Multiple Imputation

Multivariate data sets, where missing values occur on more than one variable, are often encountered in practice. List-wise deletion may result in discarding a large proportion of the data, which in turn, tends to introduce bias.

Researchers frequently use *ad hoc* methods of imputation to obtain a complete data set. The multiple imputation procedure implemented in LISREL is described in detail in Schafer (1997) and uses the EM algorithm and the method of generating random draws from probability distributions via Markov chains.

In what follows, it is assumed that data are missing at random and that the observed data have an underlying multivariate normal distribution.

Technical Details

EM algorithm:

Suppose $\mathbf{y} = (y_1, y_2, ..., y_p)^{'}$ is a vector of random variables with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and that $y_1, y_2, ..., y_n$ is a sample from **y.**

Step 1: (M-Step)

Start with an estimate of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, for example the sample means and covariances $\bar{\mathbf{y}}$ and **S** based on a subset of the data, which have no missing values. If each row of the data set contains a missing value, start with $\boldsymbol{\mu}$ = 0 and $\boldsymbol{\Sigma}$ = **I**.

Step 2: (E-Step)

Calculate $E(\mathbf{y}_{imiss} \,|\, \mathbf{y}_{iobs}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and $Cov(\mathbf{y}_{imiss} \,|\, \mathbf{y}_{iobs}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, i = 1, 2, …, N.

Use these values to obtain an update of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (M-step) and repeat steps 1 and 2 until $(\hat{\boldsymbol{\mu}}_{k+1}, \hat{\boldsymbol{\Sigma}}_{k+1})$ are essentially the same as $(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$.

Markov chain Monte Carlo (MCMC):

In LISREL, the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ obtained from the EM-algorithm are used as initial parameters of the distributions used in Step 1 of the MCMC procedure.

Step 1: (P-Step)

Simulate an estimate $\boldsymbol{\mu}_k$ of $\boldsymbol{\mu}$ and an estimate $\boldsymbol{\Sigma}_k$ of $\boldsymbol{\Sigma}$ from a multivariate normal and an inverted Wishart distribution respectively.

Step 2: (I-Step)

Simulate $\mathbf{y}_{imiss} \,|\, \mathbf{y}_{iobs}$, $i = 1, 2, ..., N$ from conditional normal distributions with parameters based on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

Replace the missing values with simulated values and calculate $\boldsymbol{\mu}_{k+1} = \bar{\mathbf{y}}$ and $\boldsymbol{\Sigma}_{k+1} = \mathbf{S}$ where $\bar{\mathbf{y}}$ and $\mathbf{S}$ are the sample means and covariances of the completed data set respectively. Repeat Steps 1 and 2 *m* times. In LISREL, missing values in row *i* are replaced by the average of the simulated values over the *m* draws, after an initial burn-in period.

References

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Monographs on Statistics and Applied Probability 72, Chapman and Hall/CRC.