Technical Report 120

On Accommodating Spatial Interactions in a Generalized Heterogeneous Data Model (GHDM) of Mixed Types of Dependent Variables

Chandra Bhat
Subodh K. Dubey
Center for Transportation Research

Abdul R. Pinjari
University of South Florida

December 2015

Data-Supported Transportation Operations & Planning Center (D-STOP)

A Tier 1 USDOT University Transportation Center at The University of Texas at Austin

D-STOP is a collaborative initiative by researchers at the Center for Transportation Research and the Wireless Networking and Communications Group at The University of Texas at Austin.

Technical Report Documentation Page

1. Report No. D-STOP/2016/120	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle On Accommodating Spatial Interactions in a Generalized Heterogeneous Data Model (GHDM) of Mixed Types of Dependent Variables		5. Report Date December 2015 6. Performing Organization Code
7. Author(s) Chandra R. Bhat, Subodh K. Dubey, Abdul R. Pinjari		8. Performing Organization Report No. Report 120
9. Performing Organization Name and Address Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 1616 Guadalupe Street, Suite 4.202 Austin, Texas 78701		10. Work Unit No. (TRAIS) 11. Contract or Grant No. DTRT13-G-UTC58
12. Sponsoring Agency Name and Address Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 1616 Guadalupe Street, Suite 4.202 Austin, Texas 78701		13. Type of Report and Period Covered 14. Sponsoring Agency Code

15. Supplementary Notes
Supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program.

16. Abstract
We develop an econometric framework for incorporating spatial dependence in integrated model systems of latent variables and multidimensional mixed data outcomes. The framework combines Bhat's Generalized Heterogeneous Data Model (GHDM) with a spatial formulation to introduce spatial dependencies through latent constructs. Monte Carlo simulation experiments on synthetic data demonstrate the efficacy of the MACML approach in recovering parameters from spatially dependent datasets, as accurately and precisely as that from aspatial data (without spatial dependency). The results also suggest that ignoring spatial dependency can lead to a substantial loss in the accuracy and efficiency of parameter estimation and in overall data fit.

17. Key Words
Spatial econometrics, Multidimensional mixed data models, Latent variables, MACML estimation

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	32	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements

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.

.

Table of Contents

List of Tables

Chapter 1. Introduction

Multi-dimensional dependent outcome models are of interest in several fields, including land-use and transportation, biology, finance, and econometrics, just to name a few. The primary motivation for modeling dependent outcomes jointly is that there may be common underlying unobserved factors (attitudes, values, and lifestyle factors) of decision-makers that impact multiple dependent outcomes simultaneously. Ignoring the jointness and considering each dimension separately invites the pitfalls of (1) inefficient estimation of covariate effects for each outcome because such an approach fails to borrow information on other outcomes (Teixeira-Pinto and Harezlak, 2013), (2) multiple statistical testing requirements for specification analysis, which even then offer relatively poor statistical power in testing and poor control of type I error rates (De Leon and Zhu, 2008), and (3) inconsistent estimates of the structural effect of one endogenous variable on another (see Bhat and Guo, 2007). The last of these problems is particularly troubling, since it leads to what is typically referred to in the econometric literature as the "sample selection" or the "endogeneity in variables" problem. That is, modeling each outcome independently with a recursive pattern of influence among the independent outcomes is tantamount to a strictly sequential decision-making process, which is not consistent with the bundled (or package) nature of multiple outcomes. For example, in a land-use and transportation context, households that are environmentally conscious (and/or auto-averse in their lifestyle) may choose to locate in transit and pedestrian friendly neighborhoods that are characterized by high land use density (the word "auto" in this paper will be used to refer to motorized vehicles in the household). Then, a cross-sectional data set may indicate low auto ownership levels in high land use density areas, but at least part of this effect can be attributed to the purely associative effect of auto-averse households choosing to own fewer autos and residing in high density areas (rather than the low auto ownership being a sole causal effect of living in a high density neighborhood). Ignoring this issue will, in general, lead to a misleading conclusion about the causal effect of land-use on auto ownership, which can, in turn, lead to misinformed land-use policies. A way out to more accurately capture causal effects is to model the choice dimensions together in a joint equations modeling framework that accounts for correlated unobserved effects as well as possible causal inter-relationships between endogenous outcomes.

To be sure, there has been a substantial amount of work in the econometric literature on the simultaneous modeling of multiple continuous variables. However, there has been relatively little emphasis on multiple non-continuous variables (see De Leon and Chough, 2013). Bhat (2015a) provides a review of the many different approaches for modeling multiple and mixed data outcomes, and proposes a relatively general modeling framework, which he refers to as the General Heterogeneous Data Model (GHDM) system.

Even as there has been increasing emphasis on mixed data outcome modeling, there also has been a growing interest in accommodating spatial (and social) dependency effects among decision-makers. This is because spatial/social interactions can be exploited by decision-makers to achieve desired system end-states. As a simple illustration of this point, consider household auto ownership, and assume that the number of autos owned by

a household influences that of the household's residential neighbors. Then, a limited-funding information campaign to reduce auto dependency (and promote the use of non-motorized modes of transportation) would do well to target individuals from different neighborhoods, rather than targeting individuals from the same neighborhood. Doing so will benefit from the "ripple wave" (or spatial multiplier) effect caused by intra-neighborhood social exchanges, so that the aggregate-level effect of the information campaign on auto ownership can be substantial. Within the context of accommodating spatial dependencies, spatial lag and spatial error-type autoregressive structures developed for continuous dependent variables are being considered for non-continuous dependent outcomes (see reviews of this literature in Elhorst, 2010, Anselin, 2010, Ferdous and Bhat, 2013, Bhat *et al*., 2014, Bhat, 2014, and Bhat, 2015b).[1] Unfortunately, in the case of non-continuous outcomes, accommodating spatial dependence, in general, leads to multidimensional integration of the order of the number of decision-makers for count and ordered-response outcomes, and of the order of the number of decision-makers times the number of alternatives minus one for nominal (unordered-response) outcomes. Typical simulation-based methods, including the frequentist recursive importance sampling (RIS) estimator (which is a generalization of the more familiar Geweke-Hajivassiliou-Keane or GHK simulator; see Beron and Vijverberg, 2004). and the Bayesian Markov Chain Monte Carlo (MCMC)-based estimator (see LeSage and Pace, 2009), become impractical if not infeasible with moderate to large estimation sample sizes (see Bhat, 2011 and Smirnov, 2010). But, recently, Bhat and colleagues have suggested a composite marginal likelihood (CML) inference approach for estimating spatial binary/ordered-response probit/count models, and the maximum approximate composite marginal likelihood (MACML) inference approach for estimating spatial unordered-response multinomial probit (MNP) models (see Bhat, 2014 for a review). These methods are easy to implement, require no simulation, and involve only univariate and bivariate cumulative normal distribution function evaluations. However, all earlier spatial model studies, regardless of the estimation technique used, have focused on a single dependent outcome for each decision maker, rather than multiple and mixed dependent outcomes for each decision maker. On the other hand, when a host of dependent outcomes are co-determined because of common underlying unobserved factors or psychological constructs (attitudes, values, lifestyles, *etc*.), it is very likely that spatial dependence will exist not across just one of those outcomes but across all the outcomes.

In the current paper, we use the important insight that the analyst can generate spatial dependence across multiple and mixed outcomes by specifying spatial dependence in the "soft" psychological construct (latent) variables. In doing so, we combine the GHDM formulation with a spatial formulation. Then, since the mixed outcomes are specified to be a function of a much smaller set of the unobserved psychological constructs in measurement equations, it immediately generates spatial dependence across all outcomes. To our knowledge, this is the first study to propose such a methodological structure for

[1] Of course, the spatial lag and spatial error specifications can be combined together in a Kelejian-Prucha specification (see Elhorst, 2010), or the spatial lag could be combined with spatially lagged exogenous variable effects in a Spatial Durbin specification (see Bhat *et al*., 2014). In all of these cases, the spatial dependence leads also to spatial heteroscedasticity in the random error terms.

multiple mixed outcomes. At the same time, from a conceptual standpoint, we are able to better disentangle true causal effects from spurious self-selection effects (because the same unobserved factors impact multiple endogenous variables) and spatial dependence effects (because of diffusion of unobserved attitudes and lifestyles based on spatial proximity). Therefore, one can use the model to more accurately examine policy impacts that involve a combination of direct causal effects, self-selection effects, and spatial diffusion effects.

Section 2 presents the formulation of the spatial GHDM model along with the MACML estimation approach. Section 3 presents a simulation experiment to examine the ability of the MACML to accurately and precisely recover parameters in a spatial GHDM model. Section 4 concludes the paper.

Chapter 2. The Spatial GHDM Model Formulation

There are two components to the model: (1) the latent variable structural equation model (SEM) system, and (2) the latent variable measurement equation model system.

2.1 Latent Variable SEM System

Let l be the index for latent variables $l = (1,2,...,L)$ and q be the index for individuals $q = (1,2,...,Q)$. Then the latent variable z_{ql}^* may be written as a linear function of covariates using a spatial auto-correlation or spatial lag structure as follows:

$$z_{ql}^* = \boldsymbol{\alpha}_l' \boldsymbol{s}_q + \eta_{ql} + \delta_l \sum_{q'=1}^{Q} w_{qq'} z_{q'l}^* \tag{1}$$

where $\boldsymbol{s}_q$ is an $(F \times 1)$ vector of observed covariates (excluding a constant), $\boldsymbol{\alpha}_l$ is the corresponding $(F \times 1)$ vector of coefficients, η_{ql} is a random error term assumed to be distributed standard normal, $\delta_l (0 < \delta_l < 1)$ is the spatial autoregressive parameter, and $w_{qq'}$ is a spatial weight matrix with $w_{qq} = 0$ and $\sum_{q' \neq q}^{Q} w_{qq'} = 1 \ \forall q$. Next, define the following notations to write Equation (1) in matrix form for all Q individuals.

$$\boldsymbol{z}_q^* = (z_{q1}^*, z_{q2}^*, ..., z_{qL}^*)' \ [(L \times 1) \text{ vector}], \boldsymbol{z}^* = \left[(\boldsymbol{z}_1^*)', (\boldsymbol{z}_2^*)', ..., (\boldsymbol{z}_Q^*)'\right]' \ [(QL \times 1) \text{ vector}],$$

$$\widetilde{\boldsymbol{s}}_q = \mathbf{IDEN}_L \otimes \boldsymbol{s}_q' \ [(L \times LF) \text{ matrix}], \widetilde{\boldsymbol{s}} = (\widetilde{\boldsymbol{s}}_1', \widetilde{\boldsymbol{s}}_2', ..., \widetilde{\boldsymbol{s}}_Q')' \ [(QL \times LF) \text{ matrix}],$$

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', ..., \boldsymbol{\alpha}_L')' \ [(LF \times 1) \text{vector}], \ \boldsymbol{\eta}_q = (\eta_{q1}, \eta_{q2}, ..., \eta_{qL})' \ [(L \times 1) \text{vector}],$$

$$\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2', ..., \boldsymbol{\eta}_Q')' \ [(QL \times 1) \text{vector}], \ \boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_L)' \ [(L \times 1) \text{vector}],$$

$$\widetilde{\boldsymbol{\delta}} = \mathbf{1}_Q \otimes \boldsymbol{\delta} \ [(QL \times 1) \text{vector}],$$

To allow correlation among the latent variables of an individual, we assume a standard multivariate normal (MVN) distribution for $\boldsymbol{\eta}_q : \boldsymbol{\eta}_q \sim \text{MVN}_L[\mathbf{0}_L, \boldsymbol{\Gamma}]$, where $\mathbf{0}_L$ is an $(L \times 1)$ column vector of zeros, and $\boldsymbol{\Gamma}$ is the correlation matrix of size $(L \times L)$. We also assume $\boldsymbol{\eta}_q$ to be independent across individuals (*i.e.*, $\text{Cov}(\boldsymbol{\eta}_q, \boldsymbol{\eta}_{q'}) = 0, \forall q \neq q'$). With this, Equation (1) may be written in matrix form for all Q individuals as follows:

$$\boldsymbol{z}^* = \mathbf{S}\widetilde{\boldsymbol{s}}\boldsymbol{\alpha} + \mathbf{S}\boldsymbol{\eta} \tag{2}$$

where $\mathbf{S} = \left[\mathbf{IDEN}_{QL} - \widetilde{\boldsymbol{\delta}}.*(\mathbf{W} \otimes \mathbf{IDEN}_L)\right]^{-1} \ [(QL \times QL) \text{matrix}]$, "$\otimes$" represents the Kronecker product, "$.*$" represents the element by element product, $\mathbf{IDEN}_{QL}$ is an identity matrix of size QL, $\mathbf{1}_Q$ is a vector of size Q with all its elements equal to 1, and $\mathbf{W}$ is a $(Q \times Q)$ row normalized weight matrix. It is now easy to see that $\boldsymbol{z}^*$ is distributed

MVN with mean $\boldsymbol{B}$ and correlation matrix $\boldsymbol{\Xi}$. That is, $\boldsymbol{z}^* \sim \text{MVN}_{QL}(\boldsymbol{B}, \boldsymbol{\Xi})$, where $\boldsymbol{B} = \mathbf{S}\tilde{\boldsymbol{s}}\boldsymbol{\alpha}$ and $\boldsymbol{\Xi} = \mathbf{S}\left[\mathbf{IDEN}_Q \otimes \boldsymbol{\Gamma}\right]\mathbf{S}'$.

2.2 Latent Variable Measurement Equation Model System

We consider a combination of continuous, ordinal, count, and nominal outcomes (indicators) of the underlying latent variable vector $\boldsymbol{z}^*$. However, these outcomes may be a function of a set of exogenous variables too.

Let h be the index for continuous outcomes $(h = 1, 2, ..., H)$. Then the continuous variable y_{qh} can be written in the usual linear regression fashion as follows:

$$y_{qh} = \boldsymbol{\gamma}_h' \boldsymbol{x}_q + \boldsymbol{d}_h' \boldsymbol{z}_q^* + \varepsilon_{qh} \tag{3}$$

where $\boldsymbol{x}_q$ is an $(A \times 1)$ vector of exogenous variables (including a constant) as well as possibly the observed values of other endogenous variables (continuous, ordinal, count variable, and nominal variables (introduced as dummy variables)), $\boldsymbol{\gamma}_h$ is the corresponding vector of coefficients, $\boldsymbol{d}_h$ is an $(L \times 1)$ vector of latent variable loadings on the h^{th} continuous outcome, and ε_{qh} is a normally distributed random error term. Next, define the following notations to write Equation (3) in a compact, matrix form for individual q.

$$\mathbf{y}_q = (y_{q1}, y_{q2}, ..., y_{qH})' \ \left[(H \times 1) \text{ vector}\right], \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2', ..., \boldsymbol{\gamma}_H')' \ \left[(H \times A) \text{ matrix}\right],$$

$$\mathbf{d} = (\mathbf{d}_1', \mathbf{d}_2', ..., \mathbf{d}_H')' \ \left[(H \times L) \text{ matrix}\right], \text{ and } \boldsymbol{\varepsilon}_q = (\varepsilon_{q1}, \varepsilon_{q2}, ..., \varepsilon_{qH})' \ \left[(H \times 1) \text{ vector}\right].$$

Now, Equation (3) may be written in matrix form for individual q as follows:

$$\boldsymbol{y}_q = \boldsymbol{\gamma}\boldsymbol{x}_q + \boldsymbol{d}\boldsymbol{z}_q^* + \boldsymbol{\varepsilon}_q. \tag{4}$$

We assume a diagonal MVN distribution for $\boldsymbol{\varepsilon}_q$: $\boldsymbol{\varepsilon}_q \sim \text{MVN}_H(\mathbf{0}_H, \boldsymbol{\Sigma})$. The non-diagonal elements of $\boldsymbol{\varepsilon}_q$ are assumed to be zero for identification purposes. Also, the $\boldsymbol{\varepsilon}_q$ terms across different individuals are assumed independent of each other.

Next, consider N ordinal outcomes (indicators) and let n be an index for ordinal outcomes $(n = 1, 2, ..., N)$. Also, let J_n be the number of categories for the n^{th} ordinal outcome $(J_n \geq 2)$ and let the corresponding index be j_n $(j_n = 1, 2, ..., J_n)$. Let $\tilde{y}_{qn}^*$ be the latent underlying variable whose horizontal partitioning leads to the observed outcome a_{qn} for the q^{th} individual's n^{th} ordinal variable. Then, in the usual ordered response formulation, for the individual q, we may write:

$$\tilde{y}_{qn}^* = \tilde{\boldsymbol{\gamma}}_n' \boldsymbol{x}_q + \tilde{\boldsymbol{d}}_n' \boldsymbol{z}_q^* + \tilde{\varepsilon}_{qn}, \ \tilde{\psi}_{q,n,a_{qn}-1} < \tilde{y}_{qn}^* < \tilde{\psi}_{q,n,a_{qn}} \tag{5}$$

where $\boldsymbol{x}_q$ is as defined earlier, $\widetilde{y}_{qn}$ is the ordinal variable outcome category, $\widetilde{\boldsymbol{\gamma}}_n$ is the corresponding vector of coefficients, $\widetilde{\boldsymbol{d}}_n$ is an $(L \times 1)$ vector of latent variable loadings on the n^{th} ordinal outcome, and $\widetilde{\varepsilon}_{qn}$ is a normally distributed random error term. For each ordinal outcome, $\widetilde{\psi}_{q,n,0} < \widetilde{\psi}_{q,n,1} < \widetilde{\psi}_{q,n,2} ... < \widetilde{\psi}_{q,n,J_n-1} < \widetilde{\psi}_{q,n,J_n}$; $\widetilde{\psi}_{q,n,0} = -\infty$, $\widetilde{\psi}_{q,n,1} = 0$, and $\widetilde{\psi}_{q,n,J_n} = +\infty$. Next, define the following notation to write Equation (5) in a compact matrix form for individual q.

$$\widetilde{\boldsymbol{y}}_q^* = (\widetilde{y}_{q1}^*, \widetilde{y}_{q2}^*, ..., \widetilde{y}_{qN}^*)' \ \left[(N \times 1) \text{ vector}\right], \widetilde{\boldsymbol{\gamma}} = (\widetilde{\boldsymbol{\gamma}}_1', \widetilde{\boldsymbol{\gamma}}_2', ..., \widetilde{\boldsymbol{\gamma}}_N')' \ \left[(N \times A) \text{ matrix}\right],$$
$$\widetilde{\mathbf{d}} = (\widetilde{\mathbf{d}}_1', \widetilde{\mathbf{d}}_2', ..., \widetilde{\mathbf{d}}_N')' \ \left[(N \times L) \text{ matrix}\right], \ \widetilde{\boldsymbol{\varepsilon}}_q = (\widetilde{\varepsilon}_{q1}, \widetilde{\varepsilon}_{q2}, ..., \widetilde{\varepsilon}_{qN})' \ \left[(N \times 1) \text{ vector}\right].$$

Also, stack the lower thresholds for the observed outcomes a_{qn} of indiviudal q $\widetilde{\psi}_{q,n,a_{qn}-1} (n = 1, 2, ..., N)$ into an $(N \times 1)$ vector $\widetilde{\boldsymbol{\psi}}_{q,\boldsymbol{low}}$ and the corresponding upper thresholds $\widetilde{\psi}_{q,n,a_{qn}} (n = 1, 2, ..., N)$ into another vector $\widetilde{\boldsymbol{\psi}}_{q,\boldsymbol{up}}$.

Now, Equation (5) may be written in matrix form for individual q as follows:

$$\widetilde{\boldsymbol{y}}_q^* = \widetilde{\boldsymbol{\gamma}} \boldsymbol{x}_q + \widetilde{\boldsymbol{d}} \boldsymbol{z}_q^* + \widetilde{\boldsymbol{\varepsilon}}_q, \ \ \widetilde{\boldsymbol{\psi}}_{q,\boldsymbol{low}} < \widetilde{\boldsymbol{y}}_q^* < \widetilde{\boldsymbol{\psi}}_{q,\boldsymbol{up}}. \tag{6}$$

For identification, we assume a diagonal multivariate normal distribution for $\widetilde{\boldsymbol{\varepsilon}}_q$ with all the diagonal elements equal to unity: $\widetilde{\boldsymbol{\varepsilon}}_q \sim \text{MVN}_N(\mathbf{0}_N, \mathbf{IDEN}_N)$. In addition, the $\widetilde{\boldsymbol{\varepsilon}}_q$ terms are assumed to be independent across individuals.

Let there be C count variables and let c be an index for count outcomes $(c = 1, 2, ..., C)$. Let k_c be the index for count value $(k_c = 0, 1, 2, ..., \infty)$ and let r_{qc} be the actual observed count value. Then, following the recasting of a count model in a generalized ordered-response probit formulation (see Bhat, 2015a), a generalized version of the negative binomial count model may be written as:

$$\breve{y}_{qc}^* = \breve{\boldsymbol{d}}_c' \boldsymbol{z}_q^* + \breve{\varepsilon}_{qc}, \ \breve{\psi}_{q,c,r_{qc}-1} < \breve{y}_{qc}^* < \breve{\psi}_{q,c,r_{qc}} \ , \tag{7}$$

$$\breve{\psi}_{q,c,r_c} = \Phi^{-1}\left[\frac{(1-\upsilon_{qc})^{\theta_c}}{\Gamma(\theta_c)} \sum_{t=0}^{r_c} \left(\frac{\Gamma(\theta_c + t)}{t!} (\upsilon_{qc})^t \right) \right] + \varphi_{c,r_c}, \ \upsilon_{qc} = \frac{\lambda_{qc}}{\lambda_{qc} + \theta_{qc}}, \text{ and } \lambda_{qc} = e^{\breve{\boldsymbol{\gamma}}_c' \boldsymbol{x}_q}. \tag{8}$$

In the above equation, $\breve{y}_{qc}^*$ is a latent continuous stochastic propensity variable associated with the count variable c that maps into the observed count r_{qc} through the $\breve{\boldsymbol{\psi}}_{q,c}$ vector (which is a vertically stacked column vector of thresholds $(\breve{\psi}_{q,c,-1}, \breve{\psi}_{q,c,0}, \breve{\psi}_{q,c,1}, \breve{\psi}_{q,c,2}, ..)'$). $\breve{\boldsymbol{d}}_c$ is a $(L \times 1)$ vector of latent variable loadings on the c^{th} count outcome, and $\breve{\varepsilon}_{qc}$ is a standard normal random error term. $\breve{\boldsymbol{\gamma}}_c$ is a column vector of coefficients corresponding to the vector $\boldsymbol{x}_q$. θ_c is a parameter that provides flexibility to the count formulation, and

is related to the dispersion parameter in a traditional negative binomial model ($\theta_c > 0 \forall c$). $\Gamma(\theta_c)$ is the traditional gamma function; $\Gamma(\theta_c) = \int_{\tilde{t}=0}^{\infty} \tilde{t}^{\theta_c - 1} e^{-\tilde{t}}\, d\tilde{t}$. The threshold terms in the $\breve{\boldsymbol{\psi}}_{q,c}$ vector satisfy the ordering condition (*i.e.*, $\breve{\psi}_{q,c,-1} < \breve{\psi}_{q,c,0} < \breve{\psi}_{q,c,1} < \breve{\psi}_{q,c,2} .... < \infty \; \forall \; c$) as long as $\varphi_{c,-1} < \varphi_{c,0} < \varphi_{c,1} < \varphi_{c,2} .... < \infty$. The φ_c terms in the thresholds provide flexibility to accommodate high or low probability masses for specific count outcomes. For identification, we set $\varphi_{c,-1} = -\infty$ and $\varphi_{c,0} = 0$ for all count variables *c*. In addition, based on empirical testing, we identify a count value e_c^* ($e_c^* \in \{0,1,2,.....\}$) above which φ_{c,k_c} ($k_c \in \{1,2,.....\}$) is held fixed at φ_{c,e_c^*} Doing so allows the count model to predict beyond the range available in the estimation sample. For later use, let $\boldsymbol{\varphi}_c = (\varphi_{c,1}, \varphi_{c,2}, \ldots \varphi_{c,e_c^*})'$ ($e_c^* \times 1$ vector) (assuming $e_c^* > 0$), $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_1', \boldsymbol{\varphi}_2', \ldots \boldsymbol{\varphi}_C')' \left[\left(\sum_c e_c^* \right) \times 1 \text{ vector} \right]$, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_C)' \; [\text{C} \times 1 \text{ vector}]$.

Next define the following notation: $\breve{\mathbf{y}}_q^* = (\breve{y}_{q,1}^*, \breve{y}_{q,2}^*, ..., \breve{y}_{q,C}^*)' \; [(C \times 1) \text{ vector}]$, $\breve{\mathbf{d}} = (\breve{\mathbf{d}}_1', \breve{\mathbf{d}}_2', ..., \breve{\mathbf{d}}_C')' \; [(C \times L) \text{ matrix}]$, $\breve{\boldsymbol{\gamma}} = (\breve{\boldsymbol{\gamma}}_1', \breve{\boldsymbol{\gamma}}_2', ..., \breve{\boldsymbol{\gamma}}_C')' \; [(C \times A) \text{ matrix}]$, $\breve{\boldsymbol{\varepsilon}}_q = (\breve{\varepsilon}_{q1}, \breve{\varepsilon}_{q2}, ..., \breve{\varepsilon}_{qC})' \; [(C \times 1) \text{ vector}]$. Also, stack the lower thresholds of observed counts for the indiviudal q $\breve{\psi}_{q,c,r_{qc}-1} (c = 1, 2, ..., C)$ into a $(C \times 1)$ vector $\breve{\boldsymbol{\psi}}_{q,low}$ and the upper thresholds $\breve{\psi}_{q,c,r_{qc}} (c = 1, 2, ..., C)$ into another vector $\breve{\boldsymbol{\psi}}_{q,up}$. Now, the latent propensity underlying the count outcomes in Equation (7) may be written in matrix form as:

$$\breve{\boldsymbol{y}}_q^* = \breve{\boldsymbol{d}} \boldsymbol{z}_q^* + \breve{\boldsymbol{\varepsilon}}_q, \;\; \breve{\boldsymbol{\psi}}_{q,low} < \breve{\boldsymbol{y}}_q^* < \breve{\boldsymbol{\psi}}_{q,up} \tag{9}$$

Similar to ordinal variables we assume that the $\breve{\boldsymbol{\varepsilon}}_q$ terms are distributed as follows: $\breve{\boldsymbol{\varepsilon}}_q \sim \text{MVN}_C(\mathbf{0}_C, \mathbf{IDEN}_C)$, with independency across individuals.

Finally, let there be G nominal (unordered-response) variables, and let g be the index for the nominal variables (g = 1, 2, 3,…, G). Also, let I_g be the number of alternatives corresponding to the g^{th} nominal variable ($I_g \geq 3$) and let i_g be the corresponding index (i_g = 1, 2, 3,…, I_g). Consider the g^{th} nominal variable and assume that the individual q chooses the alternative $m_{q,g}$. Also, assume the usual random utility structure for each alternative i_g.

$$U_{qgi_g} = \boldsymbol{b}_{gi_g}' \boldsymbol{x}_q + \boldsymbol{\vartheta}_{gi_g}' (\boldsymbol{\beta}_{gi_g} \boldsymbol{z}_q^*) + \varsigma_{qgi_g}, \tag{10}$$

where $\boldsymbol{x}_q$ is as defined earlier, $\boldsymbol{b}_{gi_g}$ is a $(A \times 1)$ column vector of corresponding coefficients, and ς_{qgi_g} is a normal error term. $\boldsymbol{\beta}_{gi_g}$ is a $(N_{gi_g} \times L)$-matrix of variables

interacting with latent variables to influence the utility of alternative i_g, and $\boldsymbol{\vartheta}_{gi_g}$ is a $(N_{gi_g} \times 1)$-column vector of coefficients capturing the effects of latent variables and its interaction effects with other exogenous variables. Let $\boldsymbol{\varsigma_{qg}} = (\varsigma_{qg1}, \varsigma_{qg2}, ... \varsigma_{qgI_g})'$ ($I_g \times 1$ vector), with $\boldsymbol{\varsigma_{qg}} \sim MVN_{I_g}(\mathbf{0}, \boldsymbol{\Lambda_g})$ and independent across individuals. Taking the difference with respect to the first alternative, the only estimable elements correspond to the covariance matrix $\breve{\boldsymbol{\Lambda}}_g$ of these error differences, $\breve{\boldsymbol{\varsigma}}_{\boldsymbol{qg}} = (\breve{\varsigma}_{qg2}, \breve{\varsigma}_{qg3}, ..., \breve{\varsigma}_{qgI_g})$ (where $\breve{\varsigma}_{qgi} = \breve{\varsigma}_{qgi} - \breve{\varsigma}_{qg1}, \forall i \neq 1$). Further, the variance term at the top left diagonal of $\breve{\boldsymbol{\Lambda}}_g$ (g=1,2,…,G) is set to 1 to account for scale invariance. $\boldsymbol{\Lambda_g}$ is constructed from $\breve{\boldsymbol{\Lambda}}_{\mathbf{g}}$ by adding a row of zeros on top and a column of zeros to the left. To proceed, define $\boldsymbol{U}_{qg} = (U_{qg1}, U_{qg2}, ..., U_{qgI_g})'$ ($I_g \times 1$ vector), $\boldsymbol{b}_g = (\boldsymbol{b_{g1}}, \boldsymbol{b_{g2}}, \boldsymbol{b_{g3}}, ..., \boldsymbol{b_{gI_g}})'$ ($I_g \times A$ matrix), and $\boldsymbol{\beta}_g = (\boldsymbol{\beta}'_{g1}, \boldsymbol{\beta}'_{g2}, ..., \boldsymbol{\beta}'_{gI_g})'$ $\left(\sum_{i_g=1}^{I_g} N_{gi_g} \times L \right)$ matrix. Also, define the $\left(I_g \times \sum_{i_g=1}^{I_g} N_{gi_g} \right)$ matrix $\boldsymbol{\vartheta_g}$, which is initially filled with all zero values. Then, position the $(1 \times N_{g1})$ row vector $\boldsymbol{\vartheta}'_{g1}$ in the first row to occupy columns 1 to N_{g1}, position the $(1 \times N_{g2})$ row vector $\boldsymbol{\vartheta}'_{g2}$ in the second row to occupy columns N_{g1}+1 to $N_{g1} + N_{g2}$, and so on until the $(1 \times N_{gI_g})$ row vector $\boldsymbol{\vartheta}'_{gI_g}$ is appropriately positioned. Further, define $\boldsymbol{\varpi}_g = (\boldsymbol{\vartheta}_g \boldsymbol{\beta}_g)$ ($I_g \times L$ matrix), $\ddot{G} = \sum_{g=1}^{G} I_g$, $\widetilde{G} = \sum_{g=1}^{G} (I_g - 1)$, $\boldsymbol{U_q} = \left(\boldsymbol{U}'_{q1}, \boldsymbol{U}'_{q2}, ..., \boldsymbol{U}'_{qG} \right)'$ ($\ddot{G} \times 1$ vector), $\varsigma_q = (\varsigma_{q1}, \varsigma_{q2}, ..., \varsigma_{qG})'$ ($\ddot{G} \times 1$ vector), $\boldsymbol{b} = (\boldsymbol{b'_1}, \boldsymbol{b'_2}, ..., \boldsymbol{b'_G})'$ ($\ddot{G} \times A$ matrix), $\boldsymbol{\varpi} = (\boldsymbol{\varpi}'_1, \boldsymbol{\varpi}'_2, ..., \boldsymbol{\varpi}'_G)'$ ($\ddot{G} \times L$ matrix), and $\boldsymbol{\vartheta} = \text{Vech}(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, ..., \boldsymbol{\vartheta}_G)$ (that is, $\boldsymbol{\vartheta}$ is a column vector that includes all elements of the matrices $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, ..., \boldsymbol{\vartheta}_G$). Then, in matrix form, we may write Equation (10) for individual q as:

$$\boldsymbol{U}_q = \boldsymbol{bx}_q + \boldsymbol{\varpi z}_q^* + \varsigma_q, \tag{11}$$

where $\varsigma_q \sim \text{MVN}_{\ddot{G}}(\mathbf{0}_{\ddot{G}}, \boldsymbol{\Lambda})$. As earlier, to ensure identification, we specify $\boldsymbol{\Lambda}$ as follows:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Lambda}_3 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Lambda}_G \end{bmatrix} (\ddot{G} \times \ddot{G} \text{ matrix}), \tag{12}$$

2.3 Reduced Form Model System

Let $E=(H+N+C)$ and $\widetilde{E}=(N+C+\widetilde{G})$. Define $\vec{\boldsymbol{y}}_q=\left(\boldsymbol{y}_q',\left[\widetilde{\boldsymbol{y}}_q^{\,*}\right]',\left[\breve{\boldsymbol{y}}_q^{\,*}\right]'\right)'$ [$E\times 1$ vector], $\vec{\boldsymbol{\gamma}}=(\boldsymbol{\gamma}',\widetilde{\boldsymbol{\gamma}}',\mathbf{0}_{AC})'$ [$E\times A$ matrix], $\vec{\boldsymbol{d}}=(\boldsymbol{d}',\widetilde{\boldsymbol{d}}',\breve{\boldsymbol{d}}')'$ [$E\times L$ matrix], and $\vec{\boldsymbol{\varepsilon}}_q=(\boldsymbol{\varepsilon}_q',\widetilde{\boldsymbol{\varepsilon}}_q',\breve{\boldsymbol{\varepsilon}}_q')'$ ($E\times 1$ vector), where $\mathbf{0}_{AC}$ is a matrix of zeros of dimension $A\times C$. Then, the equations for continuous, ordinal, and count endogenous variables (*i.e.*, Equations 4, 6 , and 9) of individual q may be brought together as follows:

$$\vec{\boldsymbol{y}}_q=\vec{\boldsymbol{\gamma}}\boldsymbol{x}_q+\vec{\boldsymbol{d}}\boldsymbol{z}_q^{*}+\vec{\boldsymbol{\varepsilon}}_q,\ \text{with}\ \ \text{Var}\left(\vec{\boldsymbol{\varepsilon}}_q\right)=\vec{\boldsymbol{\Sigma}}=\begin{bmatrix}\boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{IDEN}_{\mathbf{N}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{IDEN}_{\mathbf{C}}\end{bmatrix}\ (E\times E\ \text{matrix}) \qquad (13)$$

To combine the above equation with Equation (11) for nominal endogenous variables ($\boldsymbol{U}_q$), define $(\boldsymbol{yU})_q=\left[\vec{\boldsymbol{y}}_q',\boldsymbol{U}_q'\right]'$ $\left[(E+\vec{\mathrm{G}})\times 1\text{vector}\right]$, $\vec{\boldsymbol{b}}=(\vec{\boldsymbol{\gamma}}',\boldsymbol{b}')'$ $\left[(E+\vec{\mathrm{G}})\times A\ \text{matrix}\right]$, $\vec{\boldsymbol{c}}=(\vec{\boldsymbol{d}}',\boldsymbol{\varpi}')'$ $\left[(E+\vec{\mathrm{G}})\times L\ \text{matrix}\right]$, and $\boldsymbol{\xi}_q=(\vec{\boldsymbol{\varepsilon}}_q',\boldsymbol{\varsigma}_q')'$ $\left[(E+\vec{\mathrm{G}})\times 1\ \text{vector}\right]$. Then, the equations for all endogenous variables in the overall model system for individual q may be written compactly as:

$$(\boldsymbol{yU})_q=\vec{\boldsymbol{b}}\boldsymbol{x}_q+\vec{\boldsymbol{c}}\,\boldsymbol{z}_q^{*}+\boldsymbol{\xi}_q,\text{with}\ \ \text{Var}\left(\boldsymbol{\xi}_q\right)=\vec{\boldsymbol{\Sigma}}=\begin{bmatrix}\vec{\boldsymbol{\Sigma}} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}\end{bmatrix}\ \left[(E+\vec{G})\times(E+\vec{G})\ \text{matrix}\right] \qquad (14)$$

Now, the above Equation (14) for an individual q may be used to write a compact expression of endogenous variable equations for all Q individuals as:

$$\boldsymbol{yU}=\vec{\boldsymbol{b}}\boldsymbol{x}+\vec{\boldsymbol{c}}\boldsymbol{z}^{*}+\boldsymbol{\xi} \qquad (15)$$

where, $\boldsymbol{yU}=\left[(\boldsymbol{yU})_1',(\boldsymbol{yU})_2',\ldots,(\boldsymbol{yU})_Q'\right]'$ $\left[Q(E+\vec{\mathrm{G}})\times 1\ \text{vector}\right]$, $\boldsymbol{\xi}=(\boldsymbol{\xi}_1',\boldsymbol{\xi}_2',\ldots,\boldsymbol{\xi}_Q')'$ $\left[Q(E+\vec{\mathrm{G}})\times 1\ \text{vector}\right]$, $\boldsymbol{x}=(\boldsymbol{x}_1',\boldsymbol{x}_2',\ldots,\boldsymbol{x}_Q')'$ [$QA\times 1$ vector], $\vec{\boldsymbol{b}}=\mathbf{IDEN}_Q\otimes\vec{\boldsymbol{b}}$ $\left[Q(E+\vec{\mathrm{G}})\times QA\ \text{matrix}\right]$, and $\vec{\mathbf{c}}=(\mathbf{IDEN}_Q\otimes\vec{\mathbf{c}})$ $\left[Q(E+\vec{\mathrm{G}})\times QL\ \text{matrix}\right]$.

To develop the reduced form model system, substitute the right side of structural Equation (2) in the above equation, as below:

$$\begin{aligned}\boldsymbol{yU}&=\vec{\boldsymbol{b}}\boldsymbol{x}+\vec{\boldsymbol{c}}\left[\mathbf{S}\tilde{\boldsymbol{s}}\boldsymbol{\alpha}+\mathbf{S}\boldsymbol{\eta}\right]+\boldsymbol{\xi}\\ &=\vec{\boldsymbol{b}}\boldsymbol{x}+\vec{\boldsymbol{c}}\left[\boldsymbol{B}+\mathbf{S}\boldsymbol{\eta}\right]+\boldsymbol{\xi}\\ &=(\vec{\boldsymbol{b}}\boldsymbol{x}+\vec{\boldsymbol{c}}\boldsymbol{B})+(\vec{\boldsymbol{c}}\mathbf{S}\boldsymbol{\eta}+\boldsymbol{\xi})\end{aligned} \qquad (16)$$

Then, $\boldsymbol{yU}\sim\text{MVN}_{Q(E+\vec{G})}\left[(\vec{\boldsymbol{b}}\boldsymbol{x}+\vec{\boldsymbol{c}}\boldsymbol{B}),(\vec{\boldsymbol{c}}\boldsymbol{\Xi}\vec{\boldsymbol{c}}'+\mathbf{IDEN}_Q\otimes\vec{\boldsymbol{\Sigma}})\right]$

2.4 Model Estimation

Let $\boldsymbol{\lambda}$ be the collection of parameters to be estimated: $\boldsymbol{\lambda} = [\text{Vech}(\boldsymbol{\alpha}), \text{Vech}(\vec{\boldsymbol{\Sigma}}), \text{Vech}(\vec{\boldsymbol{b}}), \text{Vech}(\vec{\boldsymbol{c}}), \boldsymbol{\varphi}, \boldsymbol{\theta}, \boldsymbol{\delta}]$, where the operator "Vech(.)" vectorizes all the elements of the matrix/vector on which it operates. The identification issues pertaining to the estimability of these parameters in the current spatial-GHDM are the same as those discussed in Bhat (2015a) for the aspatial-GHDM, with the addition of the requirement that all elements of the vector $\boldsymbol{\delta}$ should be bounded in magnitude by the value of 1 (see Sidharthan and Bhat, 2012).

To estimate the model, we work with the latent utility differentials $u_{qgi_g m_{qg}} = \left(U_{qgi_g} - U_{qgm_{qg}}\right)$ of all non-chosen alternatives ($i_g \neq m_{qg}$) with respect to the chosen alternative (m_{qg}) for each nominal variable g and each individual q. Stack the utility differentials into a vector $\boldsymbol{u}_{qg} = \left[\left(u_{qg1m_{qg}}, u_{qg2m_{qg}}, ..., u_{qgI_g m_{qg}}\right)'; i_g \neq m_g\right]$ and then into $\boldsymbol{u}_q = \left(\left[\boldsymbol{u}_{q1}\right]', \left[\boldsymbol{u}_{q2}\right]', ..., \left[\boldsymbol{u}_{qG}\right]'\right)'$. Also, define $(\boldsymbol{yu})_q = \left[\ddot{\boldsymbol{y}}_q', \boldsymbol{u}_q'\right]' \quad \left[(E+\tilde{G})\times 1 \text{ vector}\right]$ and $\boldsymbol{yu} = \left[(\boldsymbol{yu})_1', (\boldsymbol{yu})_2', ..., (\boldsymbol{yu})_Q'\right]' \quad \left[Q(E+\tilde{G})\times 1 \text{ vector}\right]$. The distribution of the vector $\boldsymbol{yu}$ may be developed from that of $\boldsymbol{yU}$ using a matrix $\mathbf{M}$ of size $\left[Q(E+\tilde{G})\times Q(E+\vec{G})\right]$, constructed as discussed in Bhat (2015a). Then the resulting distribution is $\boldsymbol{yu} \sim \text{MVN}_{Q(E+\tilde{G})}\left[\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{\Omega}}\right]$, where $\tilde{\boldsymbol{B}} = \mathbf{M}(\vec{\boldsymbol{b}}\boldsymbol{x} + \vec{\boldsymbol{c}}\boldsymbol{B})$ and $\tilde{\boldsymbol{\Omega}} = \mathbf{M}(\vec{\boldsymbol{c}}\boldsymbol{\Xi}\vec{\boldsymbol{c}}' + \mathbf{IDEN}_Q \otimes \vec{\boldsymbol{\Sigma}})\mathbf{M}'$.

Next, partition $\boldsymbol{yu}$ into two components – one that corresponds to all the continuous variables ($\boldsymbol{y}$) and the other that corresponds to all the ordinal, count, and nominal variables ($\tilde{\boldsymbol{y}}^*, \breve{\boldsymbol{y}}^*, \boldsymbol{u}$ (utility differences)). That is, $\boldsymbol{yu} = (\boldsymbol{y}', \tilde{\boldsymbol{u}}')'$, where $\tilde{\boldsymbol{u}} = \left(\tilde{\boldsymbol{y}}^{*\prime}, \breve{\boldsymbol{y}}^{*\prime}, \boldsymbol{u}'\right)'$. Accordingly, the mean vector $\tilde{\boldsymbol{B}}$ and the variance matrix $\tilde{\boldsymbol{\Omega}}$ of $\boldsymbol{yu}$ can also be appropriately partitioned as: $\tilde{\boldsymbol{B}} = \begin{bmatrix} \tilde{\boldsymbol{B}}_y \\ \tilde{\boldsymbol{B}}_{\tilde{u}} \end{bmatrix}$ and $\tilde{\boldsymbol{\Omega}} = \begin{bmatrix} \tilde{\boldsymbol{\Omega}}_y & \tilde{\boldsymbol{\Omega}}_{y\tilde{u}} \\ \tilde{\boldsymbol{\Omega}}_{y\tilde{u}}' & \tilde{\boldsymbol{\Omega}}_{\tilde{u}} \end{bmatrix}$.

One may develop the likelihood function by decomposing the joint distribution of $\boldsymbol{yu} = (\boldsymbol{y}', \tilde{\boldsymbol{u}}')'$ into a product of marginal and conditional distributions. Specifically, the conditional distribution of $\tilde{\boldsymbol{u}}$, given $\boldsymbol{y}$, is MVN with mean $\ddot{\boldsymbol{B}}_{\tilde{u}} = \tilde{\boldsymbol{B}}_{\tilde{u}} + \tilde{\boldsymbol{\Omega}}_{y\tilde{u}}'\tilde{\boldsymbol{\Omega}}_y^{-1}(\boldsymbol{y} - \tilde{\mathbf{B}}_y)$ and variance $\ddot{\boldsymbol{\Omega}}_{\tilde{u}} = \tilde{\boldsymbol{\Omega}}_{\tilde{u}} - \tilde{\boldsymbol{\Omega}}_{y\tilde{u}}'\tilde{\boldsymbol{\Omega}}_y^{-1}\tilde{\boldsymbol{\Omega}}_{y\tilde{u}}$. Furthermore, define the threshold vectors as:

$\ddot{\boldsymbol{\psi}}_{low} = \left[\tilde{\boldsymbol{\psi}}_{low}', \breve{\boldsymbol{\psi}}_{low}', \left(-\boldsymbol{\infty}_{Q\tilde{G}}\right)'\right]'$ ($Q\tilde{E}\times 1$ vector) and $\ddot{\boldsymbol{\psi}}_{up} = \left[\tilde{\boldsymbol{\psi}}_{up}', \breve{\boldsymbol{\psi}}_{up}', \left(\mathbf{0}_{Q\tilde{G}}\right)'\right]'$ ($Q\tilde{E}\times 1$ vector),

where $-\boldsymbol{\infty}_{Q\tilde{G}}$ is a $Q\tilde{G}\times 1$-column vector of negative infinities, $\mathbf{0}_{Q\tilde{G}}$ is another $Q\tilde{G}\times 1$-column vector of zeros, and $\tilde{\boldsymbol{\psi}}_{low} = (\tilde{\boldsymbol{\psi}}_{1,low}', \tilde{\boldsymbol{\psi}}_{2,low}', ..., \tilde{\boldsymbol{\psi}}_{Q,low}')'$ ($QN\times 1$ vector),

$\tilde{\boldsymbol{\psi}}_{\boldsymbol{up}} = (\tilde{\boldsymbol{\psi}}'_{1,\boldsymbol{up}}, \tilde{\boldsymbol{\psi}}'_{2,\boldsymbol{up}}, \ldots, \tilde{\boldsymbol{\psi}}'_{Q,\boldsymbol{up}})'$ $(QN \times 1 \text{ vector})$,
$\tilde{\boldsymbol{\psi}}_{\boldsymbol{low}} = (\tilde{\boldsymbol{\psi}}'_{1,\boldsymbol{low}}, \tilde{\boldsymbol{\psi}}'_{2,\boldsymbol{low}}, \ldots, \tilde{\boldsymbol{\psi}}'_{Q,\boldsymbol{low}})'$ $(QC \times 1 \text{ vector})$, and
$\tilde{\boldsymbol{\psi}}_{\boldsymbol{up}} = (\tilde{\boldsymbol{\psi}}'_{1,\boldsymbol{up}}, \tilde{\boldsymbol{\psi}}'_{2,\boldsymbol{up}}, \ldots, \tilde{\boldsymbol{\psi}}'_{Q,\boldsymbol{up}})'$ $(QC \times 1 \text{ vector})$. Then the likelihood function may be written as:

$$L(\boldsymbol{\lambda}) = f_{QH}(\boldsymbol{y} \mid \tilde{\boldsymbol{B}}_{\boldsymbol{y}}, \tilde{\boldsymbol{\Omega}}_{\boldsymbol{y}}) \times \Pr\left[\tilde{\boldsymbol{\psi}}_{\boldsymbol{low}} \le \tilde{\boldsymbol{u}} \le \tilde{\boldsymbol{\psi}}_{\boldsymbol{up}}\right], \tag{17}$$
$$= f_{QH}(\boldsymbol{y} \mid \tilde{\boldsymbol{B}}_{\boldsymbol{y}}, \tilde{\boldsymbol{\Omega}}_{\boldsymbol{y}}) \times \int_{D_r} f_{Q\tilde{E}}(\boldsymbol{r} \mid \tilde{\boldsymbol{B}}_{\tilde{\boldsymbol{u}}}, \tilde{\boldsymbol{\Omega}}_{\tilde{\boldsymbol{u}}})\, dr.$$

In the above expression, $f_{QH}(\boldsymbol{y} \mid \tilde{\boldsymbol{B}}_{\boldsymbol{y}}, \tilde{\boldsymbol{\Omega}}_{\boldsymbol{y}})$ is a multivariate density function of dimension QH and $\Pr\left[\tilde{\boldsymbol{\psi}}_{\boldsymbol{low}} \le \tilde{\boldsymbol{u}} \le \tilde{\boldsymbol{\psi}}_{\boldsymbol{up}}\right]$ is a $Q\tilde{E}$-dimensional rectangular integral. Evaluation of such high dimensional integrals is infeasible with techniques currently available in the literature, as discussed earlier in Section 1. A possible solution to this problem is to use the composite marginal likelihood (CML) approach. In the CML approach, the maximizing function is developed as the product of low dimensional marginal densities (see Bhat, 2014 for a detailed description of the CML approach). For the spatial-GHDM model, the CML function may be written as a product of pairwise marginal densities, across all pairs of individuals, as follows:

$$L_{CML}(\boldsymbol{\lambda}) = \prod_{q=1}^{Q-1} \prod_{q'=q+1}^{Q} f_{2*H}(\boldsymbol{y}_{qq'} \mid \tilde{\boldsymbol{B}}_{qq',\boldsymbol{y}}, \tilde{\boldsymbol{\Omega}}_{qq',\boldsymbol{y}}) \times \Pr\left[\tilde{\boldsymbol{\psi}}_{qq',\boldsymbol{low}} \le \tilde{\boldsymbol{u}}_{qq'} \le \tilde{\boldsymbol{\psi}}_{qq',\boldsymbol{up}}\right] \tag{18}$$

In the above expression, $f_{2*H}(\boldsymbol{y}_{qq'} \mid \tilde{\boldsymbol{B}}_{qq',\boldsymbol{y}}, \tilde{\boldsymbol{\Omega}}_{qq',\boldsymbol{y}})$ is an MVN density function of dimension $2H$ and $\Pr\left[\tilde{\boldsymbol{\psi}}_{qq',\boldsymbol{low}} \le \tilde{\boldsymbol{u}}_{qq'} \le \tilde{\boldsymbol{\psi}}_{qq',\boldsymbol{up}}\right]$ is a $2\tilde{E}$-dimensional MVN integral. Thus, the CML approach reduces the dimensionality of integration from $Q\tilde{E}$ to $2\tilde{E}$, which can then be evaluated using the analytic approximation embedded in the MACML approach.

Chapter 3. Simulation Evaluation

In this section, we present the design of, and results from, a simulation framework to evaluate the MACML approach in terms of its parameter recovery and standard error estimation from the spatial GHDM. In addition, we provide an assessment of the potential repercussions of ignoring spatial dependency when it is present.

3.1 Design of the GHDM

For simulating the GHDM in this study, we use the same GHDM design described in Bhat (2015a). The only exception is that we include spatial dependency in the latent constructs that, in turn, generates spatial dependency in a variety of different mixed data outcomes of the GHDM through the influence of latent variables on those outcomes. The advantage of building on Bhat's aspatial GHDM setup is it provides an opportunity to compare the spatial model results with the aspatial model results in his paper. Thus, to conserve space, we provide a brief discussion of the simulation setup. The reader is referred to Bhat (2015a) for a detailed description of the simulation design.

3.1.1 Design of the Latent Variable SEM System

Consider two latent constructs: (1) *green lifestyle propensity* (*GLP*) (z_1^*) and (2) *travel freedom/privacy affinity* (TFA) (z_2^*). The first factor GLP reflects an individual's level of environmental consciousness, and specified as a function of two exogenous variables—individual's education level (s_1; $s_1 = 1$ if the individual has a bachelor's degree or higher and 0 otherwise) and gender (s_2; $s_2 = 1$ if the individual is a male adult and 0 otherwise). The second factor TFA reflects an individual's preference for privacy and a desire for control over the travel experience, and specified as a function of gender (s_2) and household income (s_3; $s_3 = 1$ if household annual income is at least \$75,000 and 0 otherwise). Below is the vector notation for the simulated SEM system, along with additional details:

$$\begin{bmatrix} z_1^* = \text{GLP} \\ z_2^* = \text{TFA} \end{bmatrix} = \begin{bmatrix} 0.8 & -0.3 & 0.0 \\ 0.0 & 0.2 & 0.5 \end{bmatrix} \times \begin{bmatrix} s_1 = \text{At least bachelor's degree} \\ s_2 = \text{Male} \\ s_3 = \text{High income} \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}, \quad (19)$$

where, $\text{Var}\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \boldsymbol{\Gamma} = \begin{bmatrix} 1.0 & \Gamma_{12} \\ \Gamma_{12} & 1.0 \end{bmatrix}$, the coefficients in the SEM are α_{11} = 0.8, α_{12} = -0.3, α_{22} = 0.2, and α_{23} = 0.5, and the correlation Γ_{12} between the two latent constructs is -0.6.

3.1.2 Design of the Measurement Equation System

The complete measurement equation system simulated for the GHDM is shown in Equation system (20), which combines the non-nominal equation system $\vec{\boldsymbol{y}}_q = \vec{\boldsymbol{\gamma}}\boldsymbol{x}_q + \vec{\boldsymbol{d}}\boldsymbol{z}_q^* + \vec{\boldsymbol{\varepsilon}}_q$ as well as the nominal equation system $\boldsymbol{U}_q = \boldsymbol{b}\boldsymbol{x}_q + \boldsymbol{\varpi}_q \boldsymbol{z}_q^* + \boldsymbol{\varsigma}_q$. The design is the same as in Bhat (2015a), with the exception that we include spatial dependency in the latent constructs that, in turn, generates spatial dependency in all the mixed data outcomes. The advantage of building on Bhat's aspatial GHDM setup is that it provides an opportunity to compare the spatial model results with the aspatial model results in his paper. A total of five non-nominal variables and two nominal variables are considered in this system. The first non-nominal variable is the individual's (log) commute distance (y_1). The next three non-nominal variables are ordinal (with a three point ordinal scale) and considered as a quantification of non-commute travel by different modes. These variables are: (1) weekly extent of non-commute travel by non-motorized transport (NM), modeled using a latent propensity variable $\widetilde{y}_1^*$; (2) weekly extent of non-commute travel by public transit (PT), modeled using a latent propensity variable $\widetilde{y}_2^*$; and (3) weekly extent of non-commute travel by motorized transport (MT), modeled using a latent propensity variable $\widetilde{y}_3^*$. The fifth non-nominal variable is a count variable labeled vehicle ownership, modeled using a latent propensity of vehicle ownership ($\breve{y}_1^*$). The nominal variables considered in the measurement equation system are: (1) residential location choice, with three alternatives – urban, suburban, and rural, and (2) commute mode choice, with three alternatives – motorized transport (MT), public transit (PT), and non-motorized modes (NM). The *utility* variables used to model these choices are $U_{1,urban}$, $U_{1,suburban}$, and $U_{1,rural}$ for residential location, and $U_{2,MT}$, $U_{2,PT}$, and $U_{2,NM}$ for commute mode choice.

The exogenous variables considered in the measurement system include: (1) Immigrant (binary variable), (2) Own house (binary variable), (3) No. of children less than 11 years in the household, and (4) No. of young adults (between 18-30yrs) in the household. In addition to the effects of these exogenous variables on various endogenous variables, Equation (20) reflects structural (*i.e.*, causal) relationships among several endogenous variables. These include the effects of: (a) commute distance on the utility of the NM commute mode; (b) urban residential location on commute distance, non-commute travel by the NM and PT modes, and on vehicle ownership; and (c) vehicle ownership on the utility of the NM commute mode.

$$
\begin{bmatrix}
y_1 = \log(\text{commute distance}) \\
\widetilde{y}_1^* = \text{NC propensity by NM} \\
\widetilde{y}_2^* = \text{NC propensity by PT} \\
\widetilde{y}_3^* = \text{NC propensity by MT} \\
\breve{y}_1^* = \text{auto own. propensity} \\
U_{1,urban} \\
U_{1,\text{suburban}} \\
U_{1,\text{rural}} \\
U_{2,\text{MT}} \\
U_{2,\text{PT}} \\
U_{2,\text{NM}}
\end{bmatrix}
=
\begin{bmatrix}
\gamma_{11} & \gamma_{12} & 0 & 0 & 0 & 0 & 0 & \gamma_{18} \\
\widetilde{\gamma}_{11} & 0 & 0 & \widetilde{\gamma}_{14} & 0 & 0 & 0 & \widetilde{\gamma}_{18} \\
\widetilde{\gamma}_{21} & 0 & 0 & 0 & 0 & 0 & 0 & \widetilde{\gamma}_{28} \\
\widetilde{\gamma}_{31} & 0 & 0 & \widetilde{\gamma}_{34} & \widetilde{\gamma}_{35} & 0 & 0 & 0 \\
\breve{\gamma}_{11} & 0 & 0 & 0 & 0 & 0 & 0 & \breve{\gamma}_{18} \\
b_{111} & b_{112} & b_{113} & 0 & 0 & 0 & 0 & 0 \\
b_{121} & 0 & 0 & b_{124} & b_{125} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b_{221} & b_{222} & 0 & 0 & 0 & 0 & 0 & b_{228} \\
b_{231} & 0 & 0 & 0 & 0 & b_{236} & b_{237} & 0
\end{bmatrix}
\times
\begin{bmatrix}
\text{Constant} \\
\text{Immigrant} \\
\text{Own house} \\
\#\ \text{children} < 11\ \text{yrs} \\
\#\ \text{young adults} \\
\text{Commute distance} \\
\text{Vehicle ownership} \\
\text{Urban dwelling}
\end{bmatrix}
$$

$$
+
\begin{bmatrix}
0 & d_{12} \\
\widetilde{d}_{11} & 0 \\
\widetilde{d}_{21} & 0 \\
0 & \widetilde{d}_{32} \\
\breve{d}_{11} & \breve{d}_{12} \\
\varpi_{111} & 0 \\
0 & 0 \\
0 & 0 \\
0 & \varpi_{212} \\
\varpi_{221} & 0 \\
\varpi_{231} & 0
\end{bmatrix}
\times
\begin{bmatrix}
Z_1^* = \text{GLP} \\
Z_2^* = \text{TFA}
\end{bmatrix}
+
\begin{bmatrix}
\varepsilon_1 \\
\widetilde{\varepsilon}_1 \\
\widetilde{\varepsilon}_2 \\
\widetilde{\varepsilon}_3 \\
\breve{\varepsilon}_1 \\
\varsigma_{11} \\
\varsigma_{12} \\
\varsigma_{13} \\
\varsigma_{21} \\
\varsigma_{22} \\
\varsigma_{23}
\end{bmatrix}
\qquad (20)
$$

Note from Equation (20) that at least one (or both) of the latent constructs is loaded onto each of the endogenous variables, so that incorporating spatial dependence in the latent constructs automatically leads to spatial dependence among all the endogenous variables. The design values of the latent construct effects (embedded in $\vec{\boldsymbol{d}}$ and $\boldsymbol{\varpi}$) and those for exogenous and endogenous outcome effects (embedded in $\tilde{\boldsymbol{\gamma}}$ and $\breve{\boldsymbol{\gamma}}$ for the non-nominal system and in $\boldsymbol{b}$ for the nominal system) are:

$\text{Vech}(\tilde{\boldsymbol{\gamma}}) = [\gamma_{11} = 1,\ \gamma_{12} = 0.5,\ \gamma_{18} = -0.3,\ \tilde{\gamma}_{11} = 1,\ \tilde{\gamma}_{14} = -0.2,\ \tilde{\gamma}_{18} = 0.6,\ \tilde{\gamma}_{21} = 1,\ \tilde{\gamma}_{28} = 0.2,\ \tilde{\gamma}_{31} = 1,\ \tilde{\gamma}_{34} = 0.4,\ \tilde{\gamma}_{35} = -0.3,\ \breve{\gamma}_{11} = 1.0,\ \breve{\gamma}_{18} = -0.5]$,

$\text{Vech}(\boldsymbol{b}) = [b_{111} = 0.2, b_{112} = 0.4, b_{113} = -0.5, b_{121} = 0.3,$

$b_{123} = 0.2, b_{124} = 0.3, b_{221} = -0.5, b_{222} = 0.3, b_{228} = 0.2, b_{231} = -0.2, b_{236} = -0.6, b_{237} = -0.4]$,

$\text{Vech}(\vec{\boldsymbol{d}}) = [d_{12} = 0.2, \tilde{d}_{11} = 0.6, \tilde{d}_{21} = 0.2, \tilde{d}_{32} = 0.3, \breve{d}_{11} = -0.5, \breve{d}_{12} = 0.5]$, and

$\text{Vech}(\boldsymbol{\varpi}) = [\varpi_{111} = 0.4, \varpi_{212} = 0.2, \varpi_{221} = 0.4, \varpi_{231} = 0.6]$.

A few other parameter values were also used to simulate the model system. The variance of the continuous endogenous variable equation, $\ddot{\Sigma}_{11} = \text{Var}(\varepsilon_1)$, is specified as 1.25. For the three ordinal endogenous variables, since we consider a three point ordinal scale for each, only one threshold needs to be estimated per ordinal variable. We specify a design value of 1.5 for this threshold for all three ordinal variables (these are the $\tilde{\psi}_{13}, \tilde{\psi}_{23}, \tilde{\psi}_{33}$ values). For the vehicle ownership count variable, we assume one flexibility parameter with a value of 0.75 (φ_1) and a dispersion parameter with a value of 2.0 (θ_1). Finally, for the error-covariance matrix of the two nominal variables, we assume a non-IID error structure for both the nominal variables (each with three alternatives), without any covariance in utilities across the two nominal variables. The specification of the error covariance matrix ($\boldsymbol{\Lambda}$) of both the nominal variables together is provided below:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.70 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.70 & 1.49 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.60 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.60 & 1.36 \end{bmatrix}$$

3.2 Design of Spatial Dependence

We generate spatial dependence in both the latent constructs using two spatial auto-regressive parameters ($\boldsymbol{\delta}$), one for each latent construct. We generate three different levels of spatial dependence: (1) no spatial dependence (*i.e.*, aspatial model) as in Bhat (2015a), with $\boldsymbol{\delta} = (0,0)$, (2) low spatial dependence using $\boldsymbol{\delta} = (0.25,0.25)$, and (3) high spatial dependence using $\boldsymbol{\delta} = (0.75,0.75)$. The aspatial model is generated as a base case

for comparison purposes. We work with two different levels of spatial dependency – low and high – since prior experience suggests difficulty in accurately recovering the parameters at high levels of spatial dependencies.

There are a total of 52 parameters to be estimated for the spatial models (50 parameters for the aspatial model). Using these parameters and the spatial GHDM framework described in Section 2, for each level of spatial dependency, we generated 50 data samples of 3000 individual observations each. For generating the weight matrix, the first step is to generate a matrix of distance between all pairs of observations. To do so, we start with a rectangular spatial configuration of size 60 × 50 (=3000) containing equi-distant square grids. Then, each individual (observation) was randomly assigned to one of the 3000 grids. Based on this spatial configuration of individuals, the distance matrix was calculated assuming the centroid of grids as the x-y coordinates of their location. Next, the elements of the weight matrix were calculated as a function of inverse of the distance between pairs of individuals. The choice of inverse of distance for deriving the weight matrix is arbitrary. One may chose other distance decay functions such as the inverse of exponential of distance, inverse of square of distance, and a simple contiguity matrix.

Finally, to keep the number of pairs in the CML function to be reasonable, we impose a threshold distance of 1.5 units beyond which spatial dependency is not considered. This helps not only in avoiding a large number of pairs of individuals in the CML function (for computational tractability), but also to ensure that every individual has atleast one pair (*i.e.*, every individual's latent constructs are influenced by atleast one other individual). Based on our spatial configuration of 3000 individuals in the sample, the threshold distance of 1.5 units results in a total of 5890 pairs of individuals.

3.3 Performance Evaluation

The performance of the MACML inference approach in recovering the parameters of the spatial and aspatial GHDM and the corresponding standard errors is evaluated as follows.

(1) Estimate the MACML parameters for the 50 datasets. Estimate the standard errors using the Godambe (sandwich) estimator.

(2) Compute the mean for each model parameter across the 50 datasets to obtain a **mean estimate**. Compute the **absolute percentage (finite sample) bias** or **(APB)** of the estimator:

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

(3) Compute the standard deviation of the mean estimate across the 50 datasets, and label this as the **finite sample standard error or FSSE** (essentially, this is the empirical standard error). Then, compute **FSSE % of true value**, which is the FSSE expressed as a percentage of the true value of the parameter.

(4) Compute the mean standard error for each model parameter across the 50 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). Then,

compute **ASE % of true value**, which is ASE expressed as a percentage of the true parameter value.

(5) Next, to evaluate the accuracy of the ASE formula as computed using the MACML inference approach 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 error as:

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

3.4 Simulation Results

An overall summary of the simulation results is presented in Table 1.[2] The table presents summaries of the above discussed evaluation measures—APB, FSSE % of true estimate, and APBASE—for each of the following cases: (1) aspatial model estimated on data with no spatial dependency, (2) spatial model estimated on low spatial dependence data, (3) spatial model estimated on high spatial dependence data, and (4) aspatial model estimated on low spatial dependence data. A block of columns is devoted for each of these four cases. The first case, which is not the focus of this paper, is only used for comparison purposes (more later). The last case is used for assessing the repercussions of ignoring spatial dependency when present. We chose low spatial dependency data for this case to evaluate the importance of accommodating spatial dependency even when present at moderate levels. Each row in the table represents the summary of evaluation measures for a specific class of parameters identified in the first column.

[2] The detailed results are available on request from the authors.

Table 1: Summary of Simulation Results

Parameters	Case(1): No spatial dependence			Case(2): Low spatial dependence			Case(3): High spatial dependence			Case(4): Ignoring spatial dependence		
	APB (%)	FSSE as a % of true value	APBASE (%)	APB (%)	FSSE as a % of true value	APBASE (%)	APB (%)	FSSE as a % of true value	APBASE (%)	APB (%)	FSSE as a % of true value	APBASE (%)
Effects of exogenous variables on latent constructs (α)	14.93	19.24	19.41	16.76	30.25	21.08	17.53	28.43	22.68	20.95	29.50	114.10
Structural equation correlation matrix parameters (Γ)	10.50	28.07	14.01	10.67	22.75	4.45	14.50	21.15	3.07	26.17	21.65	57.60
Effect of exogenous and endogenous variables on non-nominal variables ($\vec{\gamma}$)	6.05	9.23	26.38	6.14	8.75	20.52	6.11	10.19	9.91	7.55	8.82	22.00
Effect of exogenous and endogenous variables on nominal variables (b)	3.71	29.70	29.86	16.34	19.30	23.48	18.28	25.49	22.54	23.51	18.39	61.06
Effect of latent constructs on non-nominal variables ($\vec{d}$)	14.01	21.83	21.18	13.51	12.46	28.38	13.74	12.35	39.40	30.58	12.58	182.56
Effect of latent constructs on nominal variables (ϖ)	10.40	29.31	46.34	12.61	16.02	24.86	14.04	15.10	19.00	27.11	17.24	132.67
Variance of continuous variable (Σ)	10.72	2.04	9.57	10.64	2.36	32.93	8.24	1.79	61.92	21.76	1.72	109.21
Thresholds of ordinal variables ($\tilde{\psi}$)	1.61	4.07	20.54	3.49	3.18	23.90	4.05	3.46	24.24	3.30	3.17	27.71
Flexibility parameter for the count variable (φ)	6.83	9.43	25.82	16.40	4.44	23.21	29.91	3.71	40.16	29.40	4.39	167.07
Dispersion parameter for the count variable (θ)	17.20	8.63	2.64	10.40	2.98	47.89	18.45	5.14	35.29	30.45	3.13	176.08
Nominal variables error covariance matrix (Λ)	2.22	36.94	16.99	7.97	5.62	19.41	7.44	5.10	16.15	19.61	5.61	58.28
Spatial autoregressive parameters (δ)	-----	-----	-----	0.92	5.00	47.76	5.14	2.38	18.87	-----	-----	-----
Overall Average	**7.35**	**20.19**	**25.58**	**10.90**	**13.07**	**24.15**	**12.13**	**14.50**	**21.50**	**19.04**	**13.21**	**78.55**

Several observations can be made from Table 1. First, as may be observed from the magnitude of the overall average APB values (last row) for case 2 and case 3, the MACML approach does a pretty good job of recovering the parameter values for both the low and high spatial dependency cases. The overall average APB values across all parameters (under the column titled "APB") for the low and high spatial dependency cases are 10.90 and 12.13, respectively. These APB values are not too high when compared to the overall average APB of 7.35% for the model estimated on data without any spatial dependency (Case 1).[3] In other words, including spatial dependency in the model decreased the ability to recover parameters very marginally. Similarly, increasing the spatial dependency also deteriorated the APB value by only a few percentage points. This suggests that the ability of the MACML approach to recover model parameters is not very dependent on the presence and the extent of spatial dependency in the data, at least for the current setup with a sample size of 3000. This is an important result in favor of the MACML approach, because the general perception, perhaps based on experience with traditional estimation methods, is that parameter recovery degrades quickly due to the presence of, or increase in, spatial dependency in the data.

Second, not all types of parameters are recovered with a high level of accuracy. The parameters corresponding to the effects of exogenous variables on the latent variables (*i.e.*, elements of $\text{Vech}(\boldsymbol{\alpha})$), parameters corresponding to the effects of exogenous and other endogenous variables on the nominal variables (*i.e.*, elements of $\text{Vech}(\boldsymbol{b})$), parameters corresponding to the effects of latent variables on the non-nominal variables (*i.e.*, elements of $\text{Vech}(\overleftrightarrow{\boldsymbol{d}})$) and nominal variables (*i.e.*, elements of $\text{Vech}(\boldsymbol{\varpi})$), and the dispersion ($\boldsymbol{\theta}$) and flexibility ($\boldsymbol{\varphi}$) parameters corresponding to the count variable (vehicle ownership) have relatively higher APB values (than those for other parameters) in both the low and high spatial dependency cases (*i.e.*, Case 2 and Case 3). As discussed in Bhat (2015a), it is not surprising that many of these parameters are more difficult to retrieve than others because of the highly non-linear nature of their entry in the CML function. However, among all these parameters, when compared to the corresponding APBs from the aspatial data (case 1), the presence of spatial dependency leads to a substantial increase in APB values of only the parameters corresponding to the effects of exogenous and other endogenous variables on the nominal variables (*i.e.*, elements of $\text{Vech}(\boldsymbol{b})$) and those of the flexibility and dispersion parameters in the count variable. The recovery of all other parameters is not influenced substantially by spatial dependency (its presence or extent). As a matter of fact, the effects of exogenous and other endogenous variables on the non-nominal variables (continuous, ordinal, and count variables) are recovered as well as those in the aspatial model. Similarly, for all other parameters, including correlation parameters in structural equation ($\text{Vech}(\boldsymbol{\Gamma})$), variance of the continuous variables ($\text{Vech}(\boldsymbol{\Sigma})$), threshold values for ordinal variables ($\text{Vech}(\tilde{\boldsymbol{\psi}})$), covariance matrices corresponding to nominal variables ($\text{Vech}(\boldsymbol{\Lambda})$), and spatial auto-

[3] Recall that the aspatial simulation setup here is similar to that used by Bhat (2015a). His results show an overall mean APB of 6.29%, where as we show 7.35%. The difference, despite the same setup, is because the results in this paper are from only 50 simulated datasets whereas Bhat's (2015a) results are from 200 simulated datasets. We expect the APB values reported in this paper to decrease with an increase in the number of simulated datasets.

correlation parameters ($\boldsymbol{\delta}$)) , are recovered with high accuracy in both low and high spatial dependency cases.

The above results suggest little differences in the ability of the MACML approach to recover parameters from datasets with different levels of spatial dependency, except for the effects of exogenous and endogenous variables on nominal variables and the flexibility and dispersion parameters of the count variables. The relative difficulty of recovering parameters corresponding to the nominal variables is perhaps because they are characterized by multiple latent utility variables, as opposed to non-nominal variables that are characterized by at most a single underlying latent variable. However, it is not clear why the parameter recovery deteriorated (from none to low to high spatial dependency) for the flexibility and dispersion parameters in the count variables. Since the simulation framework considers only one single count variable with one flexibility parameter, it may be difficult to draw conclusions from the evidence here. Further investigation with a greater number of count variables may help shed more light on this issue.

Another notable result is that the absolute percentage bias for the spatial auto-correlation parameters ($\boldsymbol{\delta}$) is merely 0.92 for the low spatial dependency case and 5.14 for the high spatial dependency case. It suggests that the MACML approach is able to recover spatial parameters with high accuracy even for a high spatial dependency case. This is an important result for such complex models with spatial dependency and other unobserved effects, because poorly recovered spatial effects may get confounded with other parameters – especially with parameters representing unobserved effects such as those in ($\text{Vech}(\boldsymbol{\Gamma})$), ($\text{Vech}(\boldsymbol{\Sigma})$) and ($\text{Vech}(\boldsymbol{\Lambda})$) – and may lead to distorted interpretations and policy implications.

Third, the overall average values of FSSE expressed as a percentage of true value, are pretty small – 13.07 and 14.50 for the low and high spatial dependency cases, respectively. This suggests that the MACML approach exhibits notably good empirical efficiency, with the parameters being recovered with a relatively high precision. However, and for the same reasons discussed above, all sets of parameters that are relatively more difficult to recover (*i.e.*, those with relatively high APB values), such as those in ($\text{Vech}(\boldsymbol{\alpha})$), ($\text{Vech}(\boldsymbol{b})$), ($\text{Vech}(\vec{\boldsymbol{d}})$) and ($\text{Vech}(\boldsymbol{\varpi})$), also have high FSSE values (as a percentage of true value). The overall interpretation is that it is difficult to accurately and precisely recover the effects of exogenous variables on latent variables in the structural equation system, the effects of latent constructs on different outcomes in the measurement equation, and that of exogenous and other endogenous variables on nominal outcomes in spatial GHDM models.

Finally, as may be observed from the APBASE values, the asymptotic formula (Godambe's sandwich estimator) in conjunction with the CML approach for ASE is able to estimate the FSSEs reasonably well, for both low and high levels of spatial dependency. The overall APBASE values for the low and high spatial dependency cases are 24.15 and 21.50, respectively, which are not large considering that the actual FSSE values themselves are small.

In summary, the MACML approach is able to recover the parameters with a reasonable level of accuracy and precision for different levels of spatial dependency. As importantly,

the accuracy and precision in the parameter estimates do not degrade even if the spatial dependency increases to high levels, which makes the approach appealing to work with high dimensional heterogeneous datasets with high spatial dependencies. The exception is that the presence and extent of spatial dependency seems to influence the accuracy and precision of the parameter estimates corresponding to the exogenous and endogenous variables effects on nominal outcomes. The CML-based asymptotic standard error formula is fairly accurate in providing standard errors with finite samples, which should increase the analysts' confidence in making statistical inferences with the approach.

3.5 Effect of Ignoring Spatial Dependency

Here we present the effects of ignoring spatial dependencies when they are present (the last block of columns under "Case 4" in Table 1). To examine such effects, we estimated an aspatial model (by constraining the spatial auto-correlation parameters ($\boldsymbol{\delta}$) to zero) on data generated with low spatial dependency. As can be observed from the APB column for this case, the overall average APB is now 19.04, about eight percentage points higher than the overall APB (10.90) from a model in Case 2 that recognizes spatial dependency. All types of parameters, except those corresponding to the effect of exogenous variables on the non-nominal variables and thresholds of the ordinal variables, show a significant increase in the bias due to ignoring spatial dependency. Specifically, the APB values for the parameters corresponding to the effects of latent variables on the non-nominal variables ($\text{Vech}(\vec{\boldsymbol{d}})$) and nominal variables ($\text{Vech}(\boldsymbol{\varpi})$) have more than doubled. This result is expected since the spatial effects are introduced into all endogenous variables through latent constructs. Therefore, ignoring spatial effects leads to a bias in the effects of the latent variables on all endogenous variables. Further, the bias has increased substantially for the parameters corresponding to unobserved effects, including correlations in structural equation ($\text{Vech}(\boldsymbol{\Gamma})$), variance of the continuous variable ($\text{Vech}(\boldsymbol{\Sigma})$), covariance matrices of nominal variables ($\text{Vech}(\boldsymbol{\Lambda})$), as well as the dispersion parameter ($\boldsymbol{\theta}$) and flexibility ($\boldsymbol{\varphi}$) parameter in the count model. Only for the parameters corresponding to the effects of observed (exogenous and endogenous) variables on non-nominal variables, the increase in bias is not as high as that compared to other parameters.

The FSSEs (as percentages of the true values) obtained when the spatial dependency is ignored are similar to those from the spatial model (*i.e.*, Case 2). This metric, however, does not provide a measure of efficiency in estimation. It is more useful to look at the ASE values or the APBASE values, as in the last column of the table. Specifically, the APBASE values are very large for most types of parameters. This is a manifestation of a considerable loss in estimation efficiency due to ignoring spatial dependency, suggesting the unreliability of statistical inferences one can make (on the parameter estimates) from GHDM models that ignore spatial dependency.

Finally, we used the adjusted composite log-likelihood ratio test (ADCLRT) to assess the deterioration in data-fit due to ignoring spatial dependency; more precisely, to decide whether an aspatial model suffices for a given dataset with low spatial dependency (Bhat, 2011). The ADCLRT statistic, which is a modified version of the familiar log-likelihood ratio test, follows an approximate chi-squared distribution; with 2 degrees of freedom for

the two spatial autocorrelation parameters in this case. The corresponding critical chi-square value for a 0.1% level of significance is 9.21. The ADCLRT statistic value for all 50 simulated datasets is higher than 9.21 rejecting the aspatial model in favor of the spatial GHDM in all 50 samples. This highlights the importance of considering spatial dependencies in the model rather than making a-priori assumptions.

Overall, ignoring spatial dependency, even when present only to a small extent, can lead to important repercussions, such as a substantial loss in the accuracy (*i.e.*, unbiasedness) and efficiency in parameter estimation and a deterioration in overall data-fit. All of these effects will likely manifest in the form of distorted inferences and policy implications from an aspatial model.

Chapter 4. Summary and Conclusions

This paper develops a framework for incorporating spatial dependencies in integrated model systems of latent variables and multidimensional mixed data outcomes. The framework combines Bhat's Generalized Heterogeneous Data Model (GHDM) with a spatial formulation and introduces spatial dependencies through latent constructs.

For estimating the parameters of the proposed spatial GHDM framework, the paper employs the maximum approximate composite marginal likelihood (MACML) approach which reduces the dimensionality of the integrals to be evaluated independent of the extent of spatial dependence, the number of latent constructs, or the number of dependent variables in the multidimensional mixed data bundle. To evaluate the MACML approach in its ability to recover parameters of the spatial GHDM, we undertake Monte Carlo simulation experiments on synthetic data. The simulation results suggest that the MACML approach is able to recover parameters of the spatial GHDM with a pretty good level of accuracy and precision that is close enough to the level of parameter recovery achieved on datasets without any spatial dependencies. As importantly, for a majority of parameters in the model system, the overall accuracy and precision in estimation does not degrade much even when the spatial dependency increases to high levels. This makes the performance of the MACML approach less tied to the presence and level of spatial dependency and appealing for situations even with high spatial dependencies in multidimensional mixed data.

Additional simulation experiments were conducted to assess the repercussions of ignoring spatial dependencies (when present). The results suggest that, ignoring spatial dependency even when present at low levels can lead to a substantial loss in the accuracy and efficiency in parameter estimation. The bias in parameter recovery (as measured by APB values) for a majority of parameters was over 25%, while the asymptotic standard errors for many parameters were over 100% of the finite sample standard errors. The APB values due to neglected spatial dependencies more than doubled (and became close to 30%) for the effects of latent variables on different endogenous outcomes; not to mention the corresponding ASE values (as a percentage of FSSEs) were at least 130%. This suggests that ignoring spatial dependency can potentially lead to severely distorted inferences of policy implications relevant to the influence of latent constructs on endogenous variables. For example, it would be difficult to credibly assess the influence of public policy instruments (*e.g.,* educational campaigns) aimed at bringing about attitudinal and lifestyle changes for desirable travel behaviors in the population. The other effects of neglected spatial dependency include confounded effects on unobserved effects such as correlations among latent constructs, variance of continuous variable, dispersion and flexibility parameter of the count variable, and covariance among the nominal variable alternatives. Indeed, the only set of parameters that were not influenced substantially was the effects of observed (exogenous and endogenous) variables on non-nominal outcomes and the threshold parameters of ordinal outcomes. Consistent with the above results, ignoring spatial dependency resulted in statistically significant deterioration of model fit in every synthetic dataset generated for this study.

All the above results highlight the importance of accommodating spatial dependency, or at least testing for the presence of such dependencies in integrated models of mixed data

outcomes and latent variables. The good news is that, thanks to the efficacy of the MACML approach, the spatial autocorrelation parameters were recovered remarkably well when an attempt was made to accommodate spatial dependency. The relative ease with which spatial dependency can be detected helps in avoiding a-priori assumptions. Therefore, the proposed spatial GHDM framework combined with the MACML approach to estimate its parameters can potentially be a valuable tool for modeling spatial dependencies in multidimensional mixed data outcomes that are becoming of increasing interest in several fields.

References

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1), 3-25.

Beron, K.J., and Vijverberg, W.P.M. (2004). Probit in a spatial context: A Monte Carlo analysis. In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, 169-196, edited by Anselin, L., Florax, R.J.G.M., Rey, S.J., Springer-Verlag, Berlin.

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), 1-117.

Bhat, C.R. (2015a). A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B*, 79, 50-77.

Bhat, C.R. (2015b). A new spatial (social) interaction discrete choice model accommodating for unobserved effects due to endogenous network formation. *Forthcoming in Transportation*.

Bhat, C.R., and 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., Paleti, R., and Singh, P. (2014). A spatial multivariate count model for firm location decisions. *Journal of Regional Science*, 54(3), 462-502.

De Leon, A.R., and Zhu, Y. (2008). ANOVA extensions for mixed discrete and continuous data. *Computational Statistics & Data Analysis*, 52(4), 2218-2227.

De Leon, A.R., and Chough, K.C. (2013). *Analysis of Mixed Data: Methods & Applications*, CRC Press.

Elhorst, J.P. (2010a). Applied spatial econometrics: raising the bar. *Spatial Economic Analysis*, 5(1), 9-28.

Ferdous, N., and C.R. Bhat (2013). A spatial panel ordered-response model with application to the analysis of urban land-use development intensity patterns. *Journal of Geographical Systems*, 15(1), 1-29.

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

LeSage, J.P., and Pace, R.K. (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: Chapman & Hall/CRC, Taylor & Francis Group.

Sidharthan, R., and Bhat, C.R. (2012). Incorporating spatial dynamics and temporal dependency in land use change models. *Geographical Analysis*, 44(4), 321-349.

Smirnov, O.A. (2010). Modeling spatial discrete choice. *Regional Science and Urban Economics*, 40(5), 292-298.

Teixeira-Pinto, A., and Harezlak, J. (2013). Factorization and latent variable models for joint analysis of binary and continuous outcomes. In *Analysis of Mixed Data*, pp. 81-91, Chapman and Hall/CRC.