Online supplement for the paper
"Sharing the Road with Autonomous Vehicles: Perceived Safety and Regulatory Preferences"

Gopindra S. Nair, Chandra R. Bhat (corresponding author)

Mathematical formulation of GHDM model with only ordinal outcomes

Consider the case of an individual $q \in \{1, 2, ..., Q\}$. Let $l \in \{1, 2, ..., L\}$ be the index of the latent constructs and let z^*_{ql} be the value of the latent variable *l* for the individual *q*. z^*_{ql} is expressed as a function of its explanatory variables as,

$$z^*_{ql} = \boldsymbol{w}^{\mathrm{T}}_{\boldsymbol{ql}} \boldsymbol{\alpha} + \eta_{ql}\,, \tag{1}$$

where $\boldsymbol{w_{ql}}$ $(D \times 1)$ is a column vector of the explanatory variables of latent variable *l* and $\boldsymbol{\alpha}$ $(D \times 1)$ is a vector of its coefficients. η_{ql} is the unexplained error term and is assumed to follow a standard normal distribution. Equation (1) can be expressed in the matrix form as,

$$\boldsymbol{z^*_q} = \boldsymbol{w_q} \boldsymbol{\alpha} + \boldsymbol{\eta_q}\,, \tag{2}$$

where $\boldsymbol{z^*_q}$ $(L \times 1)$ is a column vector of all the latent variables, $\boldsymbol{w_q}$ $(L \times D)$ is a matrix formed by vertically stacking the vectors $(\boldsymbol{w^{\mathrm{T}}_{q1}}, \boldsymbol{w^{\mathrm{T}}_{q2}}, ..., \boldsymbol{w^{\mathrm{T}}_{qL}})$ and $\boldsymbol{\eta_q}$ $(D \times 1)$ is formed by vertically stacking $(\eta_{q1}, \eta_{q2}, ..., \eta_{qL})$. $\boldsymbol{\eta_q}$ follows a multivariate normal distribution centered at the origin and having a correlation matrix of $\boldsymbol{\Gamma}$ $(L \times L)$, i.e., $\boldsymbol{\eta_q} \sim MVN_L(\boldsymbol{0_L}, \boldsymbol{\Gamma})$, where $\boldsymbol{0_L}$ is a vector of zeros. The variance of all the elements in $\boldsymbol{\eta_q}$ is fixed as unity because it is not possible to uniquely identify a scale for the latent variables. Equation (2) constitutes the SEM component of the framework.

Let $j \in \{1, 2, ..., J\}$ denote the index of the outcome variables (including the indicator variables). Let y^*_{qj} be the underlying continuous measure associated with the outcome variable y_{qj}. Then,

$$y_{qj} = k \text{ if } t_{jk} < y^*_{qj} \le t_{j(k+1)}\,, \tag{3}$$

where $k \in \{1, 2, ..., K_j\}$ denotes the ordinal category assumed by y_{qj} and t_{jk} denotes the lower boundary of the kth discrete interval of the continous measure associated with the jth outcome.

$t_{jk} < t_{j(k+1)}$ for all *j* and all *k*. Since y_j^* may take any value in $(-\infty, \infty)$, we fix the value of $t_{j1} = -\infty$ and $t_{j(K_j+1)} = \infty$ for all *j*. Since the location of the thresholds on the real-line is not uniquely identifiable, we also set $t_{j2} = 0$. y_j^* is expressed as a function of its explanatory variables as,

$$y_{qj}^* = \boldsymbol{x}_{\boldsymbol{qj}}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{q}}^{*\mathrm{T}} \boldsymbol{d}_{\boldsymbol{j}} + \xi_{qj}, \tag{4}$$

where $\boldsymbol{x}_{\boldsymbol{qj}} (E \times 1)$ is a vector of size of explanatory variables for the continuous measure y_{qj}^*. $\boldsymbol{\beta}$ $(E \times 1)$ is a column vector of the coefficients associated with $\boldsymbol{x}_{\boldsymbol{qj}}$ and $\boldsymbol{d}_{\boldsymbol{j}}$ $(L{\times}1)$ is the vector of coefficeints of the latent variables for outcome *j*. ξ_{qj} is a stochastic error term that captures the effect of unobserved variables on the value of y_{qj}^*. ξ_{qj} is assumed to follow a standard normal distribution. Jointly, the continuous measures of the *J* outcome variables may be expressed as,

$$\boldsymbol{y}_{\boldsymbol{q}}^* = \boldsymbol{x}_{\boldsymbol{q}} \boldsymbol{\beta} + \boldsymbol{d} \boldsymbol{z}_{\boldsymbol{q}}^* + \boldsymbol{\xi}_{\boldsymbol{q}}, \tag{5}$$

where $\boldsymbol{y}_{\boldsymbol{q}}^*$ $\left(J \times 1\right)$ and $\boldsymbol{\xi}_{\boldsymbol{q}}$ $\left(J \times 1\right)$ are the vectors formed by vertically stacking y_{qj}^* and ξ_{qj} respectively of the *J* dependent variables. $\boldsymbol{x}_{\boldsymbol{q}}$ $(J \times E)$ is a matrix formed by vertically stacking the vectors $\left(\boldsymbol{x}_{\boldsymbol{q1}}^{\mathrm{T}}, \boldsymbol{x}_{\boldsymbol{q2}}^{\mathrm{T}}, ..., \boldsymbol{x}_{\boldsymbol{qJ}}^{\mathrm{T}}\right)$ and $\boldsymbol{d}$ $(J \times L)$ is a matrix formed by vertically stacking $\left(\boldsymbol{d}_{\boldsymbol{1}}^{\mathrm{T}}, \boldsymbol{d}_{\boldsymbol{2}}^{\mathrm{T}}, ..., \boldsymbol{d}_{\boldsymbol{J}}^{\mathrm{T}}\right)$. $\boldsymbol{\xi}_{\boldsymbol{q}}$ follows a multivariate normal distribution centered at the origin with an identity matrix as the covariance matrix (independent error terms). $\boldsymbol{\xi}_{\boldsymbol{q}} \sim MVN_J(\boldsymbol{0}_{\boldsymbol{J}}, \mathbf{I}_{\boldsymbol{J}})$. We assume the terms in $\boldsymbol{\xi}_{\boldsymbol{q}}$ to be independent because it is not possible to uniquely identify all the correlations between the elements in $\boldsymbol{\eta}_{\boldsymbol{q}}$ and all the correlations between the elements in $\boldsymbol{\xi}_{\boldsymbol{q}}$. Further, because of the ordinal nature of the outcome variables, the scale of $\boldsymbol{y}_{\boldsymbol{q}}^*$ cannot be uniquely identified. Therefore, the variances of all elements in $\boldsymbol{\xi}_{\boldsymbol{q}}$ is fixed to one. The reader is referred to Bhat (2015) for further nuances regarding the identification of coefficients in the GHDM framework.

Substituting Equation (2) in Equation (5), $\boldsymbol{y}_{\boldsymbol{q}}^*$ can be expressed in the reduced form as,

$$\boldsymbol{y}_{\boldsymbol{q}}^* = \boldsymbol{x}_{\boldsymbol{q}} \boldsymbol{\beta} + \boldsymbol{d}\left(\boldsymbol{w}_{\boldsymbol{q}} \boldsymbol{\alpha} + \boldsymbol{\eta}_{\boldsymbol{q}}\right) + \boldsymbol{\xi}_{\boldsymbol{q}}, \tag{6}$$

$$\boldsymbol{y}_{\boldsymbol{q}}^* = \boldsymbol{x}_{\boldsymbol{q}} \boldsymbol{\beta} + \boldsymbol{d} \boldsymbol{w}_{\boldsymbol{q}} \boldsymbol{\alpha} + \boldsymbol{d} \boldsymbol{\eta}_{\boldsymbol{q}} + \boldsymbol{\xi}_{\boldsymbol{q}}. \tag{7}$$

In the R.H.S. of Equation (7), $\boldsymbol{\eta}_q$ and $\boldsymbol{\xi}_q$ are random vectors that follow the multivariate normal distribution and the other variables are constants. Therefore, $\boldsymbol{y}_q^*$ also follows the multivariate normal distribution with a mean of $\boldsymbol{b} = \boldsymbol{x}_q\boldsymbol{\beta} + \boldsymbol{d}\boldsymbol{w}_q\boldsymbol{\alpha}$ (all the elements of $\boldsymbol{\eta}_q$ and $\boldsymbol{\xi}_q$ have a mean of zero) and a covariance matrix of $\boldsymbol{\Sigma} = \boldsymbol{d}\boldsymbol{\Gamma}\boldsymbol{d}^{\mathrm{T}} + \mathbf{I}_J$.

$$\boldsymbol{y}_q^* \sim MVN_J(\boldsymbol{b}, \boldsymbol{\Sigma}) . \tag{8}$$

The parameters that are to be estimated are the elements of $\boldsymbol{\alpha}$, strictly upper triangular elements of $\boldsymbol{\Gamma}$, elements of $\boldsymbol{\beta}$, elements of $\boldsymbol{d}$ and t_{jk} for all *j* and $k \in \{3, 4, ..., K_j\}$. Let $\boldsymbol{\theta}$ be a vector of all the parameters that need to be estimated. The maximum likelihood approach can be used for estimating these parameters. The likelihood of the q^{th} observation will be,

$$L_q(\boldsymbol{\theta}) = \int_{v_1 = t_{1y_{q1}} - b_1}^{v_1 = t_{1(y_{q1}+1)} - b_1} \int_{v_2 = t_{2y_{q2}} - b_2}^{v_2 = t_{2(y_{q2}+1)} - b_2} \dots \int_{v_J = t_{Jy_{qJ}} - b_J}^{v_J = t_{J(y_{qJ}+1)} - b_J} \phi_J(v_1, v_2, \dots, v_J \mid \boldsymbol{\Sigma}) dv_1 dv_2 \dots dv_J \ , \tag{9}$$

where, $\phi_J\left(v_1, v_2, \dots, v_J \mid \boldsymbol{\Sigma}\right)$ denotes the probability density of a *J* dimensional multivariate normal distribution centered at the origin with a covariance matrix $\boldsymbol{\Sigma}$ at the point $(v_1, v_2, \dots, v_J)$. Since a closed form expression does not exist for this integral and evaluation using simulation techniques can be time consuming, we used the One-variate Univariate Screening technique proposed by Bhat (2018) for approximating this integral. The estimation of parameters was carried out using the *maxlik* library in the GAUSS matrix programming language.

Output predictions

To predict the outcome for an individual, first the random error terms for the latent variables are drawn from a multivariate normal distribution that has a mean of $\boldsymbol{0}_L$ and a covariance matrix of $\boldsymbol{\Gamma}$. Following this the latent variables are predicted using Equation (2). Then the random error terms of the outcome variables are drawn from independent standard normal distributions. The outcomes are then predicted in a sequential manner using Equation (3) and Equation (4). Since some of the outcome variables are used as explanatory variables for other outcome variables, the sequence used for estimating the outcomes is such that the outcome that is predicted i^{th} will only have explanatory variables that are exogenous or that have been predicted in one of the previous $(i-1)$ steps.

Average treatment effects

The expression for computing the ATE is as follows,

$$A\hat{T}E_{jiAB} = \frac{1}{Q}\sum_{q=1}^{Q}\left(P_q\left(y_j = i \mid B\right) - P_q\left(y_j = i \mid A\right)\right), \tag{10}$$

where $A\hat{T}E_{jiAB}$ denotes the ATE on the *i*th level of the *j*th outcome variable by applying a treatment that changes condition of individuals from *A* to *B*, $P_q(.)$ is a function that computes the probability for the individual *q*. Since the computation of this probability is cumbersome when the outcome variable under consideration is explained by other outcome variables that are also affected by the treatment, we use Monte Carlo methods to simulate the required probabilities. For approximating the probability $P_q\left(y_j = i \mid B\right)$, we make 1000 predictions for the individual *q* where each prediction is made using a different random draw. The process for making the prediction is the same as that described in the previous section except that the variables that are affected by the condition *B* is set as per the condition irrespective of the value generated by the prediction process. The estimate for $P_q\left(y_j = i \mid B\right)$ will then be the proportion of the number of cases where y_j assumes the value *i* to the total number of predictions.

References

Bhat, C.R. (2015). A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B*, 79, 50–77. https://doi.org/10.1016/j.trb.2015.05.017

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

Latent variable loadings and thresholds of indicators

Variable (Minimum level – Maximum level)	Loading		Constant		Threshold 2|3		Threshold 3|4	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Tech-savviness indicators								
Computer use frequency (Never/Almost never – Everyday)	0.803	17.59	2.620	27.86	0.952	17.43	1.727	25.49
Frequency of internet use (Less than once a day – Many times a day)	0.853	18.83	1.794	24.99	1.272	27.61		
Uses laptop (No – Yes)	0.494	15.99	0.769	20.29				
Uses tablet or e-book reader (No – Yes)	0.287	12.23	0.049	1.85				
Used voice activated digital assistant in smart devices (No – Yes)	0.176	7.26	-0.669	-26.41				
Enthusiasm about riding in / sharing the road with AVs								
Enthusiasm about the development of AVs (Not at all enthusiastic – Very enthusiastic)	1.850	18.30	2.574	19.82	2.420	21.78	4.599	23.23
Anxiety about riding in / sharing road with AVs								
Worried about the development of AVs (Not at all worried – Very worried)	1.229	7.73	1.487	12.18	1.848	12.49	3.886	12.72