Two stage multiple imputation SEM for continuous variables

1. Moment matrices

Suppose that the rows of $\mathbf{X}(n \times p)$ are n observations of p continuous variables $x_1, x_2, \ldots, x_p$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The sample covariance matrix, $\mathbf{S}$, is an unbiased estimator of $\boldsymbol{\Sigma}$ and may be expressed as

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

where $\mathbf{x}_i$ and $\bar{\mathbf{x}}$ denote observation i and the sample mean vector of $\mathbf{x} = \left[x_1\, x_2\, \ldots\, x_p\right]'$, respectively. A typical element of a consistent estimator, $\mathbf{U}$, of the asymptotic covariance matrix, $\boldsymbol{\Upsilon}$, of the sample variances and covariances (Browne 1984) is given by

$$u_{ij,kl} = w_{ijkl} - w_{ij}w_{kl}$$

where

$$w_{ijkl} = n^{-1}\sum_{m=1}^{n}\left(x_{im} - \bar{x}_i\right)\left(x_{jm} - \bar{x}_j\right)\left(x_{km} - \bar{x}_k\right)\left(x_{lm} - \bar{x}_l\right)$$

and

$$w_{ij} = n^{-1}\sum_{m=1}^{n}\left(x_{im} - \bar{x}_i\right)\left(x_{jm} - \bar{x}_j\right)$$

where

$$\bar{x}_i = n^{-1}\sum_{m=1}^{n} x_{im}$$

The robust ML, DWLS, WLS, and ULS methods can be used to fit structural equation models for continuous variables to the sample covariance matrix by using the estimated asymptotic covariance matrix of the sample variances and covariances.

The correlation matrix, $\mathbf{P}$, of $x_1, x_2, \ldots, x_p$ is the covariance matrix of the standardized variables $z_1, z_2, \ldots, z_p$ where

$$\mathbf{P} = \mathbf{D}_\sigma^{-1}\boldsymbol{\Sigma}\mathbf{D}_\sigma^{-1}$$

and

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

where $\mathbf{D}_{\boldsymbol{\sigma}}$ denotes a diagonal matrix with the standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_p$ of $x_1, x_2, \ldots, x_p$ on the diagonal. The sample correlation matrix, $\mathbf{R}$, is an unbiased estimator of $\mathbf{P}$ and may be expressed as

$$\mathbf{R} = \mathbf{D}_{\mathbf{s}}^{-1}\mathbf{R}\mathbf{D}_{\mathbf{s}}^{-1}$$

where $\mathbf{D}_{\mathbf{s}}$ denotes a diagonal matrix with the sample standard deviations $s_1, s_2, \ldots, s_p$ of $x_1, x_2, \ldots, x_p$ on the diagonal. A typical element of a consistent estimator, $\mathbf{U}$, of the asymptotic covariance matrix, $\boldsymbol{\Upsilon}$, of the sample correlations (Steiger and Hakstian 1982) is given by

$$u_{ij,kl} = r_{ijkl} + \frac{1}{4} r_{ij} r_{kl} \left(r_{iikk} + r_{jjkk} + r_{iill} + r_{jjll} \right) - \frac{1}{2} r_{ij} \left(r_{iikl} + r_{jjkl} \right) - \frac{1}{2} r_{kl} \left(r_{ijkk} + r_{ijll} \right)$$

where

$$r_{ijkl} = \left(n-1\right)^{-1} \sum_{m=1}^{n} z_{im} z_{jm} z_{km} z_{lm}$$

and

$$r_{ij} = \left(n-1\right)^{-1} \sum_{m=1}^{n} z_{im} z_{jm}$$

and

$$z_{im} = \frac{x_{im} - \overline{x}_i}{s_i}$$

The robust DWLS, WLS, and ULS methods can be used to fit structural equation models for continuous variables to the sample correlation matrix by using the estimated asymptotic covariance matrix of the sample correlations.

2. Multiple imputation

2.1 The MCMC method

Suppose now that the n observations of the p continuous variables include missing data values with k missing data value patterns and that the joint distribution of the variables is a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The EM algorithm and the MCMC method for multiple imputation of incomplete data can be used to impute the missing data values of the continuous variables.

Suppose that $\mathbf{X}_o$ denote the observed data values. The EM algorithm (Dempster, Laird, and Rubin 1977) can be used to compute the maximum likelihood estimate of $\boldsymbol{\Sigma}$. The minus two observed-data log likelihood may be expressed as

$$-2\ln L(\boldsymbol{\Sigma} \mid \mathbf{X}_o) = \sum_{i=1}^{k} n_i \ln\left|\boldsymbol{\Sigma}_i\right| + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{x}_{oij} - \boldsymbol{\mu}_i\right)' \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x}_{oij} - \boldsymbol{\mu}_i\right)$$

where n_i denotes the number of observations of missing data value pattern $i = 1, 2, \ldots, k$, $\boldsymbol{\Sigma}_i$ denotes the population covariance matrix of missing data value pattern i, $\boldsymbol{\mu}_i$ denotes the mean vector of missing data value pattern i, and $\mathbf{x}_{oij}$ is the j^{th} vector of observed values of missing data value pattern i.

The initial estimate for the M-step is the sample covariance matrix, $\mathbf{S}$, of the complete data or $\mathbf{I}_p$ if the number of complete observations is too small. In the E-step, the conditional covariance matrices of the missing variables given the observed variables of the missing data value patterns are computed and used to compute an updated estimate $\hat{\boldsymbol{\Sigma}}^{(t+1)}$ of $\boldsymbol{\Sigma}$. Iteration of the consecutive M and E steps is terminated when the absolute difference between $\hat{\boldsymbol{\Sigma}}^{(t+1)}$ and $\hat{\boldsymbol{\Sigma}}^{(t)}$ is below the tolerance limit $\varepsilon = 10^{-5}$.

The EM estimate, $\hat{\boldsymbol{\Sigma}}$, of $\boldsymbol{\Sigma}$ is used as the initial covariance matrix of the multivariate normal distribution in the first step of the Monte Carlo Markov Chain (MCMC) method. In the first step (P-step) of the MCMC method, an estimate of $\boldsymbol{\Sigma}$ is simulated from an inverse Wishart distribution. In the I-step, observations are simulated from the conditional normal distributions of the missing variables given the observed k missing data value patterns and used to replace the missing data values. The next estimate of $\boldsymbol{\Sigma}$ is then obtained by computing the sample covariance matrix of the completed data. The P and I steps are repeated for a fixed number of times.

2.2 The FCS regression method

Suppose now that the n observations of the p continuous variables include missing data values and that a joint (multivariate) distribution of the variables exists. In this case, the Fully Conditional Specified (FCS) regression method (Brand 1999; Van Buuren 2007) can be used to impute the missing data values. The FCS regression method performs a fixed number of imputations to impute the missing data values. Each imputation consists of a filled-in phase and an imputation phase. In the filled-in phase, the missing data values are filled-in by using a sequence of regression analyses for the p continuous variables. These filled-in data are then used as the initial data for the imputation phase in which the missing data values are imputed by using a sequence of regression analyses for the p continuous variables. These imputed data are then used as the initial data for the next iteration of the imputation phase and a fixed number of iterations are executed for each imputation.

The filled-in stage fits the following p regression models sequentially to the data, namely

$$\begin{aligned}
x_1 &= \beta_{10} + e_1 \\
x_2 &= \beta_{20} + \beta_{21}x_1 + e_2 \\
x_3 &= \beta_{30} + \beta_{31}x_1 + \beta_{32}x_2 + e_3 \\
&\vdots \\
x_p &= \beta_{p0} + \beta_{p1}x_1 + \beta_{p2}x_2 + \cdots + \beta_{p,p-1}x_{p-1} + e_p
\end{aligned}$$

where the elements of $\boldsymbol{\beta} = \left[\beta_{10}\ \beta_{20} \cdots \beta_{p,p-1}\right]'$ denote unknown regression weights and $e_1, e_2, \ldots, e_p$ are p error variables. The first model is fitted to the complete data for x_1. The corresponding estimates are then used to simulate new parameter values from the posterior distributions of the parameters which in turn is used to fill-in the missing data values for x_1. The second model is then fitted to the complete data for x_2 and the filled-in data for x_1. The final model is fitted to the complete data for x_p and the filled-in data for $x_1, x_2, \ldots, x_{p-1}$. The filled-in data for $x_1, x_2, \ldots, x_p$ are used for the first iteration of the imputation phase. The simulation of the new parameter values from the posterior distributions of the parameters and the imputation of the missing data values for each of the p regression models use the same steps as outlined next for each iteration of the imputation stage.

For each iteration of the imputation stage, the following regression models are fitted sequentially either to the filled-in data or the imputed data, namely

$$x_j = \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \cdots + \beta_p x_p + e_j$$

where $j = 1, 2, \ldots, p$, the elements of $\boldsymbol{\beta}_j = \left[\beta_0\ \beta_1 \ldots \beta_{j-1}\beta_{j+1} \ldots \beta_p\right]'$ denote p unknown regression weights, and e_j denotes an error variable with variance σ_j^2. The estimated covariance matrix of the estimator $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}_j$ may be expressed as

$$\sigma_j^2 \mathbf{V}_j = \sigma_j^2 \left(\mathbf{X}'_{(j)} \mathbf{X}_{(j)}\right)^{-1}$$

where $\mathbf{X}_{(j)}$ denotes rows $1, 2, \ldots, j-1, j, \ldots, p$ of the filled-in or imputed data. New values for the parameters are then simulated from their posterior distributions as

$$\boldsymbol{\beta}_{jt} = \hat{\boldsymbol{\beta}}_j + \sigma_{tj}^2 \mathbf{V}'_{hj} \mathbf{z}$$

$$\sigma_{tj}^2 = \frac{\hat{\sigma}_j^2 \left(n_j - p\right)}{c}$$

where $\mathbf{V}_{hj}$ denotes the upper triangular matrix in the Cholesky decomposition of $\mathbf{V}_j = \mathbf{V}'_{hj} \mathbf{V}_{hj}$, $\mathbf{z}$ denotes a $p \times 1$ standard normal vector, and c is a Chi-square variable with $n_j - p$ degrees of freedom. The missing data values are then imputed as

$$x_{ijm} = \boldsymbol{\beta}'_{jt} \mathbf{x}_{i(j)} + \sigma_{tj} z$$

where x_{ijm} denotes a missing data value in row i and column j of $\mathbf{X}$, $\mathbf{x}_{i(j)}$ denotes row i of $\mathbf{X}_{(j)}$, and z is a standard normal variable.

3. Average unstandardized moment matrices

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ are m imputed data sets for the incomplete data matrix, $\mathbf{X}$, of the p continuous variables $x_1, x_2, \ldots, x_p$ and that $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m$ and $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_m$ denote the corresponding sample covariance matrices and the estimated asymptotic covariance matrices of the variances and covariances, respectively. Then, the average sample covariance matrix is

$$\bar{\mathbf{S}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{S}_i$$

and the average estimated asymptotic covariance matrix is

$$\bar{\mathbf{U}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{U}_i$$

Chung and Cai (2019) point out that $\bar{\mathbf{U}}$ only captures uncertainty based on complete data. As a result, its inverse cannot be used as a weight matrix for the robust ML, DWLS, WLS, and ULS methods for continuous structural equational modeling. A corrected weight matrix is obtained by correcting for the between-imputation variation in the estimated variances and covariances and is obtained as the inverse of

$$\hat{\boldsymbol{\Upsilon}} = \bar{\mathbf{U}} + \frac{m+1}{m(m-1)}\left[\sum_{i=1}^{m}\left(\mathbf{s}_i - \bar{\mathbf{s}}\right)\left(\mathbf{s}_i - \bar{\mathbf{s}}\right)'\right]$$

where $\mathbf{s}$ denotes the $p\times\left(p+1\right)/2$ vector consisting of the nonduplicated elements of the $p\times p$ symmetric matrix $\mathbf{S}$. $\bar{\mathbf{S}}$ and $\hat{\boldsymbol{\Upsilon}}$ can be used to fit structural equation models to the average sample covariance matrix with the robust ML, DWLS, WLS, and ULS methods. The corrected robust DWLS and ULS Chi-square test statistic proposed by Chung and Cai (2019) is given by

$$T_B = (n-1)(\mathbf{s}-\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}}))'\mathbf{V}(\mathbf{s}-\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}}))$$

where

$$\mathbf{V} = \hat{\boldsymbol{\Upsilon}}^{-1} - \hat{\boldsymbol{\Upsilon}}^{-1}\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Delta}}'\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\hat{\boldsymbol{\Upsilon}}^{-1}$$

where $\hat{\boldsymbol{\Delta}}$ denotes the Jacobian matrix of $\boldsymbol{\sigma}(\boldsymbol{\theta})$ with respect to the unknown parameters $\boldsymbol{\theta}$ of the structural equation model evaluated at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$. The small sample adjusted T_B test statistic (Yuan and Bentler 1997) is given by

$$T_{YB} = \frac{T_B}{1+nT_B/(n-1)}.$$

4. Average standardized moment matrices

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ are m imputed data sets for the incomplete data matrix, $\mathbf{X}$, of the p continuous variables $x_1, x_2, \ldots, x_p$ and that $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_m$ and $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_m$ denote the corresponding sample correlation matrices and the estimated asymptotic covariance matrices of the sample correlations, respectively. Then, the average sample correlation matrix is

$$\bar{\mathbf{R}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{R}_i$$

and the average estimated asymptotic covariance matrix is

$$\bar{\mathbf{U}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{U}_i$$

Chung and Cai (2019) point out that $\bar{\mathbf{U}}$ only captures uncertainty based on complete data. As a result, its inverse cannot be used as a weight matrix for the robust DWLS, WLS, and ULS methods for continuous structural equational modeling for correlation matrices. A corrected weight matrix is obtained by correcting for the between-imputation variation in the estimated correlations and is obtained as the inverse of

$$\hat{\boldsymbol{\Upsilon}} = \bar{\mathbf{U}} + \frac{m+1}{m(m-1)}\left[\sum_{i=1}^{m}\left(\mathbf{r}_i - \bar{\mathbf{r}}\right)\left(\mathbf{r}_i - \bar{\mathbf{r}}\right)'\right]$$

where $\mathbf{r}$ denotes the $p\times(p-1)/2$ vector consisting of the nondiagonal and the nonduplicated elements of the $p\times p$ symmetric matrix $\mathbf{R}$. $\bar{\mathbf{R}}$ and $\hat{\boldsymbol{\Upsilon}}$ can be used to fit structural equation models to the average sample correlation matrix with the robust DWLS, WLS, and ULS methods. The corrected robust DWLS and ULS Chi-square test statistic proposed by Chung and Cai (2019) is given by

$$T_B = (n-1)(\mathbf{r}-\boldsymbol{\rho}(\hat{\boldsymbol{\theta}}))'\mathbf{V}(\mathbf{r}-\boldsymbol{\rho}(\hat{\boldsymbol{\theta}}))$$

where

$$\mathbf{V} = \hat{\boldsymbol{\Upsilon}}^{-1} - \hat{\boldsymbol{\Upsilon}}^{-1}\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Delta}}'\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\hat{\boldsymbol{\Upsilon}}^{-1}$$

where $\hat{\boldsymbol{\Delta}}$ denotes the Jacobian matrix of $\boldsymbol{\rho}(\boldsymbol{\theta})$ with respect to the unknown parameters $\boldsymbol{\theta}$ of the structural equation model evaluated at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$. The small sample adjusted T_B test statistic (Yuan and Bentler 1997) is given by

$$T_{YB} = \frac{T_B}{1+nT_B/(n-1)}$$

References

Brand, J.P.L. (1999). *Development, Implementation, and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*. Unpublished Ph.D. thesis, Erasmus University.

Browne, M.W. (1984). Asymptotically Distribution-free Methods in the Analysis of Covariance Structures. *British Journal of Mathematical and Statistical Psychology*, **37**, 62-83.

Chung, S, and Cai, L. (2019). Alternative Multiple Imputation Inference for Categorical Structural Equation Modeling. *Multivariate Behavioral Research*, **54(3)**, 323–337.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

Steiger, J.H. & Hakstian, A.R. (1982). A Historical Note on the Asymptotic Distribution of Correlations. *British Journal of Mathematical and Statistical Psychology*, **36**, 157.

Van Buuren, S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods in Medical Research*, **16**, 219–242.

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, **92(438)**, 767–774.