On the Almost Exact-Equivalence of the Radial and Spherical Unconstrained Cholesky-Based Parameterization Methods for Correlation Matrices

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
Tel: 1-512-471-4535; Email: bhat@mail.utexas.edu
and
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Aupal Mondal
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Email: aupal.mondal@utexas.edu

ABSTRACT

Many multivariate modeling systems require a correlation matrix to be estimated. A requirement for these correlation matrices is that they be positive-definite, which is commonly achieved through a Cholesky parameterization of the correlation matrix. However, to conform to the unit diagonal vector in the correlation matrix, an additional unconstrained parameterization of the Cholesky elements itself is needed. The most common and well-established procedure is to use a spherical parameterization. More recently, van Oest (2021) suggested an alternative radial parameterization, suggesting that the radial parameterization may be more effective than the spherical parameterization. In this paper, we show that the two parameterization, radial and spherical, are very closely related, and, in fact, collapse to effectively the same parameterization after the adjustment of a scale parameter embedded in a logit function of both parameterizations. We confirm our theoretical position through a simulation experiment with a five-variate binary model system, and suggest that the optimal scale (from a convergence and computation speed standpoint) appears to be between the implicit scales embedded in the radial and spherical parameterizations.

Keywords: Correlation matrix, unconstrained optimization, spherical parameterization, radial parameterization, Cholesky decomposition, simulation.

1. INTRODUCTION

An important consideration in multivariate models (or even in univariate models with multiple random coefficients on exogenous variables with correlations across the coefficients) is the estimation of a positive-definite covariance matrix. This covariance matrix, along with other model parameters, are estimated using Bayesian or frequentist methods, and will generally involve optimization methods that extensively search over a large parameter space. While the constraints imposed by the need for positive-definiteness may be implemented using a constrained optimization procedure, such procedures become very cumbersome, unwieldy, and can often lead to numerical convergence problems (Pinheiro and Bates, 1996; Ferdous et al., 2010). Thus, it is almost universal practice to adopt unconstrained parameterizations of the Cholesky matrix for econometric and statistical model estimation. Unfortunately, while many such unconstrained parameterizations have been suggested and implemented for the estimation of *covariance* matrices (see, for example, Pinheiro and Bates, 1996 and McNeish and Bauer, 2020 for reviews), these do not immediately extend to the case of an unconstrained parameterization for the *correlation* matrix (because the procedures for the covariance matrix guarantee positive definiteness, but do not recognize the need for unit diagonal elements of the correlation matrix). At the same time, there are many situations where a correlation matrix needs to be estimated, such as in the estimation of a multivariate binary choice model system or a multivariate ordered response system (Dias et al., 2020; Mondal et al., 2020), where the scales of the latent variables underlying the discrete outcomes have to be normalized.[1] Besides, as highlighted by Barnard et al. (2000) and McNeish and Bauer (2020), even if the focus is on estimating a covariance matrix, there may be many contexts where breaking down the covariance matrix into a scale matrix and a correlation matrix may be advantageous from a conceptual, specification, and computational standpoint.

The unconstrained estimation of correlation matrices has seen renewed attention recently (see, for example, Bhat and Lavieri, 2018, Forrester and Zhang, 2020, and van Oest, 2021). The precise unconstrained parameterization adopted to ensure that the correlation matrix is positive definite in such applications has varied from simple one-level Cholesky decomposition schemes that write the Cholesky elements in a form conforming to the unit diagonal vector in the correlation matrix to specific multi-level Cholesky parameterization schemes. The problem with the first one-level decomposition scheme (see Srinivasan and Bhat, 2005) is that the estimation can break down, unless the code imposes a steep penalty if any of the diagonal Cholesky elements turn out to be complex (imaginary) during the search process. While a reasonable strategy, such estimations also require a good bit of hand-holding during estimation to construct a "nearest" valid correlation matrix (for example, by replacing the negative eigenvalue components in the correlation matrix

[1]Admittedly, the scale normalization itself in such models can be undertaken in one of two ways. The first is to normalize the error term scale (leading to the case of correlation matrices being estimated rather than covariance matrices), and the second is to normalize a coefficient affecting the latent variable while estimating a covariance matrix. The second approach may seem more attractive, because it does not involve additional restrictions on the covariance matrix, but for many multivariate applications that we have encountered, the first approach leads to far fewer cases of numerical instability than the latter, because it directly fixes the overall scale of the conditional unobserved error terms to be congruent across the dependent outcomes. The effectiveness of such scale congruency has been discussed in other model estimation contexts too (see Kohli et al., 2019 and McNeish and Bauer, 2020).

with a small positive value, or by adding a sufficiently high positive value to the diagonals of a matrix and normalizing to obtain a correlation matrix; see Rebonato and Jäckel, 2000, Higham, 2002, and Schöttle and Werner, 2004 for detailed discussions of these and other adjusting schemes; a review of these techniques is beyond the scope of this paper). And even then, there is no guarantee that the correlation matrix at "convergence" will be positive definite.

In the second set of multi-level decomposition schemes, three methods have been proposed in the literature, all of which have a common second-level parameterization for the Cholesky elements but differ in the third-level of parameterization. These three methods are (1) the partial correlations method (Joe, 2006), (2) the spherical parameterization method (Pinheiro and Bates, 1996 and Rebonato and Jäckel, 2000), and (3) the radial parameterization method (van Oest, 2021). Of these three methods, the second spherical parameterization method has seen the most use because of its ease and interpretability (see Madar, 2015, Pourahmadi and Wang, 2015, and Tsay and Pourahmadi, 2017). Besides, Forrester and Zhang (2020) have recently shown how the partial and the spherical parameterization methods are closely related. Thus, we will focus on the spherical parameterization method and not the partial correlation method in this paper. The third radial parameterization method was recently suggested and documented by van Oest (2021), and is presented as an easy and convenient alternative to the spherical parameterization method. However, as we show in this paper, the spherical and radial parameterizations are literally (even if not exactly) identical to each other, except for a scaling that effectively gets applied in implementation. Based on this insight, rather than focusing on the spherical and the radial parameterizations as distinct methods per se, we submit that the real question is whether scaling can improve the performance of the parameterization (whether based on the spherical or based on the radial parameterization) in terms of convergence stability and computational speed.

In this paper, we discuss the basics of the spherical parameterization scheme and the radial parameterization scheme, clearly laying out the common aspects of these parameterizations. We then demonstrate the essential similarity of these two parameterization schemes subject to a scaling factor, and compare the performance of using alternative scaling factors in a simulation study of a multivariate binary response model. Issues of convergence (or not), computation time, as well as accuracy and precision in recovering parameters are examined.

2. ALTERNATIVE DECOMPOSITION PROCEDURES

Consider an $M \times M$ correlation matrix R as follows, with $r_{i,j} = r_{j,i}$ and $-1 < r_{i,j} < 1 \; \forall \, i \neq j$:

$$R = \begin{pmatrix} 1 & r_{1,2} & r_{1,3} & \cdots & r_{1,M-1} & r_{1,M} \\ r_{2,1} & 1 & r_{2,3} & \cdots & r_{2,M-1} & r_{2,M} \\ r_{3,1} & r_{3,2} & 1 & \cdots & r_{3,M-1} & r_{3,M} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{M-1,1} & r_{M-1,2} & r_{M-1,3} & \cdots & 1 & r_{M-1,M} \\ r_{M,1} & r_{M,2} & r_{M,2} & \cdots & r_{M,M-1} & 1 \end{pmatrix} \tag{1}$$

The conditions on $r_{i,j}$ described above ensures that matrix R is a valid correlation matrix. The Cholesky decomposition of the above matrix L such that $R = LL'$ is given by:

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ l_{2,1} & \sqrt{1-l_{2,1}^2} & 0 & \cdots & 0 & 0 \\ l_{3,1} & l_{3,2} & \sqrt{1-l_{3,1}^2-l_{3,2}^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ l_{M-1,1} & l_{M-1,2} & l_{M-1,3} & \cdots & \sqrt{1-\sum_{k=1}^{M-2} l_{M-1,k}^2} & 0 \\ l_{M,1} & l_{M,2} & l_{M,3} & \cdots & l_{M,M-1} & \sqrt{1-\sum_{k=1}^{M-1} l_{M,k}^2} \end{pmatrix} \quad (2)$$

Estimating the elements above for L directly may work, but can run into problems as the diagonal gets into imaginary space when the quantity within the square root goes to a value of zero or less. Thus, it is desirable to further parameterize the Cholesky elements $l_{i,j}$ to adhere to a positive value for the diagonal quantities throughout the iterative estimation process. Note that it is sufficient to keep the diagonal entries real and positive to ensure that the corresponding matrix is positive definite and has a unique Cholesky decomposition (Martin et al., 1965, Hingham, 2009). We now discuss the spherical and radial parameterizations identified earlier.

2.1 The Spherical and Radial Parameterizations

Consider the parameterization of each element $l_{i,j} = h_{i,j}\sqrt{\prod_{k=1}^{j-1}\left(1-h_{i,k}^2\right)}$, $-1 < h_{i,j} < 1\ (i \neq j)$, $h_{i,i} = 1 \forall i,$ and $l_{i,1} = h_{i,1}\ \forall i$ (because the Cholesky refers to the lower diagonal matrix, the entries correspond to $l_{i,j}$ such that $i \geq j$, a point we will not belabor over in all following notations). The values for each column j for the $h_{i,j}$ values are built up successively from the corresponding row values for the previous j-1 columns. Thus, the parameterization looks like the below:

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ h_{2,1} & \sqrt{1-h_{2,1}^2} & 0 & \cdots & 0 & 0 \\ h_{3,1} & h_{3,2}\sqrt{1-h_{3,1}^2} & \sqrt{\left(1-h_{3,1}^2\right)\left(1-h_{3,2}^2\right)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{M-1,1} & h_{M-1,2}\sqrt{1-h_{M-1,1}^2} & h_{M-1,3}\sqrt{\left(1-h_{M-1,1}^2\right)\left(1-h_{M-1,2}^2\right)} & \cdots & \sqrt{\prod_{k=1}^{M-2}\left(1-h_{M-1,k}^2\right)} & 0 \\ h_{M,1} & h_{M,2}\sqrt{1-h_{M,1}^2} & h_{M,3}\sqrt{\left(1-h_{M,1}^2\right)\left(1-h_{M,2}^2\right)} & \cdots & h_{M,M-1}\sqrt{\prod_{k=1}^{M-2}\left(1-h_{M,k}^2\right)} & \sqrt{\prod_{k=1}^{M-1}\left(1-h_{M,k}^2\right)} \end{pmatrix} \quad (3)$$

It is easy to see that the parameterization above satisfies the condition for the diagonals in Equation (2) that $l_{i,i} = \sqrt{1 - \sum_{k=1}^{i-1} l_{i,k}^2}$. For example, $l_{3,3} = \sqrt{(1-h_{3,1}^2)(1-h_{3,2}^2)} = \sqrt{1-h_{3,1}^2-h_{3,2}^2(1-h_{3,1}^2)}$ $= \sqrt{1-l_{3,1}^2-l_{3,2}^2}$, as required.[2] The additional condition that all diagonal terms $l_{i,i} > 0$ is also satisfied as long as the condition $-1 < h_{i,j} < 1$ $(i \neq j)$ holds. Indeed, the spherical and radial parameterizations both begin with this initial parameterization of the Cholesky elements, followed by a variation in the next level parameterization for $h_{i,j}$ to keep it within the (–1,1) bound. In the spherical parameterization, the requirement for $h_{i,j}$ to be within the (–1,1) bound is achieved using the following parameterization to move from $h_{i,j}$ to unconstrained space:

$$h_{i,j} = \cos\left(\frac{\pi}{1+\exp\left(-\theta_{i,j}\right)}\right),\ i \neq j;\ \theta_{i,j} = -\infty \Rightarrow h_{i,j} = +1 \text{ and } \theta_{i,j} = +\infty \Rightarrow h_{i,j} = -1.^{3} \tag{4}$$

Important to note is that the above transformation may be re-written using the logistic distribution function as:

$$h_{i,j} = \cos\left[\pi F\left(\theta_{i,j}\right)\right],\ i \neq j, \text{ with } F(\theta_{i,j}) = \frac{1}{1+\exp(-\theta_{i,j})} \tag{5}$$

More recently, van Oest (2021) suggested an alternative radial parameterization for $h_{i,j}$.

$$h_{i,j} = \frac{\exp(\tilde{\theta}_{i,j})-1}{\exp(\tilde{\theta}_{i,j})+1} \quad (i \neq j);\ \tilde{\theta}_{i,j} = -\infty \Rightarrow h_{i,j} = -1 \text{ and } \tilde{\theta}_{i,j} = +\infty \Rightarrow h_{i,j} = +1.^{4} \tag{6}$$

The reader will note that the above transformation may also be re-written in terms of the logistic distribution function as follows:

$$h_{i,j} = F(\tilde{\theta}_{i,j}) - F(-\tilde{\theta}_{i,j}), \text{ where } F(\tilde{\theta}_{i,j}) = \frac{1}{1+\exp(-\tilde{\theta}_{i,j})}. \tag{7}$$

---

[2] Note also the corresponding reverse transformation $h_{i,j} = \frac{l_{i,j}}{\sqrt{\left(1-\sum_{k=1}^{j-1} l_{ik}^2\right)}}$ $(i \neq j)$, with $h_{i,1} = l_{i,1}\ \forall i$.

[3] The corresponding reverse transformation is $\theta_{i,j} = \ln\left(\frac{\arccos(h_{i,j})}{\pi - \arccos(h_{i,j})}\right)$, $i \neq j$.

[4] The corresponding reverse transformation in this case is $\theta_{i,j} = \ln\left(\frac{1+h_{i,j}}{1-h_{i,j}}\right)$, $i \neq j$.

In Figure 1, we plot both the spherical (Equation 5) and radial (Equation 7) transformations from the unconstrained real line to the constrained (–1,1) range, after applying a negative sign to the spherical parameterization to put both parameterizations in the same increasing mode for $h_{i,j}$ as a function of $\theta_{i,j}$ / $\tilde{\theta}_{i,j}$ ($\theta_{i,j}$ and $\tilde{\theta}_{i,j}$ increase from $-\infty$ to $+\infty$). As the plot indicates, the second radial parameterization shows a more gradual saturation to the (–1,1) bounds, which may aid better search abilities in the optimization process during estimation. In simulations with a Gaussian copula, van Oest indeed report better convergence properties with the radial parameterization than the spherical parameterization, though whether this extends more generally is an open question. Further, any differences in the two parameterizations may be attributable to scale effects in the embedded logistic distribution component than a function of the parameterizations themselves, as we discuss in the next section.

2.2 Estimating the Appropriate Scale Factors

This section shows that the radial and spherical parameterizations are effectively a scaled version of one another – an important observation that we have not seen reported in the literature so far. To demonstrate this, we first formulate a minimization problem to estimate the appropriate scale factor between the two parameterizations. Specifically, we minimize the root mean squared deviation between the two functional forms given by Equations (5) and (7). For ease of presentation, let us define $\theta_{i,j}$ (from Equation 5) as $\theta_{i,j} = \breve{\theta}_{i,j} / \sigma_{sp}$ where, σ_{sp} is the scale relevant to the spherical parameterization, and $\tilde{\theta}_{i,j}$ (from Equation 7) as $\tilde{\theta}_{i,j} = \breve{\theta}_{i,j} / \sigma_{rp}$ where, σ_{rp} is the scale relevant to the radial parameterization. $\breve{\theta}_{i,j}$ refers to the parameter space on the entire real line. Therefore, the spherical parameterization can be re-written as (following Equation 5):

$$h_{ij}^{sp} = \cos\left[\pi F\left(\breve{\theta}_{i,j} / \sigma_{sp}\right)\right],\ i \neq j \tag{8}$$

And, the radial parameterization can be re-written as (following Equation 7):

$$h_{ij}^{rp} = F(\breve{\theta}_{i,j} / \sigma_{rp}) - F(-\breve{\theta}_{i,j} / \sigma_{rp}),\ i \neq j \tag{9}$$

The minimization objective function can be developed by considering the point-wise differences between the two functional forms (h_{ij}^{sp} and h_{ij}^{rp}) at very fine points along the real line. In particular, we measure the pointwise differences between the two parameterization curves at every 1/1000th point along the real line and consider the root mean squared deviation as the objective function to be minimized. Also, as evident from the parameterization formulas as well as Figure 1, the parameterization values tend to –1 or 1 (depending on the direction of the extremity) as we move away from the center and toward the extremes; therefore, beyond a certain point, the differences between both the curves practically cease to exist and they are almost coincidental. Hence, to keep our minimization problem tractable, we consider a truncated range on the real line, say from b_{low} to b_{high}. In our case, we consider the range from –20 to 20 in

developing the objective function (note that even for more extreme ranges, the minimization problem yielded identical results).

In the formulation, we consider the radial parameterization scale (i.e., σ_{rp}) to be a fixed value and minimize the objective function with respect to the spherical parameterization scale (i.e., σ_{sp}) as the argument (of course, the minimization problem can also be developed in the exact same manner to estimate the radial parameterization scale when the scale of the spherical parameterization is fixed). With the above definitions, the root mean squared deviation can be defined as below (note that a negative sign to the spherical parameterization is inserted to put both parameterizations in the same increasing mode).

$$\Upsilon = \left(\frac{1}{N} \sum_{\substack{\breve{\theta}_{i,j}=\{b_{low},b_{low}+0.001,b_{low}+0.002,\ldots b_{high}\} \\ |\breve{\theta}_{i,j}|=N}} \left(-h^{sp}{}_{i,j} - h^{rp}{}_{i,j}\right)^2 \right)^{1/2} \tag{10}$$

Then, we estimate the spherical scaling corresponding to a specific fixed radial scale parameter as:

$$\sigma_{sp} = \arg\min \Upsilon \tag{11}$$

This minimization effort leads to a scale in the logit function embedded in the spherical parameterization as 1.6130, which provides an almost identical profile to the traditional radial parameterization (with $\sigma_{rp} = 1$). Alternatively, a similar minimization effort leads to a scale estimate of 0.6253 embedded in the radial parameterization, which provides an almost identical profile to the traditional spherical parameterization (with $\sigma_{sp} = 1$). This close tracking is shown in Figure 2, where we plot the h_{ij} profiles for both the traditional spherical parameterization (σ_{sp}=1) as well as the radial parameterization with the embedded scale of 0.6253 (that is, σ_{rp}=0.6253).

To supplement Figure 2 to ensure that the range of $\theta_{i,j} / \tilde{\theta}_{i,j}$ plotted and/or the scale used in the figures is not simply giving a visual illusion of closeness of the two parameterization distributions for h_{ij} (after appropriate scaling), we also numerically evaluate the point-wise differences between the two curves at every 1/1000th point. The regions for least differences between the two profiles, as can also be discerned from Figure 2, are the extreme ends where the curves are effectively coincidental. Therefore, we partition the real line of $\theta_{i,j} / \tilde{\theta}_{i,j}$ into three segments (since the distribution is symmetric, we focus on one side of the zero center): –20 to –10, –10 to –5, and –5 to 0. Then, we measure the point-wise differences and compute the root mean squared deviation (RMSD) separately for each of these three segments (–20 to –10, –10 to –5, and –5 to 0) and for a range of combinations of a specific radial scale and the corresponding appropriately scaled spherical scale. The results are presented in Table 1. The RMSD values for all entries are literally close to zero. For the extreme segment (–20 to –10), the differences are

effectively zero, especially for lower scale values which has a steeper slope (see the first column of Table 1). Even for the other two segments, the difference is at least beyond the third decimal place. This once again reinforces the identicality of the two distributions across all the considered scales when one of the parameterizations is appropriately scaled.

That both the parametrizations are near-equivalent in functional form should be clear from the above discussion. We next investigate, in the context of a multivariate binary response model, whether (and how) the level of scaling itself (within each parameterization) affects numerical performance (convergence rates and convergence times) and small-sample statistics performance (parameter bias, precision, coverage probability, and overall model fit) when embedded within a standard BFGS optimization method. In this first simulation experiment, we use the radial parameterization with different scale values. Also, to examine if the functional form equivalence (after appropriate scaling) between the radial and spherical parameterization carries forward to numerical performance (convergence rates and convergence times) and small-sample statistics performance (parameter bias, precision, coverage probability, and overall model fit) when the two parameterizations are embedded within the standard BFGS algorithm to estimate econometric model parameters, in a second simulation experiment, we also compare the performance of the radial parameterization with the appropriately scaled version of the spherical parameterization. Based on these experiments, we further investigate whether there is a scaling of the logistic function (as embedded in the parameterizations) that offers superior numerical and small sample statistics performances relative to the traditional spherical (σ_{sp}=1) and traditional radial (σ_{rp}=1) parameterizations.

3. DESIGN OF THE SIMULATION STUDY

Multivariate binary response models are becoming increasingly commonplace in a variety of fields, including biology, developmental toxicology, finance, economics, epidemiology, social science, and transportation (see Bhat, 2015, Visaya et al., 2015, Davenport et al., 2018, Spearing et al., 2021, Lin and Chaganty, 2021). The model structure for this model is presented next.

3.1 Model Structure

Consider an individual with I binary outcome choices to make. Let i be the index for a binary outcome (i = 1, 2, …, I, where I is the total number of binary outcomes for each individual; in the current simulation study, we consider five binary outcomes so that I = 5). In the usual latent variable representation for binary outcomes, we write the underlying latent propensity for binary outcome i, y_i^*, as a function of a (C_i×1) column vector of exogenous variables $\mathbf{x}_i$ (including a constant) and a stochastic error term vector ε_i. This latent propensity is mapped to the observed binary outcome as follows:

$$y_i^* = \boldsymbol{\beta}_i' \mathbf{x}_i + \varepsilon_i, \qquad y_i = \begin{cases} 0 & \text{if} \quad y_i^* \le 0 \\ 1 & \text{if} \quad y_i^* > 0 \end{cases} \tag{12}$$

where $\boldsymbol{\beta}_i$ is a ($C_i \times 1$) vector of coefficients to be estimated. ε_i is assumed to be a standard normal error term. More generally, in a multivariate case, we can stack all the error terms of an individual into a single vector, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots, \varepsilon_I)'$, assumed to be jointly distributed according to a multivariate standard normal distribution with a mean vector of zeros and a correlation matrix $\boldsymbol{\Sigma}$ (a generalization of this model would be to allow the univariate distributions to be a non-normal distribution, which then may be tied together using a Gaussian copula to once again create a multivariate standard normal distribution; see, for example, Bhat and Lavieri, 2018). Mathematically,

$$\boldsymbol{\varepsilon} \sim MVN\left[\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1I}\\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2I}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho_{I1} & \rho_{I2} & \rho_{I3} & \cdots & 1\end{pmatrix}\right], \text{ or } \boldsymbol{\varepsilon} \sim MVN[\mathbf{0}, \boldsymbol{\Sigma}] \tag{13}$$

The following additional vectors are defined:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1' & \mathbf{0}_{C_2}' & \mathbf{0}_{C_3}' & \cdots & \mathbf{0}_{C_I}' \\ \mathbf{0}_{C_1}' & \mathbf{x}_2' & \mathbf{0}_{C_3}' & \cdots & \mathbf{0}_{C_I}' \\ \mathbf{0}_{C_1}' & \mathbf{0}_{C_2}' & \mathbf{x}_3' & \cdots & \mathbf{0}_{C_I}' \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{C_1}' & \mathbf{0}_{C_2}' & \mathbf{0}_{C_3}' & \cdots & \mathbf{x}_I' \end{bmatrix} \quad (I \times C \text{ matrix}), \quad C = \sum_{i=1}^{I} C_i, \tag{14}$$

$$\mathbf{y}^* = (y_1^*, y_2^*, ..., y_I^*)' \ (I \times 1 \text{ vector}) \text{ and } \boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_I)' \ (C \times 1 \text{ vector}), \tag{15}$$

where $\mathbf{0}_{C_i}$ is a ($C_i \times 1$) column-vector of zeroes. Then, in vector form, we may write Equation (12) as: $\mathbf{y}^* = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The parameter vector of the multivariate binary probit model (to be estimated) is $\boldsymbol{\delta} = \left[\boldsymbol{\beta}', \left(\text{Vech}(\boldsymbol{\Sigma})\right)'\right]'$, where $\text{Vech}(\boldsymbol{\Sigma})$ refers to the vector of non-diagonal elements of $\boldsymbol{\Sigma}$.

3.2 Model Estimation

Let the actual observed outcome for individual binary outcome *i* be m_i (m_i takes the value of either zero or one). Define a vector $\boldsymbol{m} = (m_1, m_2, m_3, ..., m_I)' \ (I \times 1 \text{ vector})$, and let $\mathbf{b} = (-1)^{\boldsymbol{m}} .*(\mathbf{x}\boldsymbol{\beta})$, where $(-1)^{\boldsymbol{m}}$ refers to the exponentiation of the value '–1' by each element of $\boldsymbol{m}$ to create an $I \times 1$ vector, and the ".*" operator refers to element-by-element multiplication; so the vector $\mathbf{b}$ is also an $I \times 1$ vector). Finally, define $\boldsymbol{\Psi} = \left[(-1)^{(\boldsymbol{m}_q + \boldsymbol{m}_q')}\right].*\boldsymbol{\Sigma}$ (an $(I \times I)$ matrix). Then, exploiting

the radially symmetric nature of the multivariate normal density function, the likelihood function for the individual is:

$$L(\boldsymbol{\delta}) = \Pr(y_1 = m_1,\ y_2 = m_2,\ ...,\ y_I = m_I) = \Phi_I(\mathbf{b};\boldsymbol{\Psi}) \tag{16}$$

where $\Phi_I(\mathbf{b};\boldsymbol{\Psi})$ represents the standard multivariate normal cumulative distribution (MVNCD) function of dimension I, with the upper truncation points given by the vector $\mathbf{b}$ and the correlation matrix given by $\boldsymbol{\Psi}$. Since a closed form expression does not exist for the MVNCD function, and evaluation using simulation techniques can be time consuming, we use the Two-Variate Bivariate Screening (TVBS) technique proposed by Bhat (2018) for approximating this integral (this approach provides an efficient and tractable formulation to approximate high dimensional MVNCD integrals). Of course, during estimation, we parameterize the correlation matrix $\boldsymbol{\Sigma}$ in terms of the $\theta_{i,j}$ or the $\tilde{\theta}_{i,j}$ parameters, and first estimate these parameters. Since the spherical, and radial parameterizations are essentially the same except for scaling, in a first experiment, we arbitrarily use the radial parameterization and write:

$$F(\tilde{\theta}_{i,j}) = \frac{1}{1+\exp\left(-\dfrac{\tilde{\theta}_{i,j}}{\sigma_{rp}}\right)} \tag{17}$$

We then undertake the estimation for different values of the scale factor σ_{rp} to examine the effect of varying this scale factor. Once convergence has been achieved, we reconstruct the implied $\boldsymbol{\Sigma}$ correlation matrix (from the estimated $\tilde{\theta}_{i,j}$ parameters) and run one last iteration to get the correlation parameters and their standard errors. In the estimations, we use a quasi-Newton algorithm based on the standard BFGS update method. The line search method for step length determination is one developed by Dennis and Schnabnel (1981), labeled STEPBT. The convergence criteria used is a gradient tolerance level of 1e-4.[5]

[5] Please note that the scaling introduced through σ_{rp} in Equation (17) is embedded within the logistic function of $h_{i,j}$, which itself represents a parameterization of the Cholesky of the correlation matrix. This scaling, deeply embedded in the parameterization, is not the same as self-scaling BFGS approaches that have been suggested in the optimization process itself to limit the adverse effects of an initial poor approximate of the Hessian matrix or to overcome the ill-conditioning of the approximate Hessian at subsequent iterations. Many such self-scaling BFGS approaches have been proposed, originating first in the contribution of Oren and Luenberger, 1974), to identify efficient search directions through the scaling of the update formula of the BFGS method (see good reviews in Nocedal and Yuan, 1993, Al-Baali et al., 2014, Gao and Goldfarb, 2019, and Andrei, 2018). This scaling, undertaken so that the range of the eigenvalues of the approximate Hessian update after any iteration is dampened from those of the approximate Hessian at the end of the earlier iteration (or the initial approximate Hessian at the end of the first iteration), itself may be through scaling the approximate Hessian before use in the update formula, or by scaling/shifting the gradient change vector during the update. While some of these scaling approaches have been shown to have theoretically better superlinear convergence properties than the standard BFGS used here, such theoretical results do not always carry over to numerical performance in actual empirical contexts. In large part, this is because it is difficult to obtain a good step length in such self-scaling methods, resulting in the standard BFGS generally being superior (in terms of the number of function evaluations and overall CPU time) relative to self-scaling methods (see, for example, Al-Baali et

As discussed in the earlier section, in addition to examining the role of the scale for the radial parameterization, in a second experiment, we also undertake another set of estimations using the spherical parameterization but with the scale set at the comparable scale for the radial parameterization.

3.3 Experimental Design

To compare and evaluate the performance of different scales in the embedded logistic function, we undertake a simulation exercise for a multivariate binary choice system with five endogenous outcomes. Further, to examine the potential impact of different correlation structures, we undertake the simulation exercise for a correlation structure with low correlations and another with high correlations. For each correlation structure, the experiment is carried out for 500 independent data sets with 2000 data points each (for a total of 1000 data sets). Pre-specified values for the $\boldsymbol{\delta}$ vector are used to generate the samples, as discussed below.

In the set-up, we use a constant and an exogenous variable in each latent equation (that is, as elements of the $\mathbf{x}_i$ vector). The constants are set to $\beta_{1c} = -0.25$, $\beta_{2c} = -0.5$, $\beta_{3c} = -0.75$, $\beta_{4c} = -1$, and $\beta_{5c} = -1.25$, starting from the first outcome to the last. The exogenous variable for each of the first, second, and third binary outcomes is included as a continuous variable, while the exogenous variable for each of the last two binary outcomes is included as a dummy (binary) variable. The values for the continuous variables are drawn from standard univariate normal distributions. The parameters on the continuous variables (say β_{1v}, β_{2v}, and β_{3v}) are specified to be one across the first three outcomes. For the dummy variable in the last two binary outcomes, we draw 2000 independent values for each outcome from the standard uniform distribution. For the penultimate outcome, if the value drawn is less than 0.5, the value of '0' is assigned for the dummy variable. Otherwise, the value of '1' is assigned. The coefficient on this dummy variable (say β_{4v}) is specified to be +0.5. A similar procedure is used to construct the dummy variable for the final outcome, except that the threshold is increased to 0.7 (to create an asymmetry with more values of the dummy variable toward the value of zero than the value of one). The coefficient on this dummy variable (β_{5v}) is specified to be –0.5. Once generated, the exogenous variables are held fixed for the rest of the simulation exercise (that is, across all the 1000 data samples, the same exogenous variable values are used).

Next, we generate, for each of the 2000 observations in each of the 1000 data samples, a five-variate realization of the error term vector $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5)$ with predefined positive-definite

al., 2014, Gao and Goldfarb, 2019, Lv et al. 2020, and Runnoe, 2020). Even theoretically speaking, new results by Rodomanov and Nesterov (2021) establish that the local superlinear convergence rates from the classical BFGS method for unconstrained optimization (as is made possible by the parameterizations in our current paper) are much better than originally thought. Add to this the remarkable simplicity of the classical BFGS in practice, it is generally recognized that the BFGS method is "best" in practice (Runnoe, 2020). As Andrei (2018) also exclaims "…the BFGS method has very interesting properties and remains one of the most respectable quasi-Newton methods for unconstrained optimization". Thus, we use the same standard BFGS method for both the radial and spherical parameterization-based estimations.

low error correlation structure ($\boldsymbol{\Sigma}_{low}$) and high error correlation structure ($\boldsymbol{\Sigma}_{high}$) such that both positive and negative elements are represented in the correlation structures. A positive-definite low error correlation structure with positive and negative correlation elements is easy to create. For the high correlation structure, we use a Toeplitz matrix that enables the creation of positive-definite correlation matrices even with high positive and negative correlations. Specifically, as indicated in Bogoya et al. (2018), if a linearly decreasing sequence (of positive and negative values) is used to create a Toeplitz matrix, the matrix will remain a positive-definite matrix as long as the sum of the values in the sequence is positive. Using this result, we create a positive-definite matrix and then make small adjustments to increase the correlation magnitudes so that the resulting correlation matrix remains positive definite. The final correlation matrices used are presented below:

$$\boldsymbol{\Sigma}_{low} = \begin{bmatrix} 1.00 & 0.30 & 0.20 & -0.20 & -0.15 \\ 0.30 & 1.00 & 0.25 & -0.30 & 0.10 \\ 0.20 & 0.25 & 1.00 & -0.25 & -0.20 \\ -0.20 & -0.30 & -0.25 & 1.00 & 0.25 \\ -0.15 & 0.10 & -0.20 & 0.25 & 1.00 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{high} = \begin{bmatrix} 1.00 & 0.60 & 0.20 & -0.50 & -0.80 \\ 0.60 & 1.00 & 0.60 & 0.20 & -0.50 \\ 0.20 & 0.60 & 1.00 & 0.60 & 0.20 \\ -0.50 & 0.20 & 0.60 & 1.00 & 0.60 \\ -0.80 & -0.50 & 0.20 & 0.60 & 1.00 \end{bmatrix} \tag{18}$$

The error term realization for each observation and each binary variable is then added to the systematic component ($\boldsymbol{\beta}_i'\mathbf{x}_i$) as in Equation (12) and then translated to "observed" values of y_i. As mentioned earlier, the above data generation process is undertaken 1000 times with different realizations of the random error vector to generate 1000 different data sets. The estimation is undertaken for each data set with seven different values of the scale σ (σ = 0.2, 0.6, 0.8, 1, 1.2, 1.4, 2) in the embedded logistic function of the radial parameterization (Equation 17). Another set of estimations are then undertaken with the corresponding equivalent scale in the embedded logistic function of the spherical parameterization: Thus, we have a total of 14,000 multivariate binary model estimations.

3.4 Performance Evaluation

The performance of the models with different scales, for each of the low correlation and high correlation cases, is evaluated using multiple metrics for recovering model parameters as well as for the actual predictions. The procedure is as follows:

(1) Estimate the parameters for each of the 500 datasets for the low correlation case and 500 datasets for the high correlation case, using each of the scale values of σ. Estimate the standard errors. For each correlation case and each σ value, do the following:

(2) Compute the percentage of non-convergence estimations among the 500 datasets. Among those datasets in which convergence was achieved, undertake the following:

(3) Compute the mean estimate for each model parameter across the 500 data sets. Compute the **absolute percentage (finite sample) bias** (APB) as:

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

(4) Compute the standard deviation of each parameter estimate across the 500 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.

(5) Compute the mean standard error for each model parameter across the 500 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).

(6) 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.

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

(7) For each parameter, compute the **coverage probability (CP)** as below:

$$\text{CP} = \frac{1}{N} \sum_{r=1}^{N} I\left[\hat{\beta}_X^r - t_\alpha * \text{se}(\hat{\beta}_X^r) \le \beta_X \le \hat{\beta}_X^r + t_\alpha * \text{se}(\hat{\beta}_X^r) \right],$$

where, CP is the coverage probability, $\hat{\beta}_X^r$ is the estimated value of the parameter in dataset r, β_X is the true value of the parameter, $\text{se}(\hat{\beta}_X^r)$ is the asymptotic standard error (ASE) of the parameter in the dataset r, $I[.]$ is an indicator function which takes a value of 1 if the argument in the bracket is true (otherwise 0), N is the number of datasets (500), and t_α is the t-statistic value for a given confidence level $(1-\alpha)\times 100$. We compute CP values for 80% nominal coverage probability (i.e., $\alpha = 0.20$). CP is the empirical probability that a confidence interval contains the true parameter (i.e., the proportion of confidence intervals across the 500 datasets that contain the true parameter). CP values smaller than the nominal confidence level (80% in our study) suggest that the confidence intervals do not provide sufficient empirical coverage of the true parameter.

(8) Examine the data fit at a disaggregate level by comparing the log-likelihood values at convergence of the models. The model with the higher log-likelihood value is to be preferred, because all the models have the same number of estimated parameters. Based on the log-likelihood values for each of the 500 runs (corresponding to the 500 datasets), compute a mean log-likelihood value. In addition, also compute the average probability of correct prediction for the discrete consumption across the 500 datasets. That is, for each of the 500 datasets, compute the predicted multivariate probability of the observed discrete choice for each observation, and then compute an average across individuals. This average probability of correct prediction at a dataset-level is then averaged across the 500 datasets to obtain a single average probability of correct prediction.

(9) Finally, at the aggregate level, for each dataset, predict the aggregate share of individuals participating in each of the $2^5 = 32$ possible multivariate discrete outcomes, and compare these predicted shares with the actual percentages of observations in each multivariate combination (using the weighted MAPE statistic, which is the MAPE for each combination weighted by the actual percentage shares of observations participating in each combination). Next, compute the average of the weighted MAPE statistic across the 500 datasets.

(10) Store the run time for estimation for each data set, and compute the mean, median, and standard deviation of the run times across the 500 data sets.

4. PERFORMANCE EVALUATION RESULTS

Table 2 presents an overall summary of the small sample statistics performance and numerical performance for the seven different scaling values in the radial parameterization, and for each of the low and high correlation cases.[6] For each scaling factor, the first panel of rows presents the average APB values (across all parameters), as well as the average APB values computed separately for the mean parameters (the $\boldsymbol{\beta}$ vector) and the correlation matrix (the upper diagonal $\boldsymbol{\Sigma}$ matrix or $\text{Vech}(\boldsymbol{\Sigma})$) elements. The second through fifth panels provide the corresponding average FSSD, ASE values, APBASE, CP, and data fit measures. The final block provides information on the numerical performance for the different radial scaling values (convergence rates, and the mean, median, and standard deviations of the model estimation run times across all the converged estimations from the 500 datasets).

4.1 Accuracy of Parameter Recovery and Precision in Estimation

The APB values pretty consistently indicate the superior ability to recover the mean parameters (the $\boldsymbol{\beta}$ vector) relative to the correlation parameters; this is because the correlation parameters enter the likelihood function in a more complex non-linear fashion compared to the mean parameters, and thus are more difficult to accurately recover. However, for each of the mean and correlation parameters, the APB values do not vary by much across different scaling values and lie in the tight (and small magnitude) range of 0.675%-1.61% across all parameters. There is also no specific discernible pattern in the accuracy, even if the APB is the lowest for a scale of 1.4 in the low correlation case. Overall, for the datasets that converged, all the scaled parameterizations were able to quite accurately recover the true parameter for both the low and high correlation cases.

The values in the second, third and fourth blocks, corresponding to the FSSD, ASE and APBASE, relate to the precision in estimation. The FSSD values are useful for assessing the empirical (finite-sample) efficiency (or precision) of the different estimators, while the ASE values provide efficiency results as the sample size gets very large. The ASE values essentially provide an approximation to the FSSD values for finite samples. The APBASE values as obtained in Table

[6] The detailed results for all the cases are available in an online supplement at https://www.caee.utexas.edu/prof/bhat/ABSTRACTS/Cholesky/OnlineSupp.pdf.

2 indicate that there is no significant difference between the precision levels of the estimates across different scale factors, for both the low and high correlation categories. Remarkably, all the APBASE values are lower than 5%. Again, similar to the parameter recovery case, the scale factor does not appear to impact the accuracy or the precision of the estimates, even though it does appear that a scale of 1.0-1.2 provides APBASE values at the lowest end of the spectrum.

4.2 Coverage Probability (CP) and Data Fit Measures

The fifth block of Table 2 provides the coverage probability (CP) values for all the cases. The CP values help assess the distribution of the parameter estimates about the true parameter in terms of the empirical probability that a confidence level contains the true parameter. As one may observe from Table 2, all scale values provide good empirical coverage of the 80% nominal confidence interval, with little to no effective difference in these empirical coverages (all the values are above 80%).

The sixth block presents the data fit measures at disagreggate as well as aggregate levels. Again, the consistency in all these data fit measures across all the scale values for each of the low and high correlation cases is remarkable, with little difference across the scales. For the low correlation case, a scale of 0.8 seems to work best across all the data fit measures, while, for the high correlation case, a scale of 1.2 appears to work best. But, in general, the performance of all the models are commendable with the weighted MAPE values all being below 6% and sandwiched tightly between the 5.50%-5.82% range. The average probability of correct prediction (APCP) values are in the order of 0.19 for the low correlation case and 0.23 for the high correlation case. These average probabilities may appear low, but considering that the five outcome variables can produce a total of $2^5 = 32$ outcome combinations, these values are much higher than the probability of correct prediction in the case of a random chance assignment (=1/32=0.03125).

4.3 Computation Time and Convergence Rate

The last block in Table 2 provides the numerical performance measures (convergence rates and computation times) for the different models. All estimations were undertaken using Intel(R) Xeon(R) CPU E5-1680 v4 @3.40GHz, Windows 10 Enterprise (64 bit), 192.0 GB RAM machine. The virtue of having a gradual transformation from the real line to the {–1,1} region for the parameterizing curve is reflected in the convergence success rates. Specifically, the danger of having too low a scale factor (that is a sharp rise from –1 to +1 within a very narrow real line spectrum) is obvious, particularly for the high correlation case. For instance, only about 86% of the runs converge when a scale of 0.2 is used for the high correlation case. Having a steep transformative curve can lead to oscillations between two sub-optimal points, causing convergence to be unattainable. However, this does not imply that convergence rates will improve by unilaterally increasing the scale factor. In particular, there is little benefits to increasing the scale beyond 0.8; in fact, for the high correlation case, there is a pretty sharp degradation in the convergence rates beyond the scale of 1.0. This may be because, for some specific estimation contexts, line-search algorithms find it difficult to navigate the optimization pathway space

regardless of the flatness of the presented log-likelihood surface, especially in high-correlation cases which have higher degree of non-linearity in the parameter-to-likelihood function translation.

In terms of convergence times across converged estimations, it is not surprising that the mean and median computation times increase with the flatness of the surface introduced through higher scale factor values. The flatter the transformation curve, the slower the movement of the likelihood function toward the optimal value. Also, the computation time is higher in the case of high correlation relative to low correlation for a given scale value. This is because of two reasons. The first is the higher degree of non-linearity in the parameter-to-likelihood function translation in the high correlation case relative to the low correlation case, as mentioned in the last paragraph. The second is the more mechanical issue that the optimizing process is typically started with values of zero correlation in multivariate models, and thus reaching to the low correlations takes lesser time. In our experience, the first issue dominates the second, given the nature of the non-linear search in the maximum likelihood optimization.

In summary, good convergence rates, fast computational speed, and good data fit suggest using a scale in the radial parameterization in the 0.6-0.8 scale range. While a scale of 0.8 increases the computation time relative to the scale of 0.6, it also seems to provide better data fit measures. Overall, a scale of 0.8 is a good scale value to settle for. Interestingly, this scale value is somewhere between the implied scale of the standardized spherical parameterization in the radial parameterization (scale of 0.6253) and the standardized scale of the radial parameterization (scale of 1.0). Thus, our results point to the use of a radial parameterization with a scale of 0.80 (or the spherical parameterization with a scale of 1.2904).

4.4 Comparison of the Radial and Equivalent Spherical Parameterization

To further investigate whether the differences in the radial and spherical parameterizations are based on the embedded scale rather than on the specific parameterization per se, in a second experiment, we estimated the same models as above but now using the spherical parameterization with the scales set such that the spherical parameterization mimics the radial parameterization. This mapping of scales (approximated to two decimal places) between the scale of the radial and spherical parameterizations is as follows: 0.2 in radial is 0.32 in the spherical, 0.60 is 0.97, 0.80 is 1.29, 1.00 is 1.61, 1.20 is 1.93, 1.40 is 2.26, and 2.00 is 3.22.

The results of the spherical paramerizations using the appropriate scales indicated that, when convergence was achieved in both cases, the spherical parameterization returned the same parameters and standard errors as the radial parameterization (upto the third or fourth decimal place). These results reinforce our point that it is not the specific parameterization as much as it is the scale embedded in the logistic function that matters. Because the small sample statistics results (parameter recovery, precision, and data fit) are almost exactly identical between the radial and equivalent spherical parameterizations (that is, after appropriate scaling), we do not present the detailed small sample results for the spherical parameterization here. However, in Table 3, we do show a numerical performance comparison of the results for the radial and equivalent spherical

parameterizations in terms of convergence rates and computation time (when convergence was achieved). The row labeled "% of runs converged" for both the low and high correlation cases indicates almost literally the same convergence rates for the radial and equivalent spherical parameterizations (except for those rare instances in the order of two to twelve cases of the 500 data runs where one converged but not the other). And the mean run times and the median run times among converged runs are about the same across both the parameterizations after controlling for the scale. The standard deviation appears a little higher for the spherical parameterization relative to its corresponding radial parameterization at low scale values, especially for the high correlation case. Overall, though, our results clearly demonstrate the near-equivalence of the radial and spherical parameterizations when embedded within the standard BFGS algorithm, and suggest the use of a scaling factor (in the embedded logistic function) of 0.8 if the radial parameterization is used, or a scaling factor of 1.2904 if the spherical parameterization is used.

5. SUMMARY AND CONCLUSIONS

Many multivariate modeling systems require a correlation matrix to be estimated. A requirement for these correlation matrices is that they be positive-definite. While constrained optimization methods may be used during estimation to ensure this condition, such methods require some level of trial-and-error, and lead to difficulties in convergence. Thus, it is almost always the case that a reparameterization of the correlation matrix is undertaken to ensure positive-definiteness, while also allowing for unconstrained optimization.

In addition to maintaining positive-definiteness, the reparameterization of the correlation matrix also needs to conform to the unit diagonal vector in the correlation matrix. In this regard, the usual Cholesky decomposition applied to the covariance matrix does not immediately work for correlation matrices, because the diagonal elements of the Cholesky of a correlation matrix involve the square root of one minus the linear sum of the Cholesky elements in previous columns of the same row. During estimation, there is nothing to prevent the term within the square root of these diagonal elements to take a value of zero or even a negative number. This breaks down the estimation process. To resolve this issue, a further parameterization of the Cholesky matrix itself is undertaken. The most common and well-established procedure is to use a spherical parameterization. More recently, van Oest (2021) suggested an alternative radial parameterization, suggesting that the radial parameterization may be more effective than the spherical parameterization.

In this paper, we show that the two parameterizations, radial and spherical, have a very close relation. Both of these parameterizations involve a first parameterization (of the Cholesky) such that the off-diagonal elements are between –1 and +1. The difference between the two parameterizations is in the second level of mapping from the real line to the –1-to-+1 off-diagonal elements of the first level parameterization. We identify a logit function that is embedded within each of the two different ways of parameterizing at this second-level, and demonstrate that both parameterizations are near-identical based on the scaling used in this embedded logit function. We also show that any difference between the two parameterizations in terms of numerical

performance and small sample statistics is simply a result of the different implicit scales used in the embedded logit function, and not a function of the parameterizations themselves, at least when considered in the context of the popular standard BFGS algorithm. We further show that the optimal scale (in terms of convergence rates and computational time) of the embedded logit link is between the implicit scales of the radial and spherical parameterizations. Both the parameterizations have been coded (along with their gradient functions) in the GAUSS matrix programming language (Aptech GAUSS version 21), and are available for free download from https://www.caee.utexas.edu/prof/bhat/CodeRepository/Cholesky.html. In our experience, both parameterizations are equally easy to code and implement, and the choice of one over the other appears to be purely an issue of preference. If the radial parameterization is used, the optimal scale value in our analysis is 0.8, and if the spherical parameterization is used, the optimal scale is 1.29.

Of course, the current results for the optimal scale are based on a five-dimensional binary model system. Additional explorations in the context of other types of models may be useful, though the result from our study that there is little difference in the radial and spherical parameterizations after a scale adjustment should always hold -- there is no clear reason to believe dissimilar results would be obtained with other types of model systems when the parameters are estimated using a standard BFGS optimization method.

ACKNOWLEDGMENTS

The authors are grateful to Lisa Macias for her help in formatting this document.

Disclosure statement

The authors report there are no competing interests to declare.

References

Al-Baali, M., Spedicato, E., and Maggioni, F. (2014). Broyden's quasi-Newton methods for a nonlinear system of equations and unconstrained optimization: A review and open problems. *Optimization Methods and Software*, 29(5), 937-954. https://doi.org/10.1080/10556788.2013.856909

Andrei, N. (2018). An adaptive scaled BFGS method for unconstrained optimization. *Numerical Algorithms*, 77, 413-432. https://doi.org/10.1007/s11075-017-0321-1.

Barnard, J., McCulloch, R., and Meng, X. (2000). Modelling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4), 1281-1311.

Bhat, C.R. (2015). A new spatial (social) interaction discrete choice model accommodating for unobserved effects due to endogenous network formation. *Transportation*, 42, 879-914. https://doi.org/10.1007/s11116-015-9651-9.

Bhat, C.R. (2018). New matrix-based methods for the analytic evaluation of the multivariate cumulative normal distribution function. *Transportation Research Part B*, 109, 238-256. https://doi.org/10.1016/j.trb.2018.01.011.

Bhat, C.R., and Lavieri, P.S. (2018). A new mixed MNP model accommodating a variety of dependent non-normal coefficient distributions. *Theory and Decision*, 84, 239-275. https://doi.org/10.1007/s11238-017-9638-4.

Bogoya J.M., Grudsky S.M., and Malysheva I.S. (2018). Extreme individual eigenvalues for a class of large Hessenberg Toeplitz matrices. In: Bart H., ter Horst S., Ran A., Woerdeman H. (eds), *Operator Theory, Analysis and the State Space Approach*. Operator Theory: Advances and Applications, Vol 271. https://doi.org/10.1007/978-3-030-04269-1_4.

Davenport, C.A., Maity, A., Sullivan, P.F., and Tzeng, J.-Y. (2018). A powerful test for SNP effects on multivariate binary outcomes using kernel machine regression. *Statistics in Biosciences*, 10, 117-138 https://doi.org/10.1007/s12561-017-9189-9.

Dennis, J.E. Jr., and Schnabel, R.B. (1981). A new derivation of symmetric positive definite secant updates. In: Mangasarian, O.L., Meyer, R.R., and Robinson, S.M. (eds), *Nonlinear Programming, Vol. 4*, pp. 167-199, Academic Press, New York.

Dias, F., Lavieri, P., Sharda, S., Khoeini, S., Bhat, C.R., Pendyala, R., Pinjari, A., Ramadurai, G., and Srinivasan, K. (2020). A comparison of online and in-person activity engagement: The case of shopping and eating meals. *Transportation Research Part C*, 114, 643-656. https://doi.org/10.1016/j.trc.2020.02.023.

Ferdous, N., Eluru, N., Bhat, C.R., and Meloni, I. (2010). A multivariate ordered-response model system for adults' weekday activity episode generation by activity purpose and social context. *Transportation Research Part B*, 44(8-9), 922-943. https://doi.org/10.1016/j.trb.2010.02.002.

Forrester, P.J., and Zhang, J. (2020). Parametrising correlation matrices. *Journal of Multivariate Analysis*, 178, 104619. https://doi.org/10.1016/j.jmva.2020.104619.

Gao, W., and Goldfarb, D. (2019). Quasi-Newton methods: Superlinear convergence without line searches for self-concordant functions. *Optimization Methods and Software*, 34(1), 194-217. https://doi.org/10.1080/10556788.2018.1510927.

Higham, N.J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3), 329-343. https://doi.org/10.1093/imanum/22.3.329.

Higham, N.J. (2009). Cholesky factorization. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2), 251-254. https://doi.org/10.1002/wics.18.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10), 2177-2189. https://doi.org/10.1016/j.jmva.2005.05.010.

Kohli, N., Peralta, Y., and Bose, M. (2019). Piecewise random-effects modeling software programs. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1), 156-164. https://doi.org/10.1080/10705511.2018.1516507.

Lin, H., and Chaganty, N.R. (2021). Multivariate distributions of correlated binary variables generated by pair-copulas. *Journal of Statistical Distributions and Applications*, 8(4). https://doi.org/10.1186/s40488-021-00118-z.

Lv, J., Deng, S., and Wan, Z. (2020). An efficient single-parameter scaling memoryless Broyden-Fletcher-Goldfarb-Shanno algorithm for solving large scale unconstrained optimization problems. *IEEE Access*, 8, 85664-85674. https://doi.org/10.1109/ACCESS.2020.2992340.

Madar, V. (2015). Direct formulation to Cholesky decomposition of a general nonsingular correlation matrix. *Statistics and Probability Letters*, 103, 142-147. https://doi.org/10.1016/j.spl.2015.03.014.

Martin, R.S., Peters, G., and Wilkinson, J.H. (1965). Symmetric decomposition of a positive definite matrix. *Numerische Mathematik*, 7, 362-383.

McNeish, D., and Bauer, D.J. (2020). Reducing incidence of nonpositive definite covariance matrices in mixed effect models. *Multivariate Behavioral Research*, forthcoming. https://doi.org/10.1080/00273171.2020.1830019.

Mondal, A., Bhat, C.R., Costey, M., Bhat, A.C., Webb, T., Magassy, T., Pendyala, R.M., and Lam, W.H.K. (2020). How do people feel while walking? A multivariate analysis of emotional well-being for utilitarian and recreational walking episodes. *International Journal of Sustainable Transportation*, 15(6), 419-434. https://doi.org/10.1080/15568318.2020.1754535.

Nocedal, J., and Yuan, Y., (1993) Analysis of a self-scaling quasi-Newton method. *Mathematical Programming*, 61, 19-37. https://doi.org/10.1007/BF01582136.

Oren, S.S., and Luenberger, D.G. (1974). Self-scaling variable metric (SSVM) algorithms, part I: Criteria and sufficient conditions for scaling a class of algorithms. *Management Science*, 20(5), 845-862. https://doi.org/10.1287/mnsc.20.5.845.

Pinheiro, J.C., and Bates, D.M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6, 289-296. https://doi.org/10.1007/BF00140873.

Pourahmadi, M., and Wang, X. (2015). Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor. *Statistics & Probability Letters*, 106, 5-12. https://doi.org/10.1016/j.spl.2015.06.015.

Rebonato, R., and Jäckel, P. (2000). The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. *The Journal of Risk*, 2(2), 17-27. https://doi.org/10.21314/JOR.2000.023.

Rodomanov, A., and Nesterov, Y. (2021) New results on superlinear convergence of classical quasi-Newton methods. *Journal of Optimization Theory and Applications*, 188(3),744-769. https://doi.org/10.1007/s10957-020-01805-8.

Runnoe, J.H. (2020). Quasi-Newton methods for unconstrained optimization. Thesis, Department of Mathematics, University of California, San Diego.

Schöttle, K., and Werner, R. (2004). Improving the most general methodology to create a valid correlation matrix. *WIT Transactions on Ecology and the Environment*, 77. DOI: 10.2495/RISK040641.

Spearing, L.A., Dias, F.F., Faust, K.M., and Bhat, C.R. (2021). Determining multilevel drivers of perceiving undesirable taste and odor in tap water: Joint modeling approach. *Journal of Water Resources Planning and Management*, 147(3), 04020114. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001326

Srinivasan, S., and Bhat, C.R. (2005) Modeling household interactions in daily in-home and out-of-home maintenance activity participation. *Transportation*, 32, 523-544. https://doi.org/10.1007/s11116-005-5329-z.

Tsay, R., and Pourahmadi, M. (2017). Modelling structured correlation matrices. *Biometrika*, 104(1), 237-242. https://doi.org/10.1093/biomet/asw061.

van Oest, R. (2021). Unconstrained Cholesky-based parametrization of correlation matrices. *Communications in Statistics - Simulation and Computation*, 50(11), 3607-3613. https://doi.org/10.1080/03610918.2019.1628271.

Visaya, M.V., Sherwell, D., Sartorius, B., and Cromieres, F. (2015). Analysis of binary multivariate longitudinal data via 2-dimensional orbits: An application to the Agincourt Health and Socio-Demographic Surveillance System in South Africa. PLoS ONE, 10(4): e0123812. https://doi.org/10.1371/journal.pone.0123812.

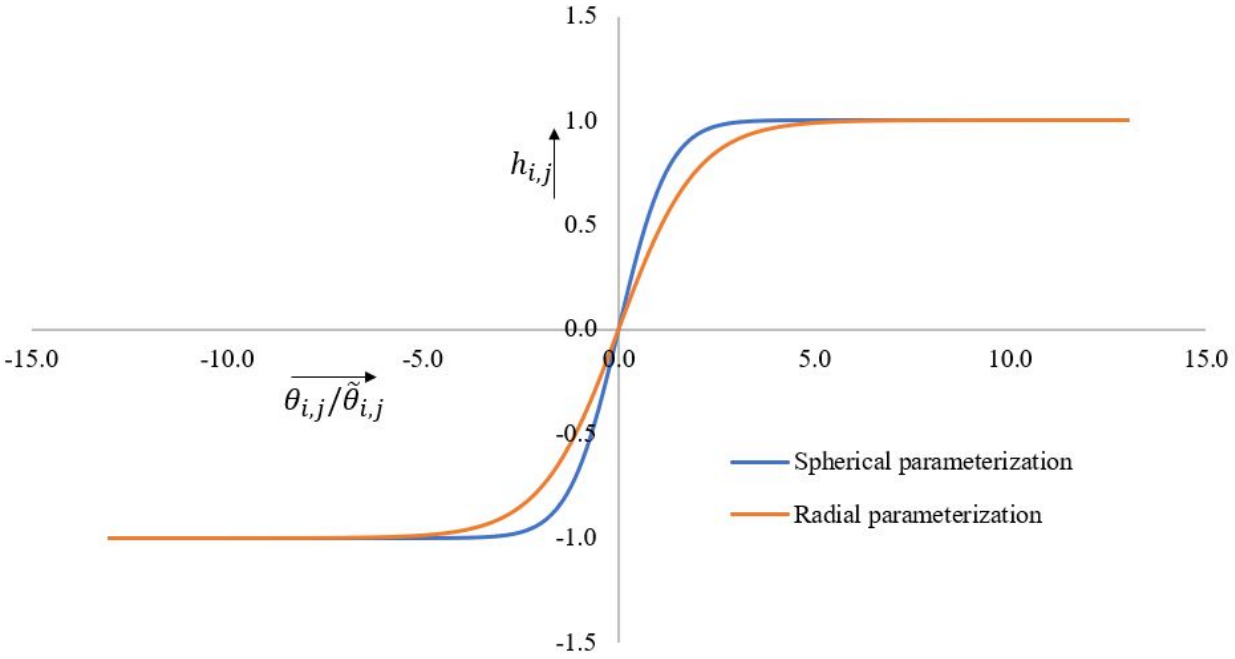

Figure 1. Traditional spherical and radial parameterization plots on the real line

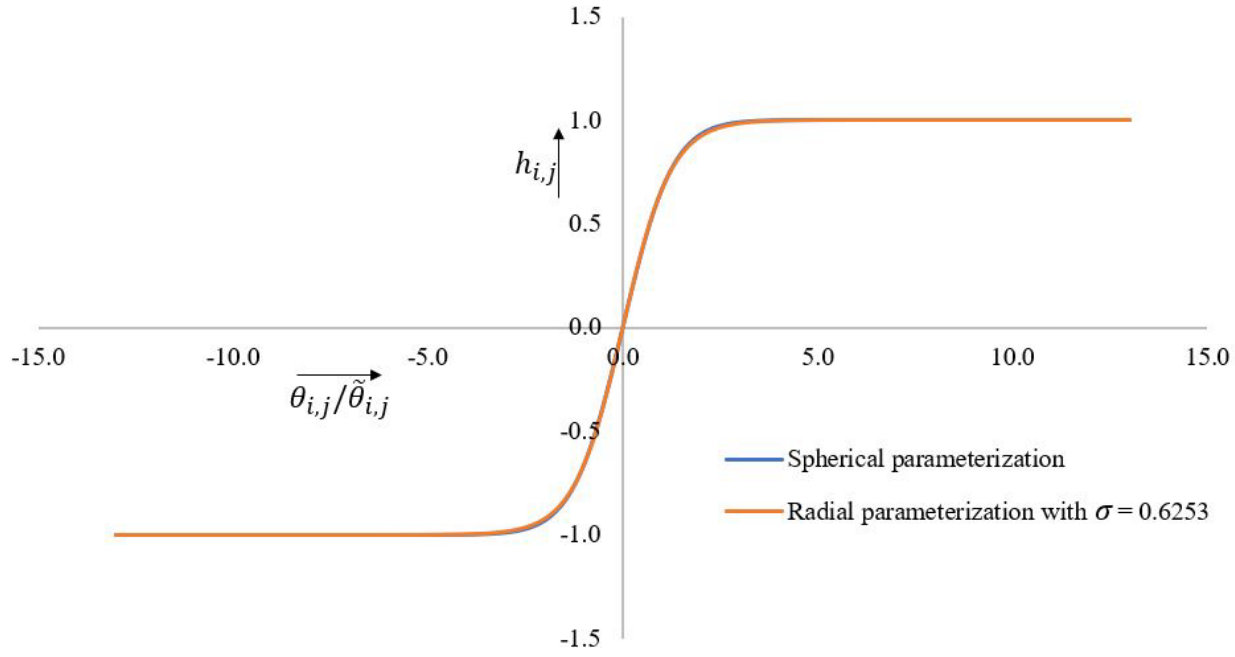

Figure 2. Scaling the radial parameterization to show its equivalence with the spherical parameterization

Table 1. Root mean squared deviation (RMSD) for pointwise comparison

Scale	RMSD		
	Segment 1 (–20.0 to –10.0)	**Segment 2 (–10.0 to –5.0)**	**Segment 3 (–5.0 to 0.0)**
Radial scale = 0.20, Spherical scale = 0.32	0	3.98E-12	2.60E-03
Radial scale = 0.60, Spherical scale = 0.97	1.93E-08	8.36E-05	4.50E-03
Radial scale = 0.80, Spherical scale = 1.29	1.33E-06	5.76E-04	5.17E-03
Radial scale = 1.00, Spherical scale = 1.61	1.63E-05	1.62E-04	5.57E-03
Radial scale = 1.20, Spherical scale = 1.93	8.32E-05	2.96E-03	5.62E-03
Radial scale = 1.40, Spherical scale = 2.26	2.56E-04	4.23E-03	5.39E-03
Radial scale = 2.00, Spherical scale = 3.22	1.62E-03	6.18E-03	4.89E3-03

Table 2. Performance evaluation results for the simulation experiment

	Low Correlation							High Correlation						
	Scale = 0.2	**Scale = 0.6**	**Scale = 0.8**	**Scale = 1.0**	**Scale = 1.2**	**Scale = 1.4**	**Scale = 2.0**	**Scale = 0.2**	**Scale = 0.6**	**Scale = 0.8**	**Scale = 1.0**	**Scale = 1.2**	**Scale = 1.4**	**Scale = 2.0**
Absolute Percentage Bias (APB)														
All parameters	1.172	0.858	0.960	1.104	0.930	0.675	0.839	1.230	1.367	1.406	1.436	1.318	1.325	1.610
Mean parameters	0.565	0.410	0.343	0.590	0.475	0.353	0.494	1.161	1.086	1.139	1.114	1.022	0.910	1.148
Correlation parameters	1.779	1.307	1.577	1.618	1.385	0.996	1.185	1.298	1.649	1.672	1.758	1.614	1.741	2.072
Finite Sample Standard Deviation (FSSD)														
All parameters	0.050	0.050	0.050	0.051	0.050	0.050	0.050	0.044	0.043	0.045	0.044	0.044	0.044	0.044
Mean parameters	0.049	0.050	0.050	0.050	0.050	0.050	0.050	0.047	0.045	0.047	0.045	0.046	0.046	0.046
Correlation parameters	0.050	0.050	0.050	0.051	0.050	0.050	0.050	0.041	0.042	0.043	0.042	0.043	0.043	0.042
Asymptotic Standard Error (ASE)														
All parameters	0.051	0.051	0.051	0.051	0.051	0.051	0.051	0.045	0.045	0.044	0.045	0.045	0.045	0.045
Mean parameters	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.047	0.047	0.045	0.047	0.046	0.046	0.047
Correlation parameters	0.052	0.051	0.052	0.052	0.051	0.052	0.051	0.043	0.044	0.042	0.043	0.044	0.044	0.044
Absolute Percentage Bias of the Asymptotic Standard Error (APBASE)														
All parameters	2.902	3.261	3.276	2.479	3.168	3.185	3.181	3.952	4.938	3.882	3.806	2.929	3.299	3.879
Mean parameters	2.078	2.414	2.497	2.490	2.275	2.885	2.358	2.386	4.604	3.472	3.787	2.001	2.516	3.103
Correlation parameters	3.726	4.108	4.055	2.469	4.060	3.486	4.005	5.517	5.272	4.292	3.825	3.856	4.083	4.655
Coverage Probability (CP)														
$CP_{80\%}$	81.84%	81.26%	81.28%	81.40%	81.40%	81.30%	81.48%	81.74%	81.50%	81.68%	81.24%	81.62%	81.84%	81.60%

	Low Correlation							High Correlation						
	Scale = 0.2	**Scale = 0.6**	**Scale = 0.8**	**Scale = 1.0**	**Scale = 1.2**	**Scale = 1.4**	**Scale = 2.0**	**Scale = 0.2**	**Scale = 0.6**	**Scale = 0.8**	**Scale = 1.0**	**Scale = 1.2**	**Scale = 1.4**	**Scale = 2.0**
Data Fit Measures														
Mean likelihood (across all datasets)	-4341.31	-4350.59	-4328.52	-4321.82	-4350.26	-4334.91	-4350.18	-3903.72	-3891.16	-3895.22	-3895.52	-3871.01	-3871.89	-3880.45
Average prob. of correct prediction	0.187	0.187	0.189	0.190	0.187	0.188	0.187	0.229	0.231	0.231	0.231	0.232	0.232	0.232
Weighted mean absolute percentage error (%)	5.649	5.741	5.579	5.708	5.730	5.510	5.737	5.696	5.751	5.708	5.680	5.621	5.820	5.725
Convergence Success Rate and Computation Time (Seconds)														
% of runs converged	96.2	96.4	97.0	97.2	98.4	98.4	98.4	86.4	91.8	94.4	94.0	91.4	86.2	83.4
Mean run time	119.8	137.9	140.4	169.1	179.1	206.5	217.8	184.1	215.5	259.7	299.8	329.9	351.2	401.4
Median run time	119.0	135.0	137.5	168.5	178.0	208.0	218.0	183.0	208.0	257.0	294.5	327.0	351.0	392.0
Standard deviation of run time	4.0	8.6	7.5	10.7	10.6	17.4	15.3	34.5	30.7	24.5	46.1	45.8	55.6	74.2

Table 3. Comparison of convergence rate and convergence time between radial and spherical parameterizations

Computation Time (Seconds) and Convergence Success Rate *Low Correlation*														
	Radial Parameterization							**Equivalent Spherical Parameterization**						
Scales	**0.20**	**0.60**	**0.80**	**1.00**	**1.20**	**1.40**	**2.00**	**0.32**	**0.97**	**1.29**	**1.61**	**1.93**	**2.26**	**3.22**
% of runs converged	96.2	96.4	97.0	97.2	98.4	98.4	98.4	96.0	96.2	96.2	97.2	98.0	97.6	99.0
Mean run time	119.8	137.9	140.4	169.1	179.1	206.5	217.8	121.2	136.5	143.5	170.5	183.7	206	219.4
Median run time	119.0	135.0	137.5	168.5	178.0	208.0	218.0	121.5	133.5	145.0	169.0	181.5	208.0	218.0
Standard deviation of run time	4.0	8.6	7.5	10.7	10.6	17.4	15.3	4.3	8.9	6.6	11.7	11.3	16.2	15.5
Computation Time (Seconds) and Convergence Success Rate *High Correlation*														
	Radial Parameterization							**Equivalent Spherical Parameterization**						
Scales	**0.20**	**0.60**	**0.80**	**1.00**	**1.20**	**1.40**	**2.00**	**0.32**	**0.97**	**1.29**	**1.61**	**1.93**	**2.26**	**3.22**
% of runs converged	86.4	91.8	94.4	94.0	91.4	86.2	83.4	86.4	93.2	94.6	95.4	88.8	84.2	81.8
Mean run time	184.1	215.5	259.7	299.8	329.9	351.2	401.4	187.5	213.7	261.7	298.1	325.2	351.6	386.2
Median run time	183.0	208.0	257.0	294.5	327.0	351.0	392.0	184.0	202.0	256.0	290.5	323.0	350.5	385.0
Standard deviation of run time	34.5	30.7	24.5	46.1	45.8	55.6	74.2	46.3	39.2	31.6	48.8	42.9	61.5	66.9