Online Supplement to

"A New Generalized Heterogeneous Data Model (GHDM) to Jointly Model Mixed Types of Dependent Variables"

Identification conditions in the case of structural inter-relationships among the latent variables in the structural equation system component of the GHDM model:

A way to develop sufficiency conditions in the case of inter-related latent variables in the GHDM model is to start from the reduced form of the structural equation system (this is the form used in the paper, which allows only covariation). Under the first four conditions listed in Section 3.1 of the paper, the reduced form of the structural equation system is identified. The question then is whether the structural form of the structural equation system (where latent variables are directly inter-related with one another) can be recovered from the reduced form of the structural equation system. This is similar to the textbook case of identification in a simultaneous linear equation system (see, for example, Greene, 2000, Chapter 16). While identification can be achieved in one of several ways to meet required rank and order conditions, an important difference between the simultaneous linear equation system and the structural form of the structural equation system in the GHDM model is that we are likely to have less information about possible equality or linear restrictions on the coefficients on exogenous variables across different latent variable equations in the latter system (another difference is that the scales of each latent variable are not identified, which has a bearing on the identification analysis as discussed below). Thus, considering an unrestricted set of coefficients on the exogenous variables, an easy way to ensure identification is to impose a recursive structure among the latent variables. For simplicity, consider a two equation recursive system for the latent variables as follows:

$$z_1^* = \boldsymbol{\alpha}_1' \boldsymbol{w} + \eta_1,$$

$$z_2^* = \boldsymbol{\alpha}_2' \boldsymbol{w} + \vartheta z_1^* + \eta_2, \text{ and } \tilde{\boldsymbol{\Gamma}} = \begin{bmatrix} 1 & \rho \\ \rho & 1-\vartheta^2 \end{bmatrix},$$

where $\boldsymbol{w}$ is a $(\tilde{D} \times 1)$ vector of observed covariates (excluding a constant), and $\tilde{\boldsymbol{\Gamma}}$ is the covariance matrix of the normally distributed errors terms η_1 and η_2 (we use the covariance structure above because the scale of each latent variable is unidentified, and so the variance of the second latent variable has to be fixed; for convenience in associating the reduced form implied by the above structural relationship to the form of the latent variable SEM used in Equation (2) of the paper, we relate the variance of η_2 to the inter-relationship coefficient ϑ). In reduced form, the equation system above may be written as:

$z_1^* = \boldsymbol{\alpha}_1' \boldsymbol{w} + \eta_1,\ \eta_1 \sim N(0,1)$

$z_2^* = \pi_2' \boldsymbol{w} + \tilde{\eta}_2,\ \pi_2 = \boldsymbol{\alpha}_2 + \vartheta \boldsymbol{\alpha}_1,\ \tilde{\eta}_2 = \vartheta \eta_1 + \eta_2,$ and $\boldsymbol{\Gamma} = cor(\eta_1, \tilde{\eta}_2) = \begin{bmatrix} 1 & \rho + \vartheta \\ \rho + \vartheta & 1 \end{bmatrix}.$

In the above system, what is identified based on the first four conditions in Section 3.1 of the paper are the vectors $\boldsymbol{\alpha}_1, \pi_2,$ and the combined correlation coefficient $\tilde{\rho} = \rho + \vartheta$. So, at issue is whether one can identify the vector $\boldsymbol{\alpha}_2$, ρ, and ϑ from the estimates of the vector $\pi_2 = \boldsymbol{\alpha}_2 + \vartheta \boldsymbol{\alpha}_1$, scalar $\tilde{\rho} = \rho + \vartheta$, and the vector $\boldsymbol{\alpha}_1$. In general, there will be an identification problem because, given $\boldsymbol{\alpha}_1$, we have $(\tilde{D} + 1)$ equations (for the vector π_2 and scalar $\tilde{\rho}$), but $(\tilde{D} + 2)$ unknowns ($\tilde{D}$ elements of vector $\boldsymbol{\alpha}_2$, ρ, and ϑ).

The situation can be resolved in one of two ways: Assume zero covariance between the error terms η_1 and η_2, in which case $\rho = 0$, the estimate of ϑ is obtained directly from the estimate of $\tilde{\rho}$ as $\vartheta = \tilde{\rho}$, and the elements of the vector $\boldsymbol{\alpha}_2$ are obtained from the elements of the vector π_2 as $\boldsymbol{\alpha}_2 = \pi_2 - \vartheta \boldsymbol{\alpha}_1$. If one adopts this approach, the inter-relationship coefficient should be maintained in the model even if it is statistically insignificant, because this is the only way that a correlation is maintained in the reduced form matrix $\boldsymbol{\Gamma}$ (which is needed for identification in the context of the first four conditions set forth in Section 3.1). In the more general case, each downstream latent variable should be related to at least one upstream latent variable in the recursive structure.

The second way to achieve identification is to allow covariance in the error terms η_1 and η_2, but ensure that at least one exogenous variable appearing in the z_1^* equation does not appear as an exogenous variable in the z_2^* equation. This is, of course, equivalent to imposing the constraint that at least one element of the vector $\boldsymbol{\alpha}_2$ is restricted to zero. Without loss of generality, let this be the first variable in the vector $\boldsymbol{w}$, so that $\alpha_{21} = 0$. Then, one can obtain the estimate of $\vartheta = \pi_{21} / \alpha_{11}$, the other elements of the vector $\boldsymbol{\alpha}_2 = \pi_2 - \vartheta \boldsymbol{\alpha}_1$, and the covariance $\rho = \tilde{\rho} - \vartheta$. In a more general recursive system, one can allow a general covariance matrix for $\tilde{\boldsymbol{\Gamma}}$ if each downstream latent variable that has an upstream latent variable on the right side does not have at least one exogenous variable appearing in the upstream latent variable equation.

Reference

Greene, W.H. (2000). *Econometric Analysis*, Prentice Hall, New Jersey.