar to that of the normal density function, but very sharply reducing the tail on the right side of the mode (see Capitanio, 2010 for a discussion of the rate of decrease in the tail distributions of the skew-normal density function). Thus, to employ a cost coefficient that is strictly constrained to the negative domain, all that the analyst needs to do is to pre-impose a very high skew parameter with a location parameter that is negative (essentially, with a very high skew parameter imposed, the probability density function drops to zero at the location parameter without any overlap on zero; that is, a skew-normal collapses to the so-called half-normal density function with no density to the right of the negative location parameter; see Azzalini, 2013). Additionally, the MVSN-mixed MNP lends itself nicely to estimation using Bhat's (2011) maximum approximate composite marginal likelihood (MACML) approach.

In this paper, we propose an even more general copula-based approach to accommodate non-normal continuous mixing distributions than that proposed in Bhat and Sidharthan (2012).[3]

[3] 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) and transportation (see Bhat, 1997). On the other hand, if a discrete distribution is considered separately for each individual random coefficient, this is essentially a non-parametric random coefficients model (see Bastin *et al*., 2010, Berry and Haile, 2014, il Kim, 2014). The non-parametric specification allows consistent estimates of the observed variable effects under broad model contexts by making regularity (for instance, differentiability) assumptions on an otherwise distribution-free density form. But the flexibility of these methods comes at a high inferential cost since consistency is achieved only in very large samples, parameter estimates have high variance, and

Specifically, the copula-based mixed random coefficients MNP model proposed in this paper allows a multivariate mixing distribution that can combine any continuous distributional shape for each coefficient, including (but not limited to) the skew-normal distribution. This extends the type of continuous multivariate distributions one may want to test, with the only restriction being that the individual coefficient distributions should be continuous. The procedure is based on generating a multivariate continuous distribution through the use of specified parametric univariate continuous coefficient distributions (that can be different for different coefficients) combined with a Gaussian Copula, and is based on Sklar's theorem (Sklar, 1959; see also Bhat and Eluru, 2009 and Joe, 2015). While one may use other copulas to join the different univariate distributions to generate a multivariate distribution, the Gaussian copula used here has many advantages. For instance, the Gaussian copula includes the case of independence across specific coefficients, allows a very flexible and wide range of dependence across coefficients, and is relatively easy to simulate relative to other copula types. It allows dependence across the random coefficients, even if the random coefficients take different marginal distributions. Most importantly, it is the best copula to work with in situations where the analyst is prepared to accept a normal density function for many coefficients, with relatively fewer coefficients specified to have non-normal parametric univariate density functions. This is because, as we will note later, the Gaussian copula requires an integral transformation of each marginal variate into a normal marginal variate. When there are many normal marginal variates, this transformation is not needed for these variates, so that these variates enter directly in the copula (see Equation (7) later), which simplifies the copula construction (with associated optimization convergence and computational speed benefits during model estimation).

The estimation of the copula model is achieved using a combination of the maximum simulated likelihood (MSL) technique (to accommodate the non-normal random coefficients) and Bhat's MACML inference approach (to accommodate all the normal random coefficients as well as the kernel normal error structure; see Bhat, 2011 and Bhat, 2014). This is the first time that a hybrid of these two inference approaches has been proposed in the literature. The combination harnesses the advantages of each of these approaches. The MSL approach is very general and can

---

the computational complexity/effort can be substantial (Mittelhammer and Judge, 2011). Overall, the continuous distribution specification dominates the literature, at least in part because it offers efficiency in the number of mixing distribution parameters to be estimated.

be used to estimate models with any distribution for the random coefficients, including the copula-based model proposed in this paper. However, the approach can be computationally very expensive to ensure good asymptotic estimator properties, and can be prohibitive and literally infeasible (in the context of the computation resources available and the time available for estimation) as the number of random coefficients increases. This is because of the rapid increase in simulation noise and degradation in the accuracy of simulation techniques at medium-to-high dimensions, leading also to convergence problems during estimation and difficulty in estimating the covariance matrix of the MSL estimator (see Bhat, 2011). On the other hand, the MACML approach is simple, computationally very efficient, and simulation-free. It easily and accurately is able to accommodate even a high number of multivariate normally distributed random coefficients, providing both more accuracy (smaller bias in parameters) and orders of magnitude of computational efficiency relative to the MSL inference approach (see Bhat *et al*., 2010, Bhat and Sidharthan, 2012, and Paleti and Bhat, 2013). The other advantage is that the smooth analytically-approximated likelihood function all but ensures convergence during maximization, and also lends itself nicely to relatively smooth second derivative functions to compute the covariance matrix of the estimator. However, the MACML estimator is restricted to normally distributed coefficients or skew-normally distributed coefficients, and does not allow more general parametric random distributions as in the proposed copula MNP model. The combination of the MSL and MACML, however, is especially well suited for the case when there are relatively few non-normally distributed coefficients (so that the simulation does not involve very high dimensions) and many normally distributed coefficients (so that the MACML computational accuracy and efficiency can be realized). However, even in the case when many or even all coefficients are non-normally distributed (with potentially different univariate non-normal distributions for each coefficient), our proposed copula approach provides a systematic parametric framework to engender dependencies (due to unobserved factors) across the non-normal coefficients (rather than pre-imposing independence assumptions on these non-normally distributed coefficients). Of course, if all the coefficients are assumed non-normal and independent, our copula-based hybrid approach collapses exactly to an MSL estimation approach where the univariate integral transforms essentially become vehicles for generating realizations from each of the non-normal univariate distributions. On the other hand, if all the coefficients are assumed to follow a multivariate normal distribution, our copula-based hybrid approach collapses exactly to the MACML estimation approach.

To summarize, in this paper, we develop a general copula-based mixed random coefficients MNP model and propose a hybrid MSL-MACML inference approach for estimation. We demonstrate the effectiveness of our inference approach through simulation exercises as well as an empirical application. The rest of this paper is structured as follows. The next section presents the basics of copula-based multivariate distributions, with an emphasis on the Gaussian copula. The third section presents the proposed model formulation and estimation procedure. Section 4 undertakes simulation exercises to assess the ability of the proposed estimation procedure to recover underlying parameters. Section 5 presents an empirical application of the model on repeated choices data. Finally, Section 6 summarizes the paper and identifies future extensions.

2. COPULA BASICS

In this section, we provide an overview of copula functions, with an emphasis on the Gaussian copula. We also use this section as preparation for the model formulation in the subsequent section. Readers interested in learning more about copula functions are referred to Trivedi and Zimmer (2007), Bhat and Eluru (2009), and Joe (2015).

The word copula, as originally coined by Sklar, 1959, originates from the Latin word "copulare", which means to tie, bond, or connect. The basic idea here is that a joint distribution can always be factored into marginal distributions tied together by a dependence function called the copula. Alternatively, a joint multivariate stochastic dependence relationship (*i.e.*, a multivariate distribution) can be generated by wrapping pre-specified marginal distributions together using an appropriately specified dependence structure called the copula. In essence, the copula approach separates the marginal distributions from the dependence structure, so that the dependence structure is unaffected by the marginal distributions assumed. This provides substantial flexibility in correlating random variables, which may not even have the same marginal distributions. The copulas themselves are multivariate distribution functions defined over the unit cube linking uniformly distributed marginal distributions, the point being that any prespecified marginal distribution can be translated into an equivalent uniform distribution using the integral transform result. So, let C be a K-dimensional copula of uniformly distributed random variables $U_1, U_2, U_3, \ldots, U_K$ with support contained in $[0,1]^K$. Then,

$$C_\theta(u_1, u_2, \ldots, u_K) = \Pr(U_1 < u_1, U_2 < u_2, \ldots, U_K < u_K), \tag{1}$$

where $\boldsymbol{\theta}$ is a parameter vector of the copula commonly referred to as the dependence parameter vector. Now, consider K random variables Y_1, Y_2, Y_3, …, Y_K, each with univariate continuous marginal distribution functions $F_k(y_k) = \Pr(Y_k < y_k)$, k =1, 2, 3, …, K. Then, by the integral transform result, and using the notation $F_k^{-1}(.)$ for the inverse univariate cumulative distribution function, we can write the following expression for each k (k = 1, 2, 3, …, K):

$$F_k(y_k) = \Pr(Y_k < y_k) = \Pr(F_k^{-1}(U_k) < y_k) = \Pr(U_k < F_k(y_k)). \tag{2}$$

A joint K-dimensional distribution function of the random variables with the continuous marginal distribution functions $F_k(y_k)$ can then be generated, using Sklar's (1973) theorem, as follows:

$$\begin{aligned} H(y_1, y_2, ..., y_K) &= \Pr(Y_1 < y_1, Y_2 < y_2, ..., Y_K < y_K) \\ &= \Pr(U_1 < F_1(y_1), U_2 < F_2(y_2), ..., U_K < F_K(y_K)) \\ &= C_{\boldsymbol{\theta}}(u_1, u_2, ..., u_K), \text{ where } u_k = F_k(y_k). \end{aligned} \tag{3}$$

To better understand the generated dependence structures between the original random variables $Y_1, Y_2, ..., Y_K$ (that is, between the elements of the $\boldsymbol{Y}$ vector, where $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_K)'$), concordance measures are used. Basically, two random variables are labeled as being concordant (discordant) if large values of one variable are associated with large (small) values of the other, and small values of one variable are associated with small (large) values of the other. One of the most popular concordance measures of dependence in the copula literature is the Spearman's ρ_S, which measures the dependence between any two random variables (Y_j, Y_k) as follows. Let $(\tilde{Y}_j, \tilde{Y}_k)$ and $(\breve{Y}_j, \breve{Y}_k)$ be independent copies of (Y_j, Y_k). That is, (Y_j, Y_k), $(\tilde{Y}_j, \tilde{Y}_k)$, and $(\breve{Y}_j, \breve{Y}_k)$ are all independent vector pairings, each with a common bivariate distribution function $F_{ij}(.,.)$ and univariate margins F_i and F_j. Then, Spearman's ρ_S is three times the probability of concordance minus the probability of discordance for the two vectors (Y_j, Y_k) and $(\tilde{Y}_j, \breve{Y}_k)$:

$$\rho_S(Y_j, Y_k) = 3\left(P\left((Y_j - \tilde{Y}_j)(Y_k - \breve{Y}_k) > 0\right) - P\left((Y_j - \tilde{Y}_j)(Y_k - \breve{Y}_k)\right) < 0\right). \tag{4}$$

The coefficient "3" is a normalization constant, since the expression in parenthesis is bounded in the region [–1/3, 1/3] (see Nelsen, 2006, pg. 161). It can be shown (see Bhat and Eluru, 2009; Joe, 2015) that the Spearman ρ_S dependence measure for a pair of continuous variables (Y_j, Y_k) is equivalent to the familiar Pearson's correlation coefficient ρ for the grades of Y_1 and Y_2, where the grade of Y_j is $F_j(Y_j)$ and the grade of Y_2 is $F_k(Y_k)$.

2.1. The Gaussian Copula

The Copula functions for use to create multivariate distributions with given marginals are themselves generated in one of several ways, including the method of inversion, geometric methods, and algebraic methods (see Nelsen, 2006; Ch. 3). The most common of these is the inversion method that starts with a known multivariate distribution, and derives a copula function from that. To generate the Gaussian copula, consider the multivariate standard normal distribution function with continuous marginal univariate standard normal distribution functions $\Phi(\tilde{d}_k) = \Pr(D_k < \tilde{d}_k)$ and a correlation matrix $\mathbf{\Gamma}$. Then, the Gaussian copula may be obtained as:

$$\begin{aligned} C_\Gamma(u_1,u_2,...,u_K) &= \Pr(U_1 < u_1, U_2 < u_2, ..., U_K < u_K) \\ &= \Pr(D_1 < \Phi^{-1}(u_1), D_2 < \Phi^{-1}(u_2), ..., D_K < \Phi^{-1}(u_K)) \\ &= \Phi_K(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_K); \mathbf{\Gamma}). \end{aligned} \tag{5}$$

Once a copula is developed, one can revert to Equation (3) to develop new multivariate distributions with arbitrary univariate margins. Thus, the multivariate distribution in Equation (3) with arbitrary marginal distribution functions and a Gaussian copula takes the following form:

$$H(y_1, y_2, ..., y_K) = \Phi_K(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_K); \mathbf{\Gamma}), \text{ where } u_k = F_k(y_k). \tag{6}$$

The Spearman's ρ_S measures for the Gaussian copula above can be written in terms of the dependence (correlation) parameters embedded in the matrix $\mathbf{\Gamma}$. Specifically, the $(\rho_S)_{jk}$ measure for the random variable pair (Y_j, Y_k) can be shown to be $(\rho_S)_{jk} = (6/\pi)\sin^{-1}(\Gamma_{jk}/2)$. Thus, $(\rho_S)_{jk}$ takes on values on [–1, 1]. The reader will note that the Gaussian copula is particularly appealing because it is comprehensive in the dependence structure in that the copula parameterizes the full range of dependence from perfect negative dependence to zero dependence to perfect positive dependence. Also, the Spearman's ρ_S values tracks the correlation parameters Γ_{jk} closely for the Gaussian copula.

Now partition the K-variate random variable vector $\boldsymbol{Y}$ into two sub-vectors $\boldsymbol{Z}$ (of size $E\times1$) and $\boldsymbol{W}$ $(L\times1)$, so that $\boldsymbol{Y} = (\boldsymbol{Z}', \boldsymbol{W}')'$. Let the elements of the $\boldsymbol{Z}$ vector each have a pre-specified but non-normal continuous parametric distribution so that $F_e(z_e) = \Pr(Z_e < z_e)$ (note that the cumulative distribution functions can vary across the elements of $\boldsymbol{Z}$). Let each element of the $\boldsymbol{W}$ vector be normally distributed with mean r_l and standard deviation η_l, so that $F(w_l) = \Pr(W_l < w_l) = \Phi(w_l^*)$, where $w_l^* = \left[\dfrac{w_l - r_l}{\eta_l}\right]$. Then, defining $u_e = F_e(z_e)$, we may write the multivariate distribution in Equation (6) as:

$$
\begin{aligned}
H(z_1, z_2, ..., z_E, w_1, w_2, ..., w_L) &= \Phi_{E+L}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_E), w_1^*, w_2^*, ..., w_L^*; \boldsymbol{\Gamma}\right) \\
&= \Phi_{E+L}\left(g_1, g_2, ..., g_E, w_1^*, w_2^*, ..., w_L^*; \boldsymbol{\Gamma}\right), \text{ where } g_e = \Phi^{-1}(u_e).
\end{aligned}
\tag{7}
$$

The important point to note is that we now have the multivariate distribution of $\boldsymbol{Y} = (\boldsymbol{Z}', \boldsymbol{W}')'$ translated to the multivariate normal distribution of $\tilde{\boldsymbol{Y}} = (\boldsymbol{G}', \boldsymbol{W}')'$, where $G_e = \Phi^{-1}[F_e(Z_e)]$ and $\boldsymbol{G} = (G_1, G_2, ..., G_E)'$. Next, partition the correlation matrix $\boldsymbol{\Gamma}$ as follows:

$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Gamma}_G & \boldsymbol{\Gamma}'_{GW} \\ \boldsymbol{\Gamma}_{GW} & \boldsymbol{\Gamma}_W \end{bmatrix}$. Immediately then, using the conditional distribution properties of the multivariate normal distribution, and defining $\boldsymbol{r} = (r_1, r_2, ... r_L)'$, $\boldsymbol{g} = (g_1, g_2, ... g_E)'$, and a diagonal $L \times L$ matrix $\boldsymbol{\Psi}$ with the l^{th} diagonal element being η_l, we are able to write the conditional distribution of the vector $\boldsymbol{W}$ conditional on $\boldsymbol{Z}$ as follows:

$$
\begin{aligned}
&\boldsymbol{W} \,|\, (\boldsymbol{Z} = \boldsymbol{z}) = \boldsymbol{W} \,|\, (\boldsymbol{G} = \boldsymbol{g}) \sim MVN_L(\boldsymbol{d}, \boldsymbol{\Omega}), \\
&\quad \boldsymbol{d} = \boldsymbol{\Psi}\boldsymbol{\Gamma}_{GW}\boldsymbol{\Gamma}_G^{-1}\boldsymbol{g} + \boldsymbol{r} \text{ and } \boldsymbol{\Omega} = \boldsymbol{\Psi}(\boldsymbol{\Gamma}_W - \boldsymbol{\Gamma}_{GW}\boldsymbol{\Gamma}_G^{-1}\boldsymbol{\Gamma}'_{GW})\boldsymbol{\Psi}.
\end{aligned}
\tag{8}
$$

This conditional distribution for $\boldsymbol{W}$ given $\boldsymbol{Z}$, while accommodating the dependence between the two random vectors, plays a central role in the estimation of the proposed Gaussian copula model, as discussed in the next section.

3. THE MODEL

Consider a repeated choice situation (or a panel situation), with the index q for the individual, $(q = 1, 2, ..., Q)$, index i for the alternative $(i = 1, 2, ..., I)$, and index t for the choice occasion. For ease in presentation, we will use the same number of choice occasions T for every individual. Extension to the case of varying number of choice occasions per individual is straightforward. Also note that the cross-sectional case corresponds to the case of T=1.

Consider the random-coefficients formulation in which the utility that an individual q associates at time period t with alternative i is written as:

$$\tilde{U}_{qti} = \boldsymbol{\beta}'_q \mathbf{x}_{qti} + \boldsymbol{\gamma}'_q \mathbf{s}_{qti} + \tilde{\boldsymbol{\varepsilon}}_{qti}, \quad (t=1, 2, 3,\ldots,T) \tag{9}$$

where $\mathbf{x}_{qti}$ is a $(E \times 1)$-column vector of exogenous attributes (without including constants), $\mathbf{s}_{qti}$ is another $(L \times 1)$-column vector of exogenous attributes (including dummy variables for constants, except in one of the I alternative utilities, say the first alternative), $\boldsymbol{\beta}_q$ is an individual-specific $(E \times 1)$-column vector of coefficients that varies across individuals based on unobserved individual attributes and with each element having a non-normal univariate distribution function $\Pr(\beta_{qe} < z_e) = F_e(z_e)$. $\boldsymbol{\gamma}_q$ is another individual-specific $(L \times 1)$-column vector of MVN-distributed coefficients that varies across individuals based on unobserved individual attributes, with each of its elements having a normal univariate distribution function $\Pr(\gamma_{ql} < w_l) = \Phi(w_l^*)$, $w_l^* = \dfrac{w_l - r_l}{\eta_l}$. Define $\boldsymbol{\alpha}_q = (\boldsymbol{\beta}'_q, \boldsymbol{\gamma}'_q)'$. The correspondence of our notations with the previous section should now be clear, with $\boldsymbol{\beta}_q$ taking the place of $\boldsymbol{Z}$, $\boldsymbol{\gamma}_q$ taking the place of $\boldsymbol{W}$, and $\boldsymbol{\alpha}_q = (\boldsymbol{\beta}'_q, \boldsymbol{\gamma}'_q)'$ corresponding to $\boldsymbol{Y} = (\boldsymbol{Z}', \boldsymbol{W}')'$. Then, following the previous section, we may write the joint cumulative multivariate distribution of $\boldsymbol{\alpha}_q = (\boldsymbol{\beta}'_q, \boldsymbol{\gamma}'_q)'$ exactly as in Equation (7) after translating it into an equivalent joint cumulative multivariate standard normal distribution of $\tilde{\boldsymbol{\alpha}}_q = (\tilde{\boldsymbol{\beta}}'_q, \boldsymbol{\gamma}'_q)'$, with $\tilde{\beta}_{qe} = \Phi^{-1}[F_e(\beta_{qe})]$ and $\tilde{\boldsymbol{\beta}}_q = (\tilde{\beta}_{q1}, \tilde{\beta}_{q2}, ..., \tilde{\beta}_{qE})'$. The correlation matrix $\boldsymbol{\Gamma}$ (of dimension $(E+L) \times (E+L)$) in Equation (7) is partitioned as $\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Gamma}_{\tilde{\beta}} & \boldsymbol{\Gamma}'_{\tilde{\beta}\gamma} \\ \boldsymbol{\Gamma}_{\tilde{\beta}\gamma} & \boldsymbol{\Gamma}_{\gamma} \end{bmatrix}$. Following Equation (8) and the definitions just preceding that equation, we write:

$$\begin{aligned} &\boldsymbol{\gamma}_q \,|\, (\boldsymbol{\beta}_q = \boldsymbol{z}_q) = \boldsymbol{\gamma}_q \,|\, (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) \sim MVN_L(\boldsymbol{d}_q, \boldsymbol{\Omega}), \\ &\quad \boldsymbol{d}_q = \boldsymbol{\Psi}\,\boldsymbol{\Gamma}_{\tilde{\beta}\gamma} \boldsymbol{\Gamma}^{-1}_{\tilde{\beta}} \boldsymbol{g}_q + \boldsymbol{r} \text{ and } \boldsymbol{\Omega} = \boldsymbol{\Psi}\,(\boldsymbol{\Gamma}_{\gamma} - \boldsymbol{\Gamma}_{\tilde{\beta}\gamma} \boldsymbol{\Gamma}^{-1}_{\tilde{\beta}} \boldsymbol{\Gamma}'_{\tilde{\beta}\gamma})\,\boldsymbol{\Psi}. \end{aligned} \tag{10}$$

The (I×1)-vector of kernel error terms, $\tilde{\boldsymbol{\varepsilon}}_{qt} = (\tilde{\varepsilon}_{qt1}, \tilde{\varepsilon}_{qt2}, \tilde{\varepsilon}_{qt3}, \ldots, \tilde{\varepsilon}_{qtI})'$, at each choice occasion is assumed to have a general covariance structure subject to identifiability considerations so that $\tilde{\boldsymbol{\varepsilon}}_{qt} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Theta})$. (note that the $\tilde{\boldsymbol{\varepsilon}}_{qt}$ error terms are considered independent across individuals and choice occasions, and $\tilde{\boldsymbol{\varepsilon}}_{qt}$ is assumed independent of $\boldsymbol{\alpha}_q = (\boldsymbol{\beta}'_q, \boldsymbol{\gamma}'_q)'$; the random vector $\boldsymbol{\alpha}_q$ is also independent across individuals). Since only utility differences matter in discrete choice models, appropriate identification conditions need to be maintained. While there are many ways to ensure identification, a common 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 $\boldsymbol{\varepsilon}_{q1} = (\varepsilon_{q21}, \varepsilon_{q31}, ..., \varepsilon_{qI1})$. Then, up to a scaling factor, the covariance matrix of $\boldsymbol{\varepsilon}_{q1}$ (say $\tilde{\boldsymbol{\Theta}}_1$) 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 $\tilde{\boldsymbol{\Theta}}_1$. $\boldsymbol{\Theta}$ is constructed from $\tilde{\boldsymbol{\Theta}}_1$ by adding a top row of zeros and a first column of zeros.

In addition to the identification condition just discussed, in the case of cross-sectional data, the elements of $\boldsymbol{\gamma}_q$ corresponding to the dummy variables for alternative-specific constants will need to be fixed, and will not have a random distribution. This is because the kernel error terms already absorb the randomness in the constants.

3.1. Model Estimation Using the Hybrid MSL-MACML Approach

With the results and identification considerations from above, we may write Equation (9) as follows:

$$\begin{aligned}
&\tilde{U}_{qti} \mid (\boldsymbol{\beta}_q = \boldsymbol{z}_q) = \tilde{U}_{qti} \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \boldsymbol{z}_q' \mathbf{x}_{qti} + \left[\boldsymbol{\gamma}_q' \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right] \; \mathbf{s}_{qti} + \tilde{\boldsymbol{\varepsilon}}_{qti},\; \boldsymbol{z}_q = (z_{q1}, z_{q2}, ... z_{qE}), \\
&z_{qe} = F_e^{-1}\left[\Phi(g_{qe})\right] = \boldsymbol{z}_q' \mathbf{x}_{qti} + \boldsymbol{d}_q' \mathbf{s}_{qti} + \tilde{\boldsymbol{\gamma}}' \mathbf{s}_{qti} + \tilde{\boldsymbol{\varepsilon}}_{qti},\; \tilde{\boldsymbol{\gamma}} \sim MVN_L(\mathbf{0}, \boldsymbol{\Omega}).
\end{aligned} \tag{11}$$

We now set out some additional notation. Define $\tilde{\mathbf{U}}_{qt} = (\tilde{U}_{qt1}, \tilde{U}_{qt2}, ..., \tilde{U}_{qtI})'$ ($I \times 1$ vector), $\tilde{\mathbf{U}}_q = (\tilde{\mathbf{U}}_{q1}', \tilde{\mathbf{U}}_{q2}', ..., \tilde{\mathbf{U}}_{qT}')'$ ($TI \times 1$ vector), $\tilde{\boldsymbol{\varepsilon}}_{qt} = (\tilde{\varepsilon}_{qt1}, \tilde{\varepsilon}_{qt2}, ..., \tilde{\varepsilon}_{qtI})'$ ($I \times 1$ vector), $\tilde{\boldsymbol{\varepsilon}}_q = (\tilde{\boldsymbol{\varepsilon}}_{q1}', \tilde{\boldsymbol{\varepsilon}}_{q2}', ..., \tilde{\boldsymbol{\varepsilon}}_{qT}')'$ ($TI \times 1$ vector), $\mathbf{x}_{qt} = (\mathbf{x}_{qt1}, \mathbf{x}_{qt2}, ..., \mathbf{x}_{qtI})'$ ($I \times E$ matrix), $\mathbf{x}_q = (\mathbf{x}_{q1}', \mathbf{x}_{q2}', ..., \mathbf{x}_{qT}')'$ ($TI \times E$ matrix), $\mathbf{s}_{qt} = (\mathbf{s}_{qt1}, \mathbf{s}_{qt2}, ..., \mathbf{s}_{qtI})'$ ($I \times L$ matrix), $\mathbf{s}_q = (\mathbf{s}_{q1}', \mathbf{s}_{q2}', ..., \mathbf{s}_{qT}')'$ ($TI \times L$ 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 Equation (11) in matrix form as:

$$\tilde{\mathbf{U}}_q \mid (\boldsymbol{\beta}_q = \boldsymbol{z}_q) = \tilde{\mathbf{U}}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \left[(\mathbf{x}_q \boldsymbol{z}_q + \mathbf{s}_q \boldsymbol{d}_q) \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right] + \left[\mathbf{s}_q \tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\varepsilon}}_q\right] \;. \tag{12}$$

From above, it is clear that $\tilde{\mathbf{U}}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)$ is multivariate normally distributed: $\tilde{\mathbf{U}}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) \sim MVN_{TI}(\mathbf{V}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q), \tilde{\boldsymbol{\Xi}}_q)$, where $\mathbf{V}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \left[(\mathbf{x}_q \boldsymbol{z}_q + \mathbf{s}_q \boldsymbol{d}_q) \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right]$ and $\tilde{\boldsymbol{\Xi}}_q = \mathbf{s}_q \boldsymbol{\Omega} \mathbf{s}'_q + (\mathbf{IDEN}_T \otimes \boldsymbol{\Theta})$.

Let the individual q choose alternative m_{qt} at the t^{th} choice occasion. Define $\mathbf{M}_q$ as an $[T \times (I-1)] \times [TI]$ block-diagonal matrix, each block diagonal being of size $(I\text{-}1) \times (I)$ and containing the matrix $\mathbf{M}_{qt}$. $\mathbf{M}_{qt}$ itself is an identity matrix of size ($I-1$) with an extra column of '-1' values added at the m_{qt}^{th} column. Let $\mathbf{B}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \mathbf{M}_q \left[\mathbf{V}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right]$ and $\boldsymbol{\Xi}_q = \mathbf{M}_q \tilde{\boldsymbol{\Xi}}_q \mathbf{M}'_q$. The parameter vector to be estimated is $\boldsymbol{\lambda} = (\boldsymbol{\delta}', \mathbf{r}', \text{Vech}(\boldsymbol{\Psi}), \text{Vech}(\boldsymbol{\Gamma}), \text{Vech}(\boldsymbol{\Theta}))'$, where $\boldsymbol{\delta}$ represents a column vector that collects all the parameters characterizing the non-normal coefficients $\boldsymbol{\beta}_q$, Vech($\boldsymbol{\Gamma}$) is a column vector obtained by vertically stacking the upper triangle elements of the matrix $\boldsymbol{\Gamma}$, Vech($\boldsymbol{\Psi}$) is another column vector obtained by vertically stacking the upper triangle elements of the matrix $\boldsymbol{\Psi}$, and Vech($\boldsymbol{\Theta}$) is a third column vector obtained by vertically stacking the estimable upper triangular elements of the matrix $\boldsymbol{\Theta}$. The likelihood contribution of individual q conditional on $\boldsymbol{\beta}_q = \boldsymbol{z}_q$ (that is, $\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q$) is as below:

$$L_q(\boldsymbol{\lambda}) \mid (\boldsymbol{\beta}_q = \boldsymbol{z}_q) = L_q(\boldsymbol{\lambda}) \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \Phi_{\tilde{J}}\left(\left[-\mathbf{B}_q^* \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right], \boldsymbol{\Xi}_q^{\;*}\right), \tag{13}$$

where $\tilde{J} = T \times (I-1)$, $\mathbf{B}_q^* \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \boldsymbol{\omega}_{\Xi_q}^{-1} \mathbf{B}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)$, $\boldsymbol{\Xi}_q^{\;*} = \boldsymbol{\omega}_{\Xi_q}^{-1} \boldsymbol{\Xi}_q \boldsymbol{\omega}_{\Xi_q}^{-1}$, and $\boldsymbol{\omega}_{\Xi_\text{q}}$ is the diagonal matrix of standard deviations of $\boldsymbol{\Xi}_\text{q}$. Finally, the unconditional likelihood contribution of individual q is:

$$L_q(\boldsymbol{\lambda}) = \int_{z=-\infty}^{z=+\infty} \left[\Phi_{\tilde{J}}\left(\left[-\mathbf{B}_q^* \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right], \boldsymbol{\Xi}_q^{\;*}\right)\right] f_E(\mathbf{z}) d\boldsymbol{z} = \int_{g=-\infty}^{g=+\infty} \left[\Phi_{\tilde{J}}\left(\left[-\mathbf{B}_q^* \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right], \boldsymbol{\Xi}_q^{\;*}\right)\right] \phi_E(\boldsymbol{g}; \boldsymbol{\Gamma}_{\tilde{\beta}}) d\boldsymbol{g}, \tag{14}$$

where $\phi_E(\boldsymbol{g};\boldsymbol{\Gamma}_{\tilde{\beta}})$ is the E-variate multivariate standard normal density function with correlation matrix $\boldsymbol{\Gamma}_{\tilde{\beta}}$, and evaluated at the vector $\boldsymbol{g}$.[4] The reader will note that the vector $\boldsymbol{\delta}$ of the moment parameters characterizing the non-normal coefficients $\boldsymbol{\beta}_q$ appears in the above function through $\mathbf{B}_q^* \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)$, which itself is a function of $\mathbf{V}_q \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \left[(\mathbf{x}_q \boldsymbol{z}_q + \mathbf{s}_q \boldsymbol{d}_q) \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q)\right]$. In the latter expression, each element of the vector $\boldsymbol{z}_q$ is computed as $z_{qe} = F_e^{-1}\left[\Phi(g_{qe})\right]$ during the integration over the vector $\boldsymbol{g}_q$ in Equation (14), and the parameters comprising $\boldsymbol{\delta}$ feature in the inverse function $F_e^{-1}(.)$. Thus, the proposed copula model allows consideration of a whole variety of non-normal multivariate random coefficient distributions, though using distributions that have a closed-form inverse function make the computation easier than when there is no closed-form. Importantly, the elements of the vector $\boldsymbol{\beta}_q$ can have different non-normal distributions. The support of each non-normal element can range from the entire real line to only the positive (or negative) half-line. While there are many distributions that have support on the entire real line, Table 1 provides a sample list of univariate marginal distributions that may be considered for elements that are strictly restricted to the positive half-line, have at least the first and second inverse moments that exist (important for willingness to pay computations where an element appears in the denominator of a ratio), and have closed-form inverse (or quantile) functions. Of these, we would particularly like to bring attention to the last of these distributions – the power log-normal distribution that has received little attention in the statistical literature and no attention at all in the context of coefficient distributions in discrete choice models. The advantage we see in this distribution relative to other distributions (including the log-normal) is that it can both allow for substantial heterogeneity (large variance parameter) while also ensuring that the skewed tail is relatively thin. This helps because convergence during estimation is much easier.[5] Figure 1 shows a comparison of the log-normal

[4] Note that, by construction, the marginal multivariate distribution function of $\boldsymbol{\beta}_q$ is the multivariate standard normal distribution function of $\tilde{\boldsymbol{\beta}}_q$; that is $F_E(\boldsymbol{\beta}_q < \mathbf{z}_q) = \Phi_E(\boldsymbol{g}_q;\boldsymbol{\Gamma}_{\tilde{\beta}})$, from which $f_E(\mathbf{z}_q) = \dfrac{dF_E(\boldsymbol{\beta}_q < \mathbf{z}_q)}{d\mathbf{z}_q} = \phi_E(\boldsymbol{g}_q;\boldsymbol{\Gamma}_{\tilde{\beta}})\dfrac{d\boldsymbol{g}_q}{d\mathbf{z}_q}$, or $f_E(\mathbf{z}_q)d\mathbf{z}_q = \phi_E(\boldsymbol{g}_q;\boldsymbol{\Gamma}_{\tilde{\beta}})d\boldsymbol{g}_q$, and Equation (14) is the result.

[5] On the other hand, the problem with the log-normal distribution to represent a coefficient such as a cost coefficient is that the tails of the distribution are directly determined by the variance term. If there is high heterogeneity in the sensitivity to cost, this immediately implies a peaking (mode) close to zero as well as a long and fat left tail (note that the cost coefficient is introduced as the negative of the log-normal distribution). The result is that, as the variance parameter of the log-normal distribution increases (for the same mean parameter), a larger fraction of individuals will have a small cost coefficient. At the same time, a small fraction of individuals will have very high cost sensitivity

and the power log-normal for identical values of μ and σ, but with different values of p in the power log-normal (when p=1 in the power log-normal, it collapses to the log-normal). Figure 1 plots the power log-normal only for p>1, which leads to thinner tails than the log-normal. The constraint p>1 can be maintained by reparametrizing p as $p = 1 + \exp(\tilde{p})$. In this sense, the power log-normal with p>1 is like the skew-normal in that it creates skew while keeping the tails thin.

The simulation approaches for evaluating the full likelihood function in Equation (14) involve integration of dimension $T \times (I-1) + E$, which can explode quickly as the number of choice occasions of the same individual increases (in the case of a cross-sectional model with only one observation per individual, T=1, and the integral dimensionality is only $(I-1)+E$). However, one can consider the following (pairwise) composite marginal likelihood function formed by taking the products (across the T choice occasions) of the joint pairwise probability of the chosen alternatives m_{qt} for the t^{th} choice occasion and $m_{qt'}$ for the t^{th} choice occasion for individual q.

$$L_{CML,q}(\boldsymbol{\lambda}) = \prod_{t=1}^{T-1} \prod_{t'=t+1}^{T} L_{CML,qtt'}(\boldsymbol{\lambda}), \tag{15}$$

where

$$L_{CML,qtt'}(\boldsymbol{\lambda}) = \int_{g_q=-\infty}^{g_q=+\infty} \left[\Phi_{\breve{J}} \left(\left[-\widetilde{\mathbf{B}}^{*}_{qtt'} \mid (\widetilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) \right], \widetilde{\boldsymbol{\Xi}}^{*}_{qtt'} \right) \right] \phi_E(\boldsymbol{g}_q; \boldsymbol{\Gamma}_{\widetilde{\beta}}) d\boldsymbol{g}_q, \tag{16}$$

where $\breve{J} = 2(I-1)$, $\widetilde{\mathbf{B}}^{*}_{qtt'} \mid (\widetilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) = \boldsymbol{\Delta}_{qtt'} \left(\mathbf{B}^{*}_q \mid (\widetilde{\boldsymbol{\beta}}_q = \boldsymbol{g}_q) \right)$, $\widetilde{\boldsymbol{\Xi}}^{*}_{qtt'} = \boldsymbol{\Delta}_{qtt'} \boldsymbol{\Xi}^{*}_q \boldsymbol{\Delta}'_{qtt'}$, and $\boldsymbol{\Delta}_{qtt'}$ is a $\breve{J} * \widetilde{J}$-selection matrix with an identity matrix of size $(I-1)$ occupying the first $(I-1)$ rows and the $[(t-1)\times(I-1)+1]^{th}$ through $[t \times (I-1)]^{th}$ columns, and another identity matrix of size $(I-1)$ occupying the last $(I-1)$ rows and the $[(t'-1)\times(I-1)+1]^{th}$ through $[t' \times (I-1)]^{th}$ columns. All other elements of $\boldsymbol{\Delta}_{qtt'}$ take the value of zero. The pairwise likelihood function now only needs the evaluation of a $[2 \times (I-1) + E]$-dimensional integral. Note also that, in a cross-sectional model (T=1), the CML likelihood function of Equation (15) has no pairings to consider and effectively collapses to the full likelihood function of Equation (14), involving the evaluation of an

because of the long and fat tail. The result can cause unusually large and small willingness to pay estimates. Further, the long and fat tail on the unbounded side of the distribution is known to cause convergence problems during estimation (Bartels *et al*., 2006).

$[(I-1)+E]$-dimensional integral. Finally, it is important to note that the same draws have to be used for the integration over $\boldsymbol{g}_q$ across all pairings corresponding to the same individual q.

The properties of the general CML estimator may be derived using the theory of estimating equations (see Bhat, 2014). Under usual regularity conditions, the maximization of the logarithm of the CML function, where the CML function across all the Q individuals is $L_{CML}(\boldsymbol{\lambda})=\prod_{q=1}^{Q} L_{CML,q}(\boldsymbol{\lambda})$, is achieved by solving the composite score equations that are themselves linear combinations of valid likelihood score functions associated with the event probabilities forming the composite log-likelihood function. Thus, the score equations immediately satisfy the requirement of being unbiased. Further, with q independent observations with panel data or repeated choice data, in the asymptotic scenario that $Q \to \infty$, a central limit theorem and a first-order Taylor series expansion can be applied in the usual way (see, for example, Godambe, 1960) to the resulting mean composite score function to obtain consistency and asymptotic normality of the CML estimator (see Section 1.4 of Bhat, 2014).

The covariance matrix is estimated as:

$$\frac{\hat{\boldsymbol{G}}^{-1}}{Q}=\frac{\left[\hat{\boldsymbol{H}}^{-1}\right]\left[\hat{\boldsymbol{J}}\right]\left[\hat{\boldsymbol{H}}^{-1}\right]'}{Q}, \text{ with}$$

$$\hat{\boldsymbol{H}}=-\frac{1}{Q}\left[\sum_{q=1}^{Q}\frac{\partial^2 \log\left[L_{CML,q}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right]_{\hat{\lambda}_{CML}}=-\frac{1}{Q}\left[\sum_{q=1}^{Q}\sum_{t=1}^{T-1}\sum_{t'=t+1}^{T}\frac{\partial^2 \log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right]_{\hat{\lambda}_{CML}}$$

$$\begin{aligned}\hat{\boldsymbol{J}}&=\frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\frac{\partial\log\left[L_{CML,q}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}\right)\left(\frac{\partial\log\left[L_{CML,q}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}\right)\right]_{\hat{\lambda}_{CML}}\\&=\frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\sum_{t=1}^{T-1}\sum_{t'=t+1}^{T}\frac{\partial\log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}\right)\left(\sum_{t=1}^{T-1}\sum_{t'=t+1}^{T}\frac{\partial\log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}\right)\right]_{\hat{\lambda}_{CML}}\end{aligned} \tag{17}$$

An alternative estimator for $\hat{\boldsymbol{H}}$ is as below:

$$\hat{\boldsymbol{H}}=\frac{1}{Q}\sum_{q=1}^{Q}\sum_{t=1}^{T-1}\sum_{t'=1}^{T}\left(\left[\frac{\partial\log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}\right]\left[\frac{\partial\log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}\right]\right)_{\hat{\lambda}_{CML}}$$

In the special case of a cross-sectional model, there are no pairings to consider and the covariance matrix collapses to the traditional inverse of the sandwich information matrix.

There are two final issues. The first is that the covariance matrices $\mathbf{\Gamma}$ and $\tilde{\mathbf{\Theta}}_1$ have to be positive definite. The simplest way to ensure the positive-definiteness of these matrices is to use a Cholesky-decomposition and parameterize the CML function in terms of the Cholesky parameters (rather than the original covariance matrices). Also, the matrix $\mathbf{\Gamma}$ is a correlation matrix, which can be maintained by writing each diagonal element (say the aa^{th} element) of the lower triangular Cholesky matrix of $\mathbf{\Gamma}$ as $\sqrt{1-\sum_{j=1}^{a-1} l_{aj}^2}$, where the l_{aj} elements are the Cholesky factors that are estimated. Using these Cholesky-parameterization, the parameters to be estimated in the model may be written as: $\boldsymbol{\lambda} = (\boldsymbol{\delta}', \mathbf{r}', \text{Vech}(\mathbf{\Psi}), \text{Vech}(\mathbf{L}_{\Gamma}), \text{Vech}(\mathbf{L}_{\Theta}))'$, where $\mathbf{L}_{\Gamma}$ is the parameterized (as above) lower Cholesky matrix of the matrix $\mathbf{\Gamma}$ and $\mathbf{L}_{\Theta}$ represents the lower Cholesky matrix of the estimable parameters of $\tilde{\mathbf{\Theta}}_1$ (as indicated earlier, $\mathbf{\Theta}$ is constructed from $\tilde{\mathbf{\Theta}}_1$).

The second issue relates to the starting parameters. In our experimentation of alternative procedures to arrive at good starting values, the following procedure worked well: (a) Assume a kernel error term covariance matrix that corresponds to an IID error structure across the alternatives with a variance of one-half for each alternative error term), (b) Estimate the parameters characterizing the marginal non-normal and normal coefficients, fixing the parameters of the copula correlation matrix and the kernel covariance matrix to their starting values discussed above, and (c) Use the coefficient vector from the estimation results in step (b) to begin the iterations for the overall estimation of the model system.

3.2. Alternative Estimation Procedure

An alternative estimation procedure is to develop the likelihood function for each individual conditional on both the $\boldsymbol{\beta}_q$ and $\boldsymbol{\gamma}_q$ vectors, and then integrate both out at the end (as opposed to the procedure in the previous section of first writing the conditional likelihood given $\boldsymbol{\beta}_q$ and then integrating out $\boldsymbol{\beta}_q$) . In this alternative procedure, using the earlier definitions, we first write $\tilde{\mathbf{U}}_{qt} \,|\, (\boldsymbol{\beta}_q = \boldsymbol{z}, \boldsymbol{\gamma}_q = \boldsymbol{w}) = \tilde{\mathbf{U}}_{qt} \,|\, (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}, \tilde{\boldsymbol{\gamma}}_q = \boldsymbol{w}^*) = \left[(\mathbf{x}_{qt}\boldsymbol{z} + \mathbf{s}_{qt}\boldsymbol{w})\right] + \tilde{\boldsymbol{\varepsilon}}_{qt}$. Next, defining $\tilde{\mathbf{A}}_{qt} = (\mathbf{x}_{qt}\boldsymbol{z} + \mathbf{s}_{qt}\boldsymbol{w})$, $\mathbf{A}_{qt} = \mathbf{M}_{qt}\tilde{\mathbf{A}}_{qt}$, $\ddot{\mathbf{\Theta}} = \mathbf{M}_{qt}\mathbf{\Theta}\mathbf{M}'_{qt}$, $\mathbf{A}^*_{qt} = \boldsymbol{\omega}_{\Xi_q}^{-1}\mathbf{A}_{qt}$, $\ddot{\mathbf{\Theta}}^* = \boldsymbol{\omega}_{\ddot{\Theta}}^{-1}\ddot{\mathbf{\Theta}}\,\boldsymbol{\omega}_{\ddot{\Theta}}^{-1}$, the likelihood function at choice occasion t conditional on $\boldsymbol{\beta}_q = \boldsymbol{z}$ and $\boldsymbol{\gamma}_q = \boldsymbol{w}$ (that is, $\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}, \tilde{\boldsymbol{\gamma}}_q = \boldsymbol{w}^*$) is

$$L_{qt}(\boldsymbol{\lambda}) \mid (\boldsymbol{\beta}_q = \boldsymbol{z}, \boldsymbol{\gamma}_q = \boldsymbol{w}) = L_{qt}(\boldsymbol{\lambda}) \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}, \tilde{\boldsymbol{\gamma}}_q = \boldsymbol{w}^*)) = \Phi_{(I-1)}\left(\left[-\mathbf{A}_{qt}^* \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}, \tilde{\boldsymbol{\gamma}}_q = \boldsymbol{w}^*)\right], \ddot{\boldsymbol{\Theta}}^*\right)$$

and the individual-level likelihood function is:

$$L_q(\boldsymbol{\lambda}) = \int_{g=-\infty}^{g=\infty} \int_{w^*=-\infty}^{w^*=\infty} \prod_{t=1}^{T} \Phi_{(I-1)}\left(\left[-\mathbf{A}_{qt}^* \mid (\tilde{\boldsymbol{\beta}}_q = \boldsymbol{g}, \tilde{\boldsymbol{\gamma}}_q = \boldsymbol{w}^*)\right], \ddot{\boldsymbol{\Theta}}^*\right) \phi_E(\boldsymbol{g}, \boldsymbol{w}^*; \boldsymbol{\Gamma}) \boldsymbol{dgdw}^*. \tag{18}$$

The above function involves the evaluation of an E+L-dimensional outer integral followed by evaluations of ($I-1$)-dimensional orthant inner integrals.

In the cross-sectional case, the estimation procedure from Section 3.1 is much more computationally efficient. This is because the estimation procedure from earlier exploits the fact that the conditional distribution of a subset of multivariate normally distributed coefficients involved in a copula-generated larger multivariate distribution, given the subset of non-normally distributed coefficients, is also multivariate normally distributed. To our knowledge, this is the first time this specific property of the multivariate Gaussian copula has been exploited in the way we do. Then, the conditional multivariate normal distribution of coefficients is combined with the kernel error multivariate normal distribution, so that the resulting multivariate normal distribution of the utilities (conditional on the non-normally distributed coefficients) has the same dimensionality as the kernel distribution of the utility error terms (that is, $I-1$). This leads to a reduction by L (the number of normally distributed coefficients) in the dimensionality of integration in the earlier estimation procedure than the one in the current section. As importantly, as indicated earlier in this paper, as the number of dimensions for integration increases, convergence problems arise in the MSL approach and the time for convergence increases substantially. On the other hand, by using the MSL approach only for the non-normal coefficients (which tend to be very few in number in most applications), and using a smooth analytic evaluation approach for the ($I-1$)-dimensional orthant multivariate distribution function (as we propose and implement in this paper using Bhat's MACML approach), convergence problems get reduced as does the computational time.

In the panel case, the full information likelihood of Equation (14) in Section 3.1 becomes difficult to impractical as the number of choice occasions per individual (*i.e.*, T) increases. However, the CML of Equation (16) in the previous section still retains substantial advantages compared to the MSL estimation technique of this section in Equation (18). This is because of three reasons. First, the $(I-1)\times 2$ orthant multivariate probability in Equation (16) is conveniently computed using the MACML procedure, which breaks this multivariate probability into solely bivariate and univariate cumulative normal distribution function computations (Bhat, 2011). Second, having a well behaved and smooth analytic expression as the integrand over which only a few non-normally distributed coefficients need to be integrated will generally lead to much superior convergence and computational properties rather than the alternative of simulating over all normal and non-normally distributed coefficients. Third, when the number of choice occasions increases, the result is that the integrand in Equation (18) becomes smaller and smaller (because it is the product of probabilities over all choice occasions), leading to potential problems in convergence (artificial scaling approaches may be devised to keep the integrand from getting too small, but this has limited use as the number of choice occasions increases). On the other hand, the CML of Equation (16) does not have this problem, because the logarithm of this equation leads to summations outside the $(I-1)\times 2$-dimensional integral. But the CML of Equation (16) also involves more and more pairings as the number of choice occasions increases. Fortunately, one can use a different CML function than that in Equation (16) in such cases. Specifically, instead of taking all pairings, one can develop a CML function that only includes a specified number of randomly chosen choice occasions (say T') to form the pairings, while leaving the others independent. For ease in presentation, assume that the choice occasions are ordered so that the randomly chosen T' pairings appear first for each individual. Then, the individual-level contribution to the CML is:

$$\widetilde{L}_{CML,q}(\boldsymbol{\lambda}) = \prod_{t=1}^{T'-1} \prod_{t'=t+1}^{T'} L_{CML,qtt'}(\boldsymbol{\lambda}) \prod_{t=T'+1}^{T} L_{CML,qt}(\boldsymbol{\lambda}) \tag{19}$$

where $L_{CML,qtt'}(\boldsymbol{\lambda})$ is defined as earlier, and

$$L_{CML,qt}(\boldsymbol{\lambda}) = \int_{g=-\infty}^{g=+\infty} \left[\Phi_{(I-1)}\left(\left[-\vec{\mathbf{B}}^{*}_{qt} \mid (\widetilde{\boldsymbol{\beta}}_q = \boldsymbol{g})\right], \ddot{\boldsymbol{\Xi}}^{*}_{qt}\right)\right]\phi_E(\boldsymbol{g}; \boldsymbol{\Gamma}_{\widetilde{\beta}})\boldsymbol{dg}\,, \tag{20}$$

with $\vec{\mathbf{B}}^{*}_{qtt'} \mid (\widetilde{\boldsymbol{\beta}}_q = \boldsymbol{g}) = \vec{\boldsymbol{\Delta}}_{qt}\left(\mathbf{B}^{*}_{q} \mid (\widetilde{\boldsymbol{\beta}}_q = \boldsymbol{g})\right), \ddot{\boldsymbol{\Xi}}^{*}_{qt} = \vec{\boldsymbol{\Delta}}_{qt}\boldsymbol{\Xi}^{*}_{q}\vec{\boldsymbol{\Delta}}'_{qt}$, and $\vec{\boldsymbol{\Delta}}_{qt}$ is a $(I-1)*\widetilde{J}$-selection matrix with an identity matrix of size $(I-1)$ occupying the first $(I-1)$ rows and the $\left[(t-1)\times(I-1)+1\right]^{th}$ through $\left[t\times(I-1)\right]^{th}$ columns. All other elements of $\vec{\boldsymbol{\Delta}}_{qt}$ take the value of zero. The covariance matrix is estimated as in Equation (17), with the following substitutions:

$$\hat{\boldsymbol{H}} == -\frac{1}{Q}\left[\sum_{q=1}^{Q}\left(\sum_{t=1}^{T'-1}\sum_{t'=t+1}^{T'}\frac{\partial^2 \log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}+\sum_{t=T'+1}^{T}\frac{\partial^2 \log\left[L_{CML,qt}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right)\right]_{\hat{\lambda}_{CML}}$$

$$\hat{\boldsymbol{J}} = \frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\frac{\partial \log\left[L_{CML,q}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}\right)\left(\frac{\partial \log\left[L_{CML,q}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}\right)\right]_{\hat{\lambda}_{CML}} \tag{21}$$

$$= \frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\sum_{t=1}^{T'-1}\sum_{t'=t+1}^{T'}\frac{\partial \log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}+\sum_{t=T'+1}^{T}\frac{\partial \log\left[L_{CML,qt}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}\right)\left(\sum_{t=1}^{T'-1}\sum_{t'=t+1}^{T'}\frac{\partial \log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}+\sum_{t=T'+1}^{T}\frac{\partial \log\left[L_{CML,qt}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}\right)\right]_{\hat{\lambda}_{CML}}$$

An alternative estimator for $\hat{\boldsymbol{H}}$ is as below:

$$\hat{\boldsymbol{H}} = \frac{1}{Q}\sum_{q=1}^{Q}\left[\sum_{t=1}^{T-1}\sum_{t'=1}^{T}\left(\left[\frac{\partial \log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}\right]\left[\frac{\partial \log\left[L_{CML,qtt'}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}\right]\right)+\sum_{t=T'+1}^{T}\left(\left[\frac{\partial \log\left[L_{CML,qt}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}}\right]\left[\frac{\partial \log\left[L_{CML,qt}(\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'}\right]\right)\right]_{\hat{\lambda}_{CML}}$$

4. SIMULATION EVALUATION

Simulations were performed for two different distributional configurations of random parameters. In both set-ups, we consider a cross-sectional mixed MNP model with four alternatives and three independent variables (a panel mixed MNP is considered in the empirical analysis). The values of each of the three independent variables for the alternatives are drawn from a standard univariate normal distribution. Once drawn, the exogenous variables are held fixed for the data set. We generate a sample of 3000 realizations of the three independent variables corresponding to a situation of 3000 choice occasions.

We allow random coefficients on all the three independent variables. In the first set of simulations, two of the three coefficients are assumed to be realizations from power log-normal distributions with identical location parameters (μ_1 and μ_2=0.5), identical scale parameters (σ_1 and σ_2=1.0), and identical power terms (p_1 and p_2= 5, considered fixed).[6] The last coefficient is assumed to be a realization from a normal distribution with mean r = 0.5 and standard deviation η = 1.5. In the second set of simulations, the three coefficients are assumed to be realizations of

[6] As discussed earlier, the log-normal distribution *a priori* fixes the power term to 1. Here, while we can estimate the power term, our experience suggested that the optimization algorithms took longer with much more convergence difficulty than if the power term was fixed. That is, the best way to estimate a model with a power log-normal term appears to be to estimate the model at different fixed values of the power term, and then compare the data fits across the different optimization function values (corresponding to different fixed values of the power term) to determine the best value for the power term. That is the reason we fix the power term at the value of three in the simulation estimations here, while estimating the means (μ_1 and μ_2) and the scale parameters (σ_1 and σ_2).

different distributions: power log-normal (μ_1=0.5, σ_1=1.0 and $p = 5$), exponential (μ_2=1.0 and σ_2=0.8) and normal (r = 0.5 and η=1.5) respectively. In both cases all the parameters except for p are freely estimated. The reason for testing two settings of simulations with different distributional configurations is to evaluate the performance of the model in recovering parameters vis-à-vis different distribution shapes (tail length).

To ensure the positivity of the scale parameters σ_1 and σ_2, we parameterize them as $\sigma_j = \exp(\tilde{\sigma}_j),\ j = 1,2$ in estimation. The first two random coefficients in the above setup constitute the $\boldsymbol{\beta}_q$ vector in the notation of Section 3, with $\boldsymbol{\delta} = (\mu_1, \sigma_1, p_1, \mu_2, \sigma_2)$. The normal distribution scale parameter for the third coefficient is also parameterized as $\eta = \exp(\tilde{\eta})$ in estimation (technically, because of the symmetric nature of the normal distribution, one can let the standard deviation to be free, and simply change the sign if it is estimated to be negative; but we prefer the parametrization from the beginning to help the optimization process along a single line search direction). In the notation of Section 2.1, $\mathbf{r} = (r)$ and $\text{Vech}(\boldsymbol{\Psi}) = (\eta)$. All of these coefficients are tied together through the dependency (correlation) matrix of the Gaussian copula. The correlation structure used in the first and in the second sets of simulations is as follows:

$$1)\quad \boldsymbol{\Gamma} = \begin{bmatrix} 1.0 & 0.6 & 0.4 \\ 0.6 & 1.0 & 0.4 \\ 0.4 & 0.4 & 1.0 \end{bmatrix} = \mathbf{L}_\Gamma \mathbf{L}'_\Gamma = \begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 0.600 & 0.800 & 0.000 \\ 0.400 & 0.200 & 0.894 \end{bmatrix} \begin{bmatrix} 1.0000 & 0.600 & 0.400 \\ 0.0000 & 0.800 & 0.200 \\ 0.0000 & 0.0000 & 0.894 \end{bmatrix}$$

$$2)\quad \boldsymbol{\Gamma} = \begin{bmatrix} 1.0 & 0.4 & 0.6 \\ 0.4 & 1.0 & 0.6 \\ 0.6 & 0.6 & 1.0 \end{bmatrix} = \mathbf{L}_\Gamma \mathbf{L}'_\Gamma = \begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 0.400 & 0.917 & 0.000 \\ 0.600 & 0.393 & 0.697 \end{bmatrix} \begin{bmatrix} 1.0000 & 0.400 & 0.600 \\ 0.0000 & 0.917 & 0.393 \\ 0.0000 & 0.0000 & 0.697 \end{bmatrix}$$

Again, as indicated earlier, to maintain positive definiteness, we work with the Cholesky decomposition elements of the correlation matrix of the Gaussian copula. Thus, there are three Cholesky matrix elements to be estimated in $\mathbf{L}_\Gamma$ corresponding to the non-diagonal elements in the matrices above (note that the diagonal elements are simply a function of the non-diagonal elements and are not estimated directly, because $\boldsymbol{\Gamma}$ is a correlation matrix with unit diagonals). Collectively, then, $\text{Vech}(\mathbf{L}_\Gamma) = (l_{\Gamma 1}, l_{\Gamma 2}, l_{\Gamma 3})' = (0.6, 0.4, 0.2)'$ for the first set and $(0.4, 0.6, 0.393)'$ for the second set. The important point to note is that the specification above generates dependence across the different distributions.

With the preliminaries above, the vector $\boldsymbol{\alpha}_q = (\boldsymbol{\beta}'_q, \boldsymbol{\gamma}'_q)'$ is generated as follows for the first case in which two of the coefficients follow a power lognormal distribution and the third follows a normal distribution: (a) First draw a three-variate realization of $(\tilde{\boldsymbol{\beta}}_q, \tilde{\boldsymbol{\gamma}}_q)$ from the multivariate standard normal distribution of three dimensions with a mean vector of all zero elements and correlation matrix $\boldsymbol{\Gamma}$, (b) Obtain the realization of β_{qj} as $F_1^{-1}\left[\Phi(\tilde{\beta}_{qj})\right] = \exp\left[-\sigma_j \Phi^{-1}\left[\left[1-\Phi(\tilde{\beta}_{qj})\right]^{1/p}\right]\right] + \mu_j$, $j = 1,2$, (c) Obtain the realization of the one-dimensional vector $\boldsymbol{\gamma}_q$ as $\boldsymbol{\gamma}_q = \boldsymbol{\Psi}\tilde{\boldsymbol{\gamma}}_q + \mathbf{r}$,where $\boldsymbol{\Psi}$ is the one-dimensional (in this simulation case) diagonal matrix with the element η as the scale parameter, and $\mathbf{r}$ is the one-dimensional mean location parameter. For the second case where the first coefficient follows a power lognormal distribution and the second coefficient follows an exponential distribution, the same procedure as above is followed to generate the first coefficient (the power lognormal) and the third coefficient (the normal). But the second coefficient β_{q2} is developed from the normal draw $\tilde{\beta}_{q2}$ as follows: $F_2^{-1}\left[\Phi(\tilde{\beta}_{q2})\right] = -\sigma_2 \ln\left[1-\Phi(\tilde{\beta}_{q2})\right] + \mu_2$.

In both simulation settings, we allow a general covariance matrix for the kernel error term vector $\tilde{\boldsymbol{\varepsilon}}_q$ with a covariance specification for $\boldsymbol{\Theta}$ as follows:

$$\boldsymbol{\Theta} = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.500 & 0.500 \\ 0.000 & 0.500 & 1.000 & 0.600 \\ 0.000 & 0.500 & 0.600 & 1.413 \end{bmatrix}$$

$$= \mathbf{L}_{\Theta}\mathbf{L}'_{\Theta} = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.500 & 0.866 & 0.000 \\ 0.000 & 0.500 & 0.404 & 0.998 \end{bmatrix} \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.500 & 0.500 \\ 0.000 & 0.000 & 0.866 & 0.404 \\ 0.000 & 0.000 & 0.000 & 0998 \end{bmatrix}$$

Note that, as discussed in Section 3, the first row and first column are all normalized to zero, and the second diagonal element is normalized to 1 for identification. To maintain positive definiteness, we work with the Cholesky decomposition elements of $\boldsymbol{\Theta}$, with two Cholesky matrix elements to be estimated in $\mathbf{L}_{\Theta} = (l_{\Theta 5} = 0.404$ and $l_{\Theta 6} = 0.998)$.[7] Collectively,

[7] The specification for the differenced covariance matrix above may be viewed as being derived from a specification where the error terms for the first three alternatives are independent and distributed with a variance of 0.5, while the

$\text{Vech}(\mathbf{L}_{\Theta}) = (l_{\Theta 5}, l_{\Theta 6})'$. A multivariate draw of $\tilde{\boldsymbol{\varepsilon}}_q$ is obtained by drawing I multivariate normally distributed random numbers in the usual way, given that $\tilde{\boldsymbol{\varepsilon}}_q \sim \text{MVN}(0, \mathbf{IDEN}_1 \otimes \boldsymbol{\Theta})$ for the cross-sectional case.

To generate the dependent variable values in the simulation for given independent variable values for each individual (that is, for given $\mathbf{x}_q$ and $\mathbf{s}_q$ values), we track back to the matrix form of Equation (9) and write:

$$\tilde{\mathbf{U}}_q = \mathbf{x}_q \boldsymbol{\beta}_q + \mathbf{s}_q \boldsymbol{\gamma}_q + \tilde{\boldsymbol{\varepsilon}}_q .$$

Once the multivariate realizations of $\boldsymbol{\alpha}_q = (\boldsymbol{\beta}'_q, \boldsymbol{\gamma}'_q)'$ and $\tilde{\boldsymbol{\varepsilon}}_q$ are drawn, the utility of each alternative at each choice occasion is computed, and the alternative with the highest utility at each choice occasion is then identified as the chosen alternative.

The above data generation process is undertaken, for each simulation setting, 200 times with different realizations of the $\boldsymbol{\alpha}_q$ and $\tilde{\boldsymbol{\varepsilon}}_q$ vectors to generate 200 different data sets, each with 3000 choice occasions as mentioned earlier. The hybrid MSL-MACML inference approach of Equation (14) is applied to each of the 200 data sets to estimate data specific values of $\boldsymbol{\lambda}$. In this approach, MSL is used to integrate out the non-normal coefficients and for this procedure we use 50 draws per individual from the Halton sequence. The MACML approach is employed to evaluate the MVNCD function that is the integrand in Equation (14). In the MACML procedure, a single random permutation is generated for each individual (the random permutation varies across individuals, but is the same across iterations for a given individual), and the multivariate normal cumulative distribution (MVNCD) function is approximated using the resulting conditional probability sequence.

4.1. Performance Evaluation

For both simulation settings, the performance of the hybrid MSL-MACML approach in recovering parameters of the model is evaluated as follow.

(1) Estimate the parameters for the 200 datasets. Estimate the standard errors.

---

last error term has a variance of 0.913 and is correlated with the error term of the third alternative with a covariance of 0.1. In the simulation experiment estimations, to focus on the random coefficients, we fix the variances of the first three alternatives to 0.5 and impose independence among the first three alternatives, but estimate the variance of the fourth error term and the covariance between the third and fourth alternatives, which translates to the two Cholesky parameters $l_{\Theta 5} = 0.404$ and $l_{\Theta 6} = 0.998$.

(2) Compute the mean estimate for each model parameter across the 200 data sets. Compute the **absolute percentage bias** (APB) as: $APB = \left| \frac{\text{mean estimate - true value}}{\text{true value}} \right| \times 100$.

(3) Compute the standard deviation of each parameter estimate across the 200 datasets, and label this as the **finite sample standard deviation or FSSD** (essentially, this is the empirical standard error). Compute the FSSD as a percentage of the true value of each parameter.

(4) Compute the mean standard error for each model parameter across the 200 datasets, and label this as **the asymptotic standard error or ASE** (essentially this is the standard error of the distribution of the estimator as the sample size gets large, and is a theoretical approximation to the FSSD).

(5) Next, to evaluate the accuracy of the asymptotic standard error formula for the finite sample size used, compute the **absolute percentage bias of the asymptotic standard error (APBASE)** for each parameter relative to the corresponding finite sample standard deviation.

$$APBASE = \left| \frac{\text{ASE - FSSD}}{\text{FSSD}} \right| \times 100$$

4.2. Simulation Results

Summaries of the performance measures for the first and the second simulation settings are presented in Table 2 and Table 3, respectively. The tables provide the true value of the parameters, followed by the parameter estimates and the standard error estimates. Overall, the results show that the proposed method recovers parameters very well with the average of the absolute percentage bias (APB) in both cases being lower than 5% (see the last row under the APB column). Further, the asymptotic standard error from the method also quite closely reflects the finite sample standard deviation, as evident from the APBASE estimates, whose average (across all parameters) is less than 9% (see last row under the APBASE column).

Several other observations may be made from the results. In the first setting (the case in which two coefficients are assumed to be realizations from power log-normal distributions and one coefficient is assumed to be a realization from a normal distribution; see Table 2), the third copula correlation parameter $(l_{\Gamma 3})$ presents a high APB value of 18.19%. However, this result is rather deceiving because the true estimate for this parameter is 0.20 and the finite sample bias is only 0.036; that is the APB value is being inflated in percentage simply because of the small magnitude of the true value of the parameter. Interestingly, this parameter estimate also presents the highest APBASE value of the table (19.72%). In fact, the APBASE is relatively high for all the copula correlation parameters relative to other parameters, suggesting that the copula correlations are the most difficult to precisely estimate. This is not surprising, because the copula correlation parameters are the ones that occur most non-linearly in the CML function of Equation (20).

In the second simulation setting (corresponding to the three different distributions of power log-normal, exponential and normal for the coefficients; see Table 3), the copula correlation coefficient $l_{\Gamma 1}$ presents the highest APB (10.22%). This represents the correlation between the power log-normal and exponential distributions. This is to be expected, given the relatively non-linear and complicated manner in which the copula correlation enters into the optimization function for retrieving the parameters. This is also reflected in the high APBASE value (19.65%) for this copula correlation, reinforcing the notion that not only is it difficult to accurately retrieve this parameter, but so is the precision of recovery of the parameter. But it is to be noted that even these are not egregiously high biases. The other parameter showing a very high APBASE (34.48%) is that corresponding to the standard deviation of the exponential distribution (σ_2). It is indeed interesting that the two parameters ($l_{\Gamma 1}$ and σ_2) that are most difficult to recover (from an accuracy and/or precision standpoint) involve the exponential distribution. These results are a consequence of the long tail of the exponential distribution, a reason that also typically makes estimation using a traditional log-normal distribution (that also has a long tail) rather unstable and imprecise. As in the first simulation setting, we again find that the copula parameters are the ones that are the most difficult to precisely pin down.

Finally, the copula correlation parameter $l_{\Gamma 2}$ in Table 3, which represents the correlation between the power log-normal and the normal univariate marginals in the second setting, has a much smaller APB (0.20%) than $l_{\Gamma 3}$ in the first setting (Table 2), which also represents the correlation between power log-normal and normal. This result confirms that the high APB of this parameter in the first setting was due to its small magnitude and not poor recovery.

5. AN EMPIRICAL APPLICATION

In this section, we illustrate the use of the proposed model for an empirical application on a commuter mode choice dataset containing repeated choices from the same individuals. The dataset is drawn from a web-based stated preference survey from Austin. The purpose of the survey was to examine the demographic, employment, and overall travel characteristics of Austin area commuters, and to identify the possible effects on commute mode share of adding a commuter rail as a new transportation option. Four alternative modes are presented to the respondent as commuting options: drive alone, shared ride, bus and the commuter rail. Each respondent provides the mode she or he would choose to use on four repeated choice occasions, with different attribute values for each of several attributes, including travel time and travel cost. Additional details about the survey and the stated preference design can be found in Bhat (2004) and Bhat and Sardesai (2006). There are 322 individuals in the sample and a total of 1288 choice occasions. The mode share across all choice occasions is: 45.34% drive alone, 13.43% shared ride, 5.67% bus and 35.56% commuter rail. While the commuter rail share is very high, the reader will note that this is purely a stated preference survey in which commuter rail, and the shared ride and bus modes, were included by design as available options for all individuals, to maximize the information we were able to extract about the relative tradeoffs between travel time and travel cost. Besides, for the same reason, the SP choice scenarios involved an increase over the current scenario for the respondent in drive alone travel times and costs. The obvious overstatement in non-drive alone mode choice because of the SP design may be controlled for if one wants to make predictions of future modal shares, as undertaken by Bhat and Sardesai (2006). But the emphasis in this paper is on the distributions of the travel time and cost coefficients (and the resulting value of time), not on the predictions of modal shares.

5.1. Valuation of Travel Time Savings

The valuation of travel time savings (VTTS) is a central element in transportation planning and analysis. As indicated by Small (2012), “its theoretical meaning and its empirical measurement are fundamental to travel demand modeling, social cost analysis, pricing decisions, project evaluation, and the evaluation of many public policies”. Small proceeds to discuss in detail the many uses of VTTS, which we will not elaborate on here for presentation conciseness. But the important point is that, while there is general agreement that no one would be interested in wasting money on daily travel (such as commuting), there is some (at least small) possibility that individuals would want to extend their travel time on at least some travel occasions. Cirillo and Axhausen (2006) provide a conceptual justification for this in the short term, because, while pure travel time is valued negatively by individuals, there is a comingling of this (dis-) utility of travel with the potentially positive utility from secondary activities that may be undertaken during daily travel (such as being able to listen to relaxing music in the privacy of one's vehicle, or the joy of being in movement per se).

For our analysis, it suffices to note that theoretical considerations require that the cost coefficient (the denominator in the VTTS computation) should be always negative (and cannot even take the value of zero in its domain, because this causes a singularity problem in the computation of VTTS). That is, we need a bounded distribution for the cost coefficient that does not straddle the zero value. However, we do not impose this as an absolute requirement for the travel time coefficient (the numerator in the VTTS computation), especially in the short-term context of daily travel. That is, we allow for the possibility of the travel time coefficient to be unbounded, leading to potentially negative or zero VTTS values, though our expectation is that there will be a relatively small fraction of individuals for whom VTTS will not be positive.

In our estimation specifications, we considered several bounded distributions for travel cost, as well as an unbounded normal distribution and several bounded distributions for travel time. However, the power log normal distribution (with p=5) consistently came out to be the best bounded distribution in our empirical context, for both the cost and time coefficients. For completeness, in the next section, we present the results for all possible combinations of fixed, log-normal (the distribution that has been typically used in the literature for bounded distributions), and power log-normal coefficients (with p=5) for cost and time, supplemented by a possible normal distribution for travel time (but not for cost). This leads to the presentation of twelve models with different distributional combinations for the cost and time coefficients. In each of these models, the cost variable is used as cost over personal income. That is, the cost coefficient is actually a coefficient with the stipulated distribution divided by personal income. We use such a specification because it is intuitive and also because it came out to be consistently superior to the simple cost specification. In addition, we tested for a random covariance structure for the baseline constants (except for one alternative, which is the base alternative) to capture heterogeneity across individuals in modal preferences (as well as individual-level dependence in the unobserved modal preferences). The repeated nature of our data allows such an individual-level covariance structure in modal preferences in addition to the choice occasion-level heterogeneity captured by the covariance of the choice-occasion-level kernel-error terms. But we have only four choice occasions per individual, which can be inadequate to tease out a full covariance matrix capturing inter-individual intrinsic preference differences. In any case, in our analysis, this generic individual-level covariance terms consistently turned out to be statistically insignificant whether or not the time and cost coefficient heterogeneities were introduced. And the fit of the model with only the generic individual-level covariance structure was worse than the model with only heterogeneity in the cost or time coefficients. Effectively, the time and cost variables show a good amount of variation (both across individuals as well as within the choice occasions of the same individual) because of the SP design, thus allowing individual-level heterogeneity to be captured on these variables. This is also an efficient way of capturing individual-level heterogeneity, given the relatively small number of individuals in the sample. But there is simply not sufficient information it appears to pin down the many covariance terms characterizing the intrinsic individual-level heterogeneity effects.

Also to be noted here is that, as soon as we introduced any random coefficient (for the cost and/or the time coefficient), the estimated covariance matrix of the differenced error terms $\varepsilon_{qi1} = (\tilde{\varepsilon}_{qi} - \tilde{\varepsilon}_{q1})$ could not be distinguished from a matrix of ones on the diagonal and 0.5 values on the off-diagonals. That is, we could not rule out an IID covariance matrix for the original kernel error terms, and so all the models presented in the next section use an IID kernel formulation. The implication is that, in the current empirical context, any utility covariances at the individual-level or at the choice occasion-level may be structurally traced to individual-level random coefficients in the cost and/or time coefficients.

5.2. Empirical Results

In this section, we first discuss data fit and VTTS considerations, and then present the full model results for the preferred model specification.

5.2.1. Data Fit and VTTS Estimates

The third main column of Table 4 shows the composite marginal likelihood (CML) values at convergence for the twelve model specifications discussed earlier. The same variable specification was retained for all the twelve models, with the only difference being in the distributional assumptions for the cost and time coefficients (indeed, the same set of variables came out to be appropriate from a statistical significance perspective for all twelve models). The many models may be compared with each other based either on a nested test (when one model is a restricted version of the other) or a non-nested test. In Table 4, all the models with a fixed cost and/or fixed time coefficient are restricted versions of appropriate other models (for example, a model with a fixed cost coefficient and a normally distributed travel time coefficient is a restricted version of all models with a specified random distribution on the cost coefficient and a normally distributed travel time coefficient). In these cases, the restrictive models can be compared with the corresponding unrestricted models using the adjusted composite marginal likelihood ratio test (ADCLRT; see Bhat, 2014). However, for all cases of comparisons between two models with one or both random coefficients with different distributional assumptions, one needs to use a non-nested statistical test. This can be done using the composite likelihood information criterion (CLIC) introduced by Varin and Vidoni (2005) may be used. The CLIC takes the following form[8]:

$$\log L^{*}_{CML}(\hat{\boldsymbol{\theta}}) = \log L_{CML}(\hat{\boldsymbol{\theta}}) - tr\left[\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}})^{-1}\right] \tag{22}$$

The model that provides a higher value of CLIC is preferred. Technically, the CLIC statistic can also be used to compare nested models, though it has less power than the ADCLRT statistic. So, for presentation ease, in Table 4, we only show the CLIC statistic for each of the estimated models (however, each restricted model was rejected in favor of its unrestricted versions based on the ADCLRT test).

Several important observations may be made from Table 4. First, the model with fixed cost and time coefficients (Model 1) is rejected soundly relative to random coefficients on one or both of travel cost and travel time. This clearly suggests the presence of random individual-level heterogeneity in taste to cost/time. Second, models with a fixed coefficient on one of the travel time or travel cost variables (Models 2, 3, 4, 5, and 9) fare much more poorly than models with both coefficients randomly distributed (Models 6, 7, 8, 10, 11, 12). This finding supports the notion that it is not advisable to a priori fix a coefficient simply to make WTP computations easier (see

[8] This penalized log-composite likelihood is nothing but the generalization of the usual Akaike's Information Criterion (AIC). In fact, when the candidate model includes the true model in the usual maximum likelihood inference procedure, the information identity holds (*i.e.*, $\boldsymbol{H}(\boldsymbol{\theta})=\boldsymbol{J}(\boldsymbol{\theta})$) and the CLIC in this case is exactly the AIC [$= \log L_{ML}(\hat{\boldsymbol{\theta}}) -$ (# of model parameters)].

Train and Weeks, 2005, for related reasons for not fixing the cost coefficient). Third, the results show that the models with a fixed coefficient on travel time and a bounded distribution on travel cost (models 5 and 9) are superior from a data fit standpoint relative to those that specify a fixed coefficient on cost and a random coefficient on travel time (models 2, 3, and 4; note that the CLIC statistic ranges from -4320 to -4315 for the first set of models compared to the range from -4388 to -4374 for the second set of models). The implication is that there is much more individual-level heterogeneity related to cost sensitivity rather than associated with time sensitivity. Fourth, between any pair of models with the same distributional assumption for the travel time coefficient, a power-lognormal distribution for the cost coefficient does better than the traditional log-normal distribution (compare models 5 versus 9, 6 versus 10, 7 versus 11, and 8 versus 12). As indicated earlier, the proposed model allows a flexible parametric distributional form for the random coefficients. Our results suggest that researchers may want to try the power-lognormal as an alternative to the log-normal, especially given potential convergence problems originating from the long tail of the log-normal. Fifth, the model with a power log-normal cost coefficient distribution and a normal time coefficient distribution (model 10) provides the best data fit, but also implies that, for some individuals, there is a positive valuation of travel time, leading to a negative VTTS value (please see additional discussion later). Based on the mean and standard deviation of the normal travel time coefficient, 29% of individuals are predicted to have a negative VTTS (this is, interestingly, in the same range as that obtained by Cirillo and Axhausen, 2006).

Table 4, in addition to providing data fit measures, also provides median VTTS estimates for three annual personal income categories: low income (US $15,000, the minimum value in the sample), medium income (US $50,000, the median value), and high income (US $150,000, a high income value). In the table, we provide the median VTTS estimates because it is a better central measure to compare across the models. The VTTS median estimate is computed by drawing 20,000 realizations from the bivariate copula distribution of the time and cost coefficients, computing the implied VTTS for each bivariate realization by taking the ratio of the time to cost draws, and then computing the median value across the 20,000 realizations (for presentation efficiency, we will refer to the median estimate as the VTTS estimate from hereon), Again, many observations stand out from the VTTS estimates. First, and as expected, for every model, the VTTS increases proportionally with income, which is a result of the "cost over income" specification in the models. Second, for each of the three income values, the VTTS estimate from the model in which both the cost and time coefficients are fixed (model 1) and the model in which the cost coefficient is fixed and the time coefficient is normally distributed (model 2) produce the highest estimates. These VTTS estimates are higher than the implied wage rate for each income category (based on full-time work with 52 weeks and 40 hours per week, the wage rates for the low, medium, and high income categories are \$7.2 per hour, \$24 per hour and \$72 per hour). These certainly seem out of the realm of reasonableness. On the other hand, the lowest VTTS values are obtained in the models that involve a lognormally or power-lognormally distributed cost coefficient and a random time coefficient (models 6, 7, 8, 10, 11, and 12). In these models, the VTTS values are about one-third of the wage rate. These models also have a superior data fit relative to other models. The VTTS estimates for the models with one fixed coefficient and the other being randomly distributed (models 3, 4, 5, and 9) lie somewhere in-between, with an implied value of about 63% of the wage rate. <u>Third</u>, a further exploration of the VTTS distributions (rather than simply the median VTTS estimates) reveals that there are differences in the distributions even between models providing similar median VTTS estimates. Thus, among models 6, 7, 8, 10, 11, and 12, which all use a log-normally or power-lognormally distributed cost coefficient with a randomly distributed time coefficient, those that use a log-normal distribution for one coefficient and a power-lognormal distribution for the other (models 7, 8 and 11) belong to one group (labeled Group 1) with a sharp spike in the VTTS distribution. This is shown in Figure 2, where, to avoid clutter, we show the VTTS distribution only for model 11 as the representative model for this group (model 11 has the best data fit in Group 1). The sharp spike for this group is, of course, a manifestation of the log-normal distribution used for one of the two coefficients. Also, because of the strictly bounded nature of the distribution for both the cost and time coefficients, we get only positive VTTS values. On the other hand, models 6 and 10, which use a normal travel time coefficient and a log-normal or power log-normal cost coefficient, also have similar VTTS profiles, but that are very different from Group 1. In Figure 2, we show the VTTS profile for model 10 as the representative model for Group 2 comprising models 6 and 10. As should be obvious, this group allows negative VTTS values (as discussed earlier, of the order of 29% of the distribution) and also has the lowest spike. Similarly, there are also VTTS distribution differences among the four models with one fixed coefficient and the other being randomly distributed (models 3, 4, 5, and 9). Specifically, the VTTS profiles for models 3, 5, and 9 are similar, with that of model 9 shown in Figure 1 as the representative of this Group 3 set of models. Group 3 is identified by a spike between that of Groups 1 and 2, and the longest right tail of all groups. Finally, models 4 and 12 also have similar profiles, but are also sufficiently different to be shown separately in Figure 2. While model 12 uses a power lognormal distribution for both the cost and the time coefficients, model 4 uses a power lognormal only for the time coefficient with a fixed coefficient on cost (rendering the VTTS to be power lognormal). The VTTS profiles of both models 4 and 12 start off similarly on the left edge with a spike of the same order of magnitude, but then the profile for model 4 moves more toward that of Group 3 with the long tail. This leads to the much higher median VTTS value from Model 4 compared to model 12. Overall, while different reasonable analysts can come to different conclusions, we believe that Model 12 represents the best combination of data fit, median VTTS value as a percentage of wage rate, and the shape of the VTTS profile. While models 6 and 10 provide a better data fit, the percentage predicted to have a negative VTTS is just too high in our opinion, as is the implied very high variance across individuals in their VTTSs. [9]

Overall, the results indicate that there needs to be much more emphasis in the literature on VTTS profiles, rather than simply statistics of the VTTS values such as means, medians, and standard deviations. The profiles, which play a critical role in consumer welfare analyses, can be very different even when the imputed VTTS mean or median values are similar. Of course, the only way that different profiles can be considered is by allowing flexible marginal non-normal distributions on specific individual coefficients, which is precisely what our proposed copula model enables the analyst to estimate using a convenient and practically feasible hybrid MACML-MSL inference technique.

[9] There has been a healthy discussion and debate in the literature (see, for example, Ory and Mokhtarian, 2005; Cirillo and Axhausen, 2006) on the issue of whether or not some individuals associate a positive valuation to travel time as opposed to the predominantly held view that people are averse to higher travel times. Of course, there is also the issue that this may be very context dependent, including, for example, the length of the travel time being considered (see, for example, Pinjari and Bhat (2006), who suggest that the sensitivity to travel time is non-linear over travel time). In this paper, we do not engage in this line of debate. The purpose here is to present, and demonstrate an application of, a flexible copula model and its estimation that can be gainfully employed to estimate different combinations of multivariate random coefficient distributions to then guide the final model structure and specification, based on theoretical considerations (for example, which coefficients should have bounded distributions and which can have unbounded distributions), intuitive considerations (the reasonableness of trade-off values obtained and their profiles over the population), and statistical data fit considerations.

5.2.2. Estimation Results for the Preferred Model (Model 12)

For completeness, Table 5 presents the estimation results for the preferred model, which are generally consistent with the vast literature now on commute travel mode choice (see, for example, Bhat and Sardesai, 2006; Paleti *et al*., 2013; Ho and Mulley, 2015; Wang, 2015). The alternative-specific constants in the first row panel do not have any substantive interpretations; they simply control for the sample values of the exogenous variables and the sample shares, though the overall negative signs on all the non-drive alone modes are consistent with the high mode share of the drive alone alternative in the sample. Individuals who earn a higher share of total household income are less likely to use the currently available non-solo auto modes (share-ride and bus) relative to those who earn a lower share of total household income, suggesting that those who wield more market power in the household have "first choice rights" over modes that are viewed as flexible, fast, and comfortable. Also, non-work activity stops made during the mid-day and/or during the commute encourage the use of the car mode (either drive alone or shared-ride). Women are less likely to commute by bus relative to men, though this variable is statistically significant at only the 83% confidence level. Finally, the table provides the parameters for the travel cost and travel time distributions, which formed the basis for much of the discussion in the earlier section. A point that should be noted here is that the copula parameter came out to be statistically insignificant in the current empirical analysis. But the situation could be different in other empirical contexts. In any case, the only way to test the presence and intensity of the copula parameter is to estimate the multivariate Gaussian copula model introduced in this paper

6. SUMMARY AND CONCLUSIONS

In this paper, we propose a mixed multinomial probit model that is able to accommodate a general covariance structure for the kernel error terms as well as a very flexible continuous parametric multivariate structure for unobserved individual heterogeneity. The latter is introduced using a Gaussian copula approach that ties different continuous univariate mixing distributions into a joint multivariate distribution. The individual univariate mixing distributions can be bounded or unbounded, allowing the incorporation of theoretical considerations that require specific coefficients to span only the half-line. In addition, our proposed approach includes the case of independence across specific coefficients, allows a flexible and wide range of dependence across coefficients, and is easy to work with. The estimation of the model is achieved using a combination of the maximum simulated likelihood (MSL) technique (to accommodate the non-normal random coefficients) and Bhat's MACML inference approach (to accommodate all the normal random coefficients as well as the kernel normal error structure; see Bhat, 2011 and Bhat, 2014). To our knowledge, this is the first time that a copula-based mixed MNP model has been proposed in the literature, along with an associated hybrid MSL-MACML inference approach that is ideally suited for the case when there are few non-randomly distributed coefficients (so that the MSL simulation does not involve very high dimensions) and many normally distributed coefficients (so that the MACML computational accuracy and efficiency can be realized). For the non-normal coefficients, the use of univariate distributions that have a closed-form inverse function facilitates quick estimation. Of these, we would particularly like to highlight our consideration of the power log-normal distribution that has not been considered earlier in discrete choice models. The advantage of this distribution relative to other distributions on the half-line (including the log-normal) is that it can both allow for substantial heterogeneity (large variance parameter) and also ensure that the skewed tail is relatively thin, which helps convergence.

We demonstrate the effectiveness of our inference approach through simulation exercises as well as an empirical application. The simulations involve cross-sectional choice data with a sample size of 3000, and two configurations of three random coefficients. The first includes two power log-normal coefficients and one normal coefficient, while the second considers one each of power log-normal, exponential and normal coefficients. Overall, the simulation results indicate that the proposed method allows for accurate parameter recovery. Further, the asymptotic standard errors from the method also quite closely reflect the finite sample standard deviations. One finding, however, is that it appears to be more difficult to recover the copula parameters characterizing the dependence between pairs of univariate margins, especially between pairs of non-normal univariate margins. Also, the simulation results suggest that distributions with very long tails (such as the exponential and lognormal) make it particularly difficult to recover variance parameters and corresponding copula parameters of dependence with other margins. However, even in these cases, the method performs quite well. Future simulation studies should study the performance of the proposed method in more detail, in relation to varying sample sizes, both cross-sectional and repeated choice data, different degrees of copula dependence, an array of different bounded and unbounded univariate margins, and varying numbers of random coefficients.

The empirical application focuses on a repeated choice commute travel mode stated preference data collected in Austin, Texas. The results reiterate the importance of the power lognormal distribution as a strong contender (and alternative) to the traditional lognormal distribution and other bounded distributions for the travel cost coefficient. Additionally, the preferred model with a power lognormal distribution for the cost coefficient (with p=5) and a power lognormal distribution for the time coefficient yields a median value of travel time savings that is about a third of the wage rate. Further our results do suggest there is much more individual-level heterogeneity related to cost sensitivity rather than associated with time sensitivity.

Of course, the emphasis of this paper has been on presenting a new copula-based discrete choice model structure and an associated inference approach. Much still needs to be done in terms of investigating ways to obtain good starting parameters for the copula model with different marginal distributions, and develop structured optimization algorithms for the quick estimation of models with power lognormal and other margins (for example, is there a better way to optimize rather than fix p values and estimate different specifications repeatedly, which can become cumbersome when there are many random coefficients). Besides, additional research needs to compare the performance and effectiveness of the proposed copula-based model with other non-parametric ways to introduce taste heterogeneity. There is also room for testing different distributional assumptions that were not included in this study. For example, future research should test the use of log-uniform and log-triangle distributions that, similar to the power log-normal, are bounded above zero and have thin tails that allow for heterogeneity while facilitating convergence. We hope that this new flexible parametric approach will offer researchers and practitioners another way of accommodating heterogeneity in a general and efficient manner in choice models, and open up a new stream of empirical applications with bounded and non-normal distributions.

ACKNOWLEDGMENTS

This research was partially supported by the U.S. Department of Transportation through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center. The first author would like to acknowledge support from a Humboldt Research Award from the Alexander von Humboldt Foundation, Germany. The authors are grateful to Lisa Macias for her help in formatting this document, and to two anonymous referees who provided useful comments on an earlier version of the paper.

REFERENCES

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

Azzalini, A., 2013. *The Skew-normal and Related Families (Vol. 3)*. Cambridge University Press, Cambridge, U.K.

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 nonparametric random utility models with an application to the value of time in heterogeneous populations. *Transportation Science* 44(4), pp.537-549.

Berry, S.T., Haile, P.A., 2014. Identification in differentiated products markets using market level data. *Econometrica* 82(5), 1749-1797.

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

Bhat, C.R., 2004. Austin commuter survey: findings and recommendations. Technical Report, Department of Civil, Architectural & Environmental Engineering, The University of Texas at Austin. http://www.ce.utexas.edu/prof/bhat/reports/austin_commuter_survey_report.doc

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., 2014. The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models. *Foundations and Trends in Econometrics* 7(1), Now Publishers Inc.

Bhat, C.R., Eluru N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, 43(7), 749-765.

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., Sardesai R., 2006. The impact of stop-making and travel time reliability on commute mode choice. *Transportation Research Part B* 40(9), 709-730.

Bhat, C.R., Sidharthan R., 2012. A new approach to specify and estimate non-normally mixed multinomial probit models. *Transportation Research Part B* 46(7), 817-833.

Bhat, C.R., Dubey, S.K., Nagel, K., 2015. Introducing non-normality of latent psychological constructs in choice modeling with an application to bicyclist route choice. *Transportation Research Part B* 78, 341-363.

Bhat, C.R., Sener I.N., Eluru N., 2010. A flexible spatially dependent discrete choice model: Formulation and application to teenagers' weekday recreational activity participation. *Transportation Research Part B* 44(8-9), 903-921.

Capitanio, A., 2010. On the approximation of the tail probability of the scalar skew-normal distribution. *Metron* 68(3), 299-308.

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.

Cirillo, C., Axhausen, K.W., 2006. Evidence on the distribution of values of travel time savings from a six-week diary. *Transportation Research Part A* 40(5), 444-457.

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

Godambe, V.P., 1960. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* 31(4), 1208-1211.

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

Ho, C., Mulley, C., 2015. Intra-household interactions in tour-based mode choice: The role of social, temporal, spatial and resource constraints. *Transport Policy* 38, 52-63.

il Kim, K., 2014. Identification of the distribution of random coefficients in static and dynamic discrete choice models. *The Korean Economic Review* 30(2), 191-216.

Joe, H., 2015. *Dependence Modeling with Copulas*. CRC Press, Taylor and Francis, Boca Raton, FL.

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

Luce, R.D., Suppes, P., 1965. *Preference, utility, and subjective probability*. In Luce, R.D., Bush R.R., Galanter, E.H. (eds.), *Handbook of Mathematical Psychology*, Vol. 3, 249-410, Wiley, New York.

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

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

Mittelhammer, R.C., Judge, G., 2011. A family of empirical likelihood functions and estimators for the binary response model. *Journal of Econometrics* 164(2), 207-217.

Nelsen, R.B., 2006. *An Introduction to Copulas*. 2nd Edition, Springer, New York.

Paleti, R., Bhat C.R., 2013. The composite marginal likelihood (CML) estimation of panel ordered-response models. *Journal of Choice Modelling* 7, 24-43.

Paleti, R., Bhat, C., Pendyala, R., 2013. Integrated model of residential location, work location, vehicle ownership, and commute tour characteristics. *Transportation Research Record: Journal of the Transportation Research Board* 2382, 162-172.

Pinjari, A. and Bhat, C., 2006. Nonlinearity of response to level-of-service variables in travel mode choice models. *Transportation Research Record: Journal of the Transportation Research Board* 1977, 67-74.

Revelt, D., Train, K., 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics* 80(4), 647-657.

Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris, 8, 229-231.

Sklar, A., 1973. Random variables, joint distribution functions, and copulas. *Kybernetika* 9(6), 449-460.

Small, K.A., 2012. Valuation of travel time. *Economics of Transportation* 1(1), 2-14.

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

Train, K., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In Scarpa, R., Alberini, A. (eds.), *Applications of Simulation Methods in Environmental and Resource Economics*, Ch. 1, 1-16, Springer, Dordrecht, The Netherlands.

Trivedi, P.K., Zimmer, D.M., 2007. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics* 1(1), Now Publishers Inc.

Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. *Biometrika* 92(3), 519-528.

Wang, R., 2015. The stops made by commuters: evidence from the 2009 US National Household Travel Survey. *Journal of Transport Geography* 47, 109-118.

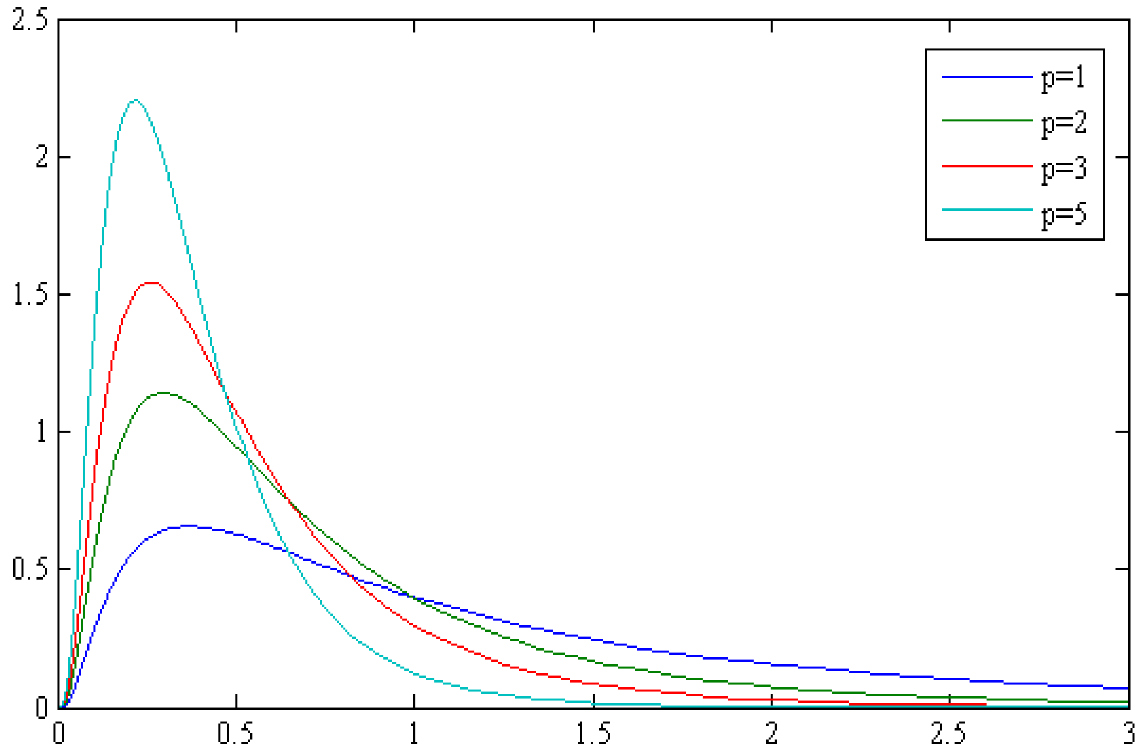
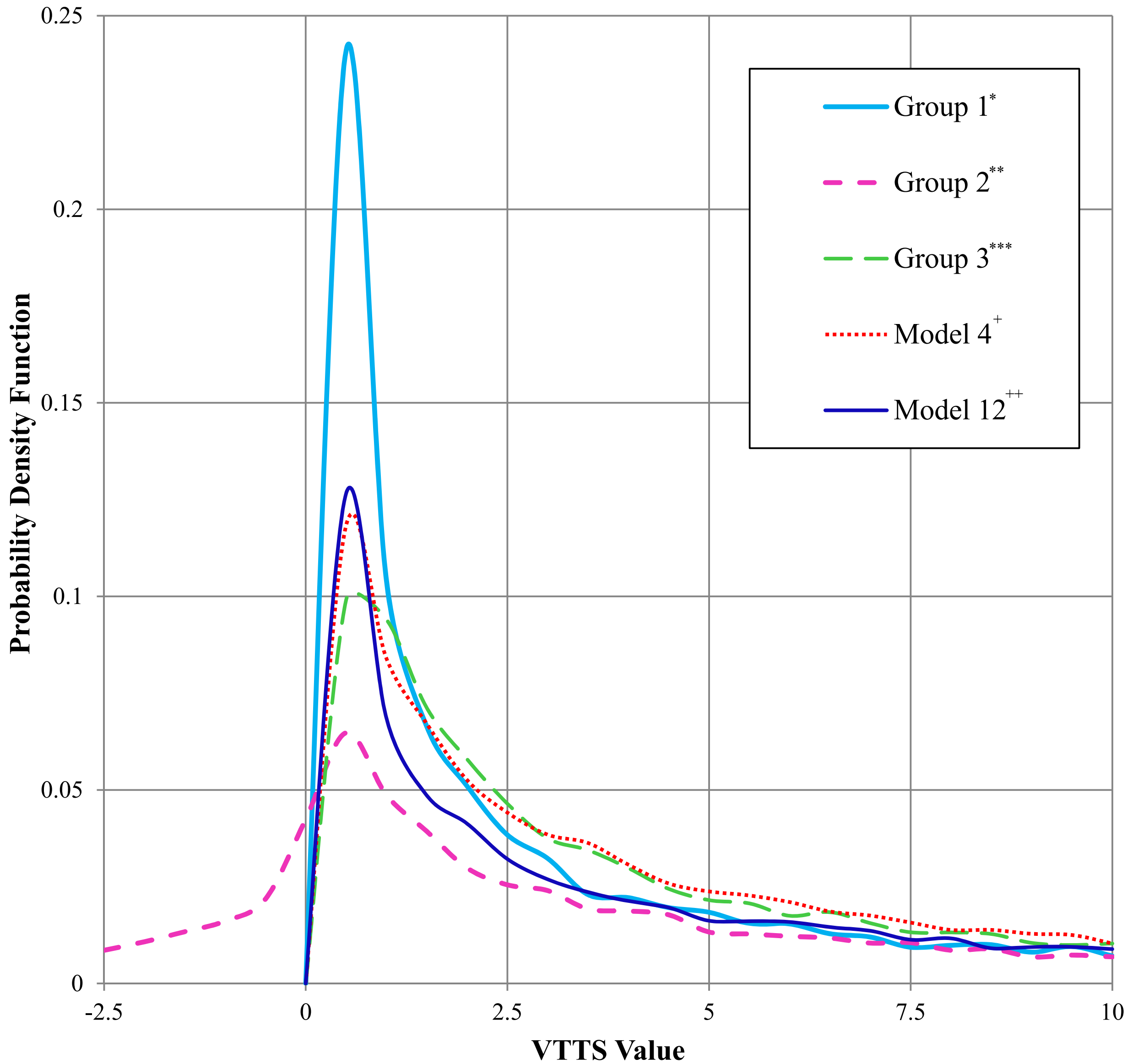


Figure 1: Comparison of the log-normal (*p*=1) and the power log-normal distributions for identical values of μ and σ (μ=0 and σ=1)



*	Includes	Model 7-	Log-normal cost coefficient, Log-normal time coefficient
		Model 8-	Log-normal cost coefficient, Power log-normal time coefficient
		Model 11-	Power log-normal cost coefficient, Log-normal time coefficient
**	Includes	Model 6-	Log-normal time coefficient, Normal time coefficient
		Model 10-	Power log-normal cost coefficient, Normal time coefficient.
***	Includes	Model 3-	Fixed cost coefficient, Log-normal time coefficient
		Model 5-	Log-normal cost coefficient, Fixed time coefficient
		Model 9-	Power log-normal cost coefficient, Fixed time coefficient
+	Includes	Model 4-	Fixed cost coefficient, Power log-normal time coefficient
++	Includes	Model 12-	Power log-normal cost coefficient, Power log-normal time coefficient

Figure 2: Resulting VTTS distributions for different groups of models that have random coefficents for cost or time, or both.

Table 1: Sample distributions with closed-form inverse cumulative distribution functions and that are bounded on the half-line

Distribution Name	Density Function $f_{\beta_{qe}}(Z_e) = \text{Prob}[\beta_{qe} = Z_e]$	Cumulative Distribution Function $F_{\beta_{qe}}(Z_e) = \text{Prob}[\beta_{qe} < Z_e]$	Inverse CDF $F^{-1}_{\beta_{qe}}(g_e)$	General Notes
Exponential	$\frac{1}{\sigma} e^{-\left(\frac{Z_e - \mu}{\sigma}\right)}$	$1 - e^{-\left(\frac{Z_e - \mu}{\sigma}\right)}$	$-\sigma \ln(1 - g_e) + \mu$	$z_e \geq 0, \sigma > 0, \mu \geq 0$ Mean $= \sigma + \mu$, Median $= \sigma \ln(2) + \mu$, Mode $= \mu$, Range: μ to ∞, Std. Dev $= \sigma$, All inverse moments exist if $\mu > 0$ No inverse moments exist if $\mu = 0$
Rayleigh	$\left(\frac{Z_e - \mu}{\sigma^2}\right) e^{-\left[\frac{1}{2}\left(\frac{Z_e - \mu}{\sigma}\right)^2\right]}$	$1 - e^{-\left[\frac{1}{2}\left(\frac{Z_e - \mu}{\sigma}\right)^2\right]}$	$\sigma\sqrt{-2\ln(1 - g_e)} + \mu$	$z_e \geq 0, \sigma > 0, \mu \geq 0$ Mean $= \sigma\sqrt{\frac{\pi}{2}} + \mu$, Median $= \sigma\sqrt{2\ln(2)} + \mu$, Mode $= \sigma + \mu$, Range: μ to ∞, Std. Dev $= \sigma\sqrt{\frac{4 - \pi}{2}}$, All inverse moments exist if $\mu > 0$ No inverse moments exist if $\mu = 0$
Weibull	$\left(\frac{\gamma}{\alpha}\right)\left(\frac{Z_e - \mu}{\alpha}\right)^{\gamma - 1} e^{-\left[\left(\frac{Z_e - \mu}{\alpha}\right)^{\gamma}\right]}$	$1 - e^{-\left[\left(\frac{Z_e - \mu}{\alpha}\right)^{\gamma}\right]}$	$\alpha[-\ln(1 - g_e)]^{1/\gamma} + \mu$	$z_e \geq 0, \sigma > 0, \gamma > 0, \mu \geq 0$ Mean $= \sigma\Gamma(\gamma^{-1} + 1) + \mu$, Median $= \sigma[\ln(2)]^{1/\gamma} + \mu$, Mode $= \begin{cases} = \mu & if\ 0 < \gamma \leq 1 \\ = \alpha[(1 - \gamma^{-1})]^{1/\gamma} + \mu & if\ \gamma > 1 \end{cases}$, Range: μ to ∞, Std. Dev $= \sigma[\Gamma(1 + 2\gamma^{-1}) - \{\Gamma(1 + \gamma^{-1})\}^2]$, $\Gamma(a) = \int_{t=0}^{\infty} t^{a-1} e^{-t} dt$ All inverse moments exist if $\mu > 0$ Inverse k^{th} inverse moments exist if $\mu = 0$ and $\gamma > k$ If $\gamma = 1$, Weibull collapses to exponential If $\gamma = 2$, Weibull collapses to Rayleigh with $\alpha = \sqrt{2}\sigma$

Distribution Name	Density Function $f_{\beta_{qe}}(Z_e) = \text{Prob}[\beta_{qe} = Z_e]$	Cumulative Distribution Function $F_{\beta_{qe}}(Z_e) = \text{Prob}[\beta_{qe} < Z_e]$	Inverse CDF $F^{-1}_{\beta_{qe}}(g_e)$	General Notes
Log-Normal	$\frac{1}{Z_e\sigma}\,\phi\left(\frac{\ln Z_e - \mu}{\sigma}\right)$	$\Phi\left(\frac{\ln Z_e - \mu}{\sigma}\right)$	$e^{[\sigma\phi^{-1}(g_e)+\mu]}$	$z_e \geq 0, \sigma > 0$ Mean $= e^{\left(\mu+\frac{1}{2}\sigma^2\right)}$, Median $= e^{\mu}$ Mode $= e^{\mu-\sigma^2}$, Range: Strictly positive Real line, Std. Dev $= e^{\mu}\sqrt{e^{\sigma^2}\left(e^{\sigma^2}-1\right)}$, All inverse moments exist
Power Log-Normal	$\left(\frac{p}{Z_e\sigma}\right)\phi\left(\frac{\ln Z_e - \mu}{\sigma}\right)\left\{\Phi\left[-\left(\frac{\ln Z_e - \mu}{\sigma}\right)\right]\right\}^{p-1}$	$1-\left\{\Phi\left[-\left(\frac{\ln Z_e - \mu}{\sigma}\right)\right]\right\}^{p}$	$e^{\left[-\sigma\Phi^{-1}\left[(1-g_e)^{1/p}\right]+\mu\right]}$	$z_e \geq 0, \sigma > 0, p > 0$ Mean $= \int_0^1 e^{\left[-\sigma\Phi^{-1}\left(y^{1/p}\right)+\mu\right]}dy$, Median $= e^{\left[-\sigma\Phi^{-1}\{0.5^{1/p}\}+\mu\right]}$ Mode is solution to: $1+\left(\frac{\ln Z_e - \mu}{\sigma^2}\right) + \left(\frac{p-1}{\sigma}\right)\phi\left(\frac{\ln Z_e - \mu}{\sigma}\right)\left[\Phi\left\{-\left(\frac{\ln Z_e - \mu}{\sigma}\right)\right\}\right]^{-1} = 0$ Range: Strictly positive Real line, Std. Dev $= \sqrt{\left[\left\{\int_0^1 e^{\left[-2\sigma\Phi^{-1}\left(y^{1/p}\right)+\mu\right]}dy\right\} - \text{Mean}^2\right]}$, If $p = 1$, power lognormal collapses to lognormal

Table 2: Simulation results for 200 samples of 3000 observations: Two power log-normal and one normal random parameter

Parameter		True value	Parameter Estimates		Standard Error		
			Mean Estimate	APB	ASE	FSSD	APBASE
Power log-normal	μ_1	0.500	0.475	5.09%	0.163	0.164	0.92%
	μ_2	0.500	0.467	6.53%	0.165	0.164	1.02%
	σ_1	1.000	0.986	1.38%	0.178	0.171	4.51%
	σ_2	1.000	0.972	2.82%	0.184	0.169	8.54%
Normal	r	0.500	0.491	1.75%	0.070	0.065	7.29%
	η	1.500	1.495	0.37%	0.139	0.128	8.63%
Copula Correlation	$l_{\Gamma 1}$	0.600	0.592	1.34%	0.184	0.160	15.44%
	$l_{\Gamma 2}$	0.400	0.381	4.63%	0.111	0.129	14.07%
	$l_{\Gamma 3}$	0.200	0.236	18.19%	0.158	0.132	19.72%
Kernel Covariance	$l_{\Theta 5}$	0.404	0.413	2.29%	0.149	0.168	11.27%
	$l_{\Theta 6}$	0.998	0.980	1.80%	0.101	0.094	7.42%
Overall Average		-	-	4.20%	0.146	0.140	8.98%

Table 3: Simulation results for 200 samples of 3000 observations: One power log-normal, one exponential and one normal random parameter

Parameter		True value	Parameter Estimates		Standard Error		
			Mean Estimate	APB	ASE	FSSD	APBASE
Power log-normal	μ_1	0.500	0.502	0.45%	0.156	0.152	2.82%
	σ_1	1.000	1.000	0.00%	0.167	0.162	3.29%
Exponential	μ_2	1.000	1.013	1.28%	0.119	0.126	6.09%
	σ_2	0.800	0.792	1.05%	0.281	0.209	34.48%
Normal	r	0.500	0.498	0.47%	0.074	0.070	5.95%
	η	1.500	1.504	0.26%	0.138	0.139	0.25%
Copula Correlation	$l_{\Gamma 1}$	0.400	0.441	10.22%	0.213	0.178	19.65%
	$l_{\Gamma 2}$	0.600	0.604	0.60%	0.095	0.101	6.18%
	$l_{\Gamma 3}$	0.393	0.394	0.20%	0.155	0.141	9.99%
Kernel Covariance	$l_{\Theta 5}$	0.404	0.393	2.66%	0.164	0.165	0.64%
	$l_{\Theta 6}$	0.998	0.973	2.54%	0.106	0.100	5.99%
Overall Average		-	-	1.80%	0.152	0.140	8.67%

Table 4: Data fit and Value of Travel Time Savings (VTTS) for different models

Model Number	Coefficient distribution		Composite marginal log-likelihood (CML) value at convergence	CLIC statistic	Median VTTS Value (computed by simulation using 20,000 bivariate simulation realizations)		
	Cost[1]	Time			Low annual income (US $15,000)	Medium annual income (US $50,000)	High annual income (US $150,000)
1	Fixed	Fixed	-4390.02	-4420.11	8.72	29.08	87.24
2	Fixed	Normal	-4355.96	-4387.57	10.47	34.89	104.67
3	Fixed	Log-normal	-4365.14	-4376.27	4.68	15.59	46.78
4	Fixed	Power log-normal[2]	-4364.73	-4375.73	4.56	15.19	45.59
5	Log-normal	Fixed	-4319.15	-4330.32	4.47	14.89	44.68
6	Log-normal	Normal	-4281.47	-4294.56	2.91	9.79	29.14
7	Log-normal	Log-normal	-4300.15	-4313.22	2.44	8.15	24.45
8	Log-normal	Power log-normal	-4299.58	-4312.79	2.32	7.73	23.18
9	Power log-normal	Fixed	-4315.64	-4326.65	4.59	15.31	45.92
10	Power log-normal	Normal	-4277.88	-4290.96	2.47	8.23	24.69
11	Power log-normal	Log-normal	-4296.69	-4309.71	2.48	8.28	24.85
12	Power log-normal	Power log-normal	-4296.08	-4309.09	2.30	7.68	23.04

[1] The cost variable is introduced in all specifications as cost/personal income.
[2] For all Power log-normal distributions p=5.

Table 5: Empirical results of model with power lognormal cost coefficient and power lognormal time coefficient

(coefficients provide the impact of variables on the utility of alternatives)

Variable name	Parameter Estimates	
	Coef.	t-stat
Alternative specific constants		
Shared ride	-0.595	-9.67
Bus	-1.160	-10.35
Commuter rail	-0.174	-4.16
Personal income divided by household income		
Shared ride/Bus	-0.263	-3.24
Individual makes non-work mid-day stops		
Drive alone	0.072	2.51
Individual makes commute stops		
Bus	-0.431	-3.83
Commuter rail	-0.119	-2.70
Female		
Bus	-0.130	-1.38
Level of service variables		
Trip cost (dollars) divided by personal annual income (dollars divided by 10^5)		
Mean	1.941	8.16
Standard deviation	2.802	26.98
Travel time (hours)		
Mean	2.096	6.29
Standard deviation	2.570	13.16
Copula Correlation	0.120	0.48