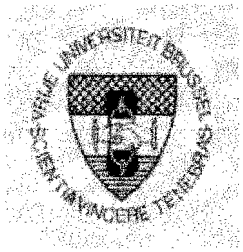


Interuniversity papers in demography

Missing at random data problems and maximum likelihood structural equation modelling

G. Verleye

VRIJE UNIVERSITEIT BRUSSEL
Centrum voor Statistiek en Wiskunde

IPD-Working Paper 1997-3

(Paper presented at the Annual Meeting of the Population Association of America, Washington DC, March 1997)

Interface Demography, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
Tel: 32-2-629.20.40 Fax: 32-2-629.24.20 E-mail: esvbalck@vnet3.vub.ac.be
Website: http://www.vub.ac.be/DEMO
Vakgroep Bevolkingswetenschappen, Universiteit Gent, Sint-Pietersnieuwstraat 49, B-9000 Gent, Belgium
Tel: 32-9-264.42.41 Fax: 32-9-264.42.94 E-mail: sandy.vanlaer@rug.ac.be
Website:http://allserv.rug.ac.be/~jlievens

Verleye G.
Vrije Universiteit Brussel
Centrum voor Statistiek en Wiskunde
Pleinlaan 2
1050 Brussels
Belgium

MISSING AT RANDOM DATA PROBLEMS AND MAXIMUM LIKELIHOOD STRUCTURAL EQUATION MODELLING

ABSTRACT

The aim of this paper is to present the results of a performance study that compares five missing data solutions in the context of structural equation modelling (SEM). By means of a 5 factor simulation approach with multiple numerical and graphical evaluations, 7 research hypotheses are tested. A new and easy applicable method to handle multiple imputed data sets is also presented.

Acknowledgements:

I thank R. Pepermans, M. Despontin , A. Boomsma , J. Arbuckle, K. Bollen, D. Rubin, L. Chuanhai and P. Allison for the useful comments and assistance. This research was supported by Vrije Universiteit Brussel OZR grants.

1. INTRODUCTION

Social scientists use concepts in their theories. Such concepts can be intelligence, memory functions, educational level, socio-economic status and attitudinal components. As a means to validate theories that assume various relationships between such concepts and observed variables, the SEM approach is a most valuable instrument to bring together complex empirical methods and theories. In this case, the latent variables are the concepts for which there is a measurement model that links the concepts to their indicator variables. These indicators are permitted to be unperfect measures in the sense that the SEM technique allows measurement error in the indicator variables. The structural model in SEM then relates the concepts in a linear model. The evaluation of the model happens both on the level of individual parameter estimates and on the level of the entire model (the various goodness of fit measures). An interesting feature of SEM is the capability to model simultaneous relationships, feedback loops and mediating effects. This makes SEM a powerful method that analyzes variables and concepts in their mutual relationships. This implies that the degree of abstraction is more realistic since relationships between concepts and variables can be modelled in an environment of variables respecting complex patterns of relations. An excellent example of this is a SEM application in the study of intellectual development and can be found in Rudinger et al. (1989). Special cases of the LISREL model are well known statistical approaches such as factor analysis models, the confirmatory factor analysis technique, regression models, path analysis, econometric models, panel and wave models. Today the LISREL model is generally recognized and applications are acknowledged in other than social sciences domains. An example is the Hauspie et al. (1996) longitudinal study of Lublin children where determinants of growth in body length from birth to 6 years of age are modelled with SEM.

Social sciences data often contain missing data. The information may be absent for several reasons: respondents can refuse to answer questions, information can get lost, the data collection schedule can be designed so that several parts of the data are missing by design. In a repeated measurements exercise, subjects may not longer attend follow-up sessions and attrition takes place,...

The question dealt with in this paper is how to handle the missing at random data when SEM techniques are applied to such data sets. This study tries to answer that question by means of a performance study relying on Monte Carlo simulation techniques. In the next section we take a closer look at missing data and alternative solutions for the missing data problem in SEM. In the third section an analysis is designed to evaluate the performance of several missing data solutions. The results are presented in section four. In section five, suggestions for further research are presented.

2. MISSING DATA: CONCEPTS AND PREVIOUS PERFORMANCE STUDIES OF MISSING DATA TREATMENT

2.1 Missing data mechanisms and SEM

A very useful conceptualisation of missing data mechanisms is presented in Rubin (1976) and Little and Rubin (1989). "Missing data are called Missing at random (MAR) if the probability of having a particular pattern of missing data in a sample is independent of the values of the data that are missing, but may be dependent on the values of the data that are observed. Missing data are called Observed at random (OAR) if the probability of having a particular pattern of missing data in a sample is independent of the values of the data that are observed, but may be dependent on the values of the data that are missing. If the missing data are independent of the values that are observed as well as the values that are missing the condition is called Missing Completely at random (MCAR)."
An interpretation of the example by Little and Rubin (1989) in a psychological context is a two variable example where years of education is fully observed and the attitude toward an advertisement is only partially observed. If the probability that the attitude is recorded is equal for all individuals regardless of education or the attitude itself, the missingness is MCAR. If the probability of the attitude being recorded varies according to the education of the respondents and does not vary according to the attitude level within an educational level group, we have MAR but not OAR data. Finally, if the probability that the attitude is recorded varies according to the attitude within each educational level group, we have *systematic* missing values, hence neither MAR nor OAR data. Basically, MCAR means that the observations with any missing values are a simple random sub-sample of the full sample. When the data are MCAR then the mechanism is called *ignorable* for both sampling-based and likelihood-based inferences. The mechanism is ignorable for likelihood-based inferences and not for sampling-based inferences if the data are MAR. If the data are neither OAR nor MAR then the mechanism is called *non-ignorable* (Little and Rubin, 1987). This implies that likelihood methods will yield valid statistical inferences (parameter estimates, standard errors) even when the missingness depends on the observed values, as long as the missingness is independent of the values of what is missing.

The strong MCAR conditions are not often met in the real world. In the case of panel research, sample attrition is unlikely to occur completely at random since the means of variables for respondents to follow-up waves appear to be different from the means for non-respondents. Essentially, in order to prove that the missing data process is MCAR, a requirement of many missing data techniques, non-respondents should be traced and probed for their values. If the completed data are a random sub-sample of the sample, we have a MCAR process. The search for and questioning of non-respondents is seldom possible and realistic. Since likelihood based procedures (see further) require only the missing data to be MAR, how could one find out whether the missing data are MAR? Marini, Olsen and Rubin (1987) skip the question and state that the analysis should proceed under the MAR assumption because if this assumption is not made, we are forced to build special models of the way non-respondents differ from the respondents (as in Little and Rubin, 1987). Much like the MCAR hypothesis, such special models are untestable with the

data at hand. Allison (1987, p. 77) formulates: "As with other statistical assumptions, however, the missing-at-random assumption may be a useful approximation even when it is believed to be false". According to Rubin (1976), there is no general test for the MAR assumption because the observed data are always consistent with some MAR model.

Estimating the parameters of a SEM with three different approaches (in this case pairwise deletion, listwise deletion and the full information method) to handle the not MCAR missing data yields different estimates (Verleye, 1996). This and other personal experiences motivate to take a closer look at the quality of SEM estimates when there are MCAR or MAR data by designing a multi factor simulation study. An other example by Rovine (1994) illustrates the same problem: three different approaches to estimate the covariance matrix in the presence of missing data yield three different results.

2.2 Previous research: reflections and conclusions

After analysis of previous research (Glasser (1964), Afifi and Elashoff (1966, 1967), Haitovsky (1968), Timm (1970), Beale and Little (1975), Gleason and Staelin (1975), Kim and Curry (1977), Finkbeiner (1979), Brown (1983), Malhotra (1987), Brown (1994) and Arbuckle (1995a, 1995b)) a number of reflections can be formulated:

- From the previous studies it is clear that Monte Carlo approaches together with real data re-sampling studies are limited in the number of factors that can simultaneously be studied. The number of factors in these studies is at maximum 4 to 5.
- Most studies use artificial data. Only 1 study cited (Arbuckle) is based on (re-sampling of) empirical data. Both approaches have positive and negative elements that should be carefully analyzed before a study actually takes place. The use of complete sets of empirical data implies inevitably that the distribution of the variables is not multivariate normal. Because the results of such performance studies are guidelines for the researcher who has a missing data problem with real data, studies based on simulated data sets should include both non-normal and normal data as a design factor. This is especially so if methods are evaluated that are derived under the normal model.
- Every study cited includes the factor amount of missing data. Since it is very likely that the performance of any method may depend on this factor, each study should include this factor.
- The large majority of the previous simulation work has been done on missing at random data. Older papers do not differentiate between MCAR or MAR processes but in most cases it appears that the missing data is created according to a MCAR process. For practical and theoretical reasons it could sometimes be fruitful to compare MCAR with MAR results since ML techniques can work under the weaker MAR condition while simpler techniques require the process to be MCAR. Including both MCAR and MAR enables to compare the differential performance of various techniques under the more realistic MAR condition. As mentioned before the solution of systematic missing data problems requires special techniques so the methods that work for MCAR and MAR problems can hardly be expected to work properly under systematic missing data patterns without modifications.
- The factor number of (independent) variables appears in 2 studies. Without minimizing the potential role of this factor, we believe that the redundancy factor (e.g. average intercorrelation) is more relevant if a choice between design factors must be made because the estimation of the missing values is based on the information in the predictor variables rather than the number of such variables. Social sciences data also show a degree of redundancy especially in the context of attitude measurement where scales are constructed based on the degree of intercorrelation of variables. This degree of redundancy factor is included in three studies (Timm, Gleason and Staelin, Brown 1983).
- The research already done contains a high variability in the number of replications. This number can range between 10 and 300. In order to obtain a sharp and clear picture a large number of replications is desired although this multiplies the amount of work to be done. Since it merely increases computer time this consequence is of minor importance. A second consequence of a large number of replications is that minor differences between research conditions or methods will lead to significant tests mainly due to the large number of cases in each cell of the design. Statistical power is indeed an

interesting property but can be a burden if every ANOVA interaction term becomes significant mainly due to the very large cell-sizes.

- Sample sizes are variable in 3 studies and it is an interesting factor especially if the goal of the study is to advice the practitioner on which method to choose given his/her sample size. Sofar the inclusion of this factor leads to answers that are linked to the statistical properties of the methods under study.
- One remark concerning the evaluation of the various methods is about the choice of criteria. While it is tempting to use one indicator for the whole of the parameter estimates (e.g. average absolute bias), differences in bias between individual parameter estimates (error variances, regression weights,...) both in direction (under versus overestimation) and magnitude should be inspected.

Because the different studies use a mixture of several evaluative criteria combined with different designs, care should be taken if the results are to be compared. Moreover, some promising approaches seem to work optimal in there very own context only. The regression imputation is an example where findings show that it works best when used for estimation of regression estimates. Even the principal components method shows lower efficiency when applied to factor analysis problems (according to the analysis by Finkbeiner). Some findings however are present in multiple studies and can therefore be generalized as conclusions so far:

- ML approaches are the most efficient under the broad range of design factors. Their overall superiority holds even in the small sample case. A most interesting feature of recent ML methods is that they work under the weaker MAR condition.
- In general, imputation methods such as mean substitution and hot-deck imputation do not yield efficient estimates. Perhaps multiple imputation does a better job.
- Complete cases or listwise deletion is very wasteful as from moderate levels of missing data on. It remains a valid option when there are few MCAR missing data.
- The available information method, also known as pairwise deletion is generally applicable for MCAR missing data problems. Research indicates that this method can be used if the correlations between the variables are rather low. In this case regression methods and ML techniques cannot outperform the pairwise method because of the lower redundancy in the data.
- Direct estimation of the model parameters, in contrast to indirect estimation, seems to be a valid option.
- Assigning means to missing values is poor in comparison with pairwise and listwise deletion, regression methods, principal components solutions and ML techniques.

3. METHODS AND RESEARCH DESIGN

In the next section five different approaches are presented that will be used to handle non-systematic missing data problems. For many people that apply SEM, 'listwise' and 'pairwise' deletion are the two available options (apart from mean substitution available in many commercial statistical programs or similar response pattern imputation in PRELIS (Jöreskog and Sörbom, 1993)). The three other methods are recent ML techniques that have, to our knowledge, not yet been compared in one performance study.

3.1 Two quick procedures

3.1.1 The complete cases method

This case, often called "listwise deletion", uses the N_L cases where all K variables are observed. Under the MCAR assumption, the complete cases are a random sub-sample of the original cases and discarding cases does not bias estimates. If the MCAR condition is satisfied, this approach has many advantages. Standard complete data analysis methods can be applied without modifications (Little and Rubin, 1987, p.40). Univariate statistics can be compared since all such parameters are computed on the same number of cases (Little and Rubin, 1987, p.40). Listwise deletion leads to a consistent estimator of the parameter vector ω, using maximum likelihood, as long as $N_L \rightarrow \infty$ when $N \rightarrow \infty$ (Bollen, 1987). When

$N_L > K$, the S_L (the observed covariance matrix calculated using the N_L complete cases) is positive-definite, provided the model implied covariance matrix Σ is positive-definite (Dijkstra, 1981). However, N_L , the number of cases used in the analysis, is always smaller than N , the total number of cases. The loss of cases and information can be severe. The analysis of S_L , the covariance matrix for the remaining cases, leads to less efficient estimators than the analysis of S using the full sample (Bollen, 1987, p.370). Applied to the example from section 2.1, it is clear that if the missing data are such that the attitude variable is less missing for cases with more education, the marginal distribution based on complete cases analysis of the attitude-toward-the-ad variable will be distorted by the overrepresentation of people with more education. The estimation of the correlation coefficient is subject to bias because of the overrepresentation.

One may test the MCAR assumption in the case of the analysis of complete cases through a comparison of the distribution of the N_L cases with the distribution of the available cases for each variable (Little and Rubin, 1987). A discrepancy between the two distributions is an indication that the data are not MCAR because the sub-sample based on the N_L cases is not a random sample out of the N cases.

3.1.2 The available information method

Next to the complete cases method, a second procedure called "available information", or "pairwise deletion" is often used in covariance type estimates. In the case of pairwise calculation of the covariance matrix, the measures of the covariance for X and Y are based on the N_P cases where N_P stands for the number of cases for which we have values for X and Y at the same time. Although alternative computational versions exist (see Little and Rubin, 1987), the covariance estimate is obtained as:

$$S_{XY}^{N_P} = \frac{1}{N_P} \sum_{i=1}^{N_P} (x_i - \bar{x}^{N_P})(y_i - \bar{y}^{N_P}).$$

If the missing data process is MCAR, then this estimate is consistent Although this approach uses all information available, its practical utility is limited for at least two reasons. First of all, if the pairwise deletion method is used to estimate a covariance matrix, one cannot provide the sample size for the entire matrix since this sample size can be different for each pair of variables. This creates a problem to determine the value of N toward the SEM program. The chi-square tests of model fit and the estimated asymptotic standard errors are sensitive to the choice of N (see Bollen, 1987). A second problem is that if the number of variables K increases, the resulting covariance matrix S could not be positive-definite. SEM programs such as LISREL can handle non-positive definite covariance matrices by adding a constant times the diagonal of S to S (see Jöreskog and Sörbom, 1989).

3.2 Three Maximum likelihood approaches for ignorable missing data

The advantages of maximum likelihood based approaches to be treated in this section relative to complete cases and available information are twofold: complete cases and available cases procedures are consistent but not efficient, providing that the missing data process is MCAR. If the data are only MAR, the estimates can be biased. ML estimates are consistent and efficient under the less restrictive MAR condition. The three methods that follow are potentially fruitful in treating MCAR and MAR missing data problems for three reasons. First of all they can be applied both with MAR and MCAR missing data. Second, a review of the literature shows the potential of the ML approach. Finally, these 3 ML methods are very distinct, so the common promising ML characteristic shared is applied in different ways. This

might lead to interesting findings when missing data of various kinds in data sets with different characteristics are treated with these methods.

3.2.1 The Expectation Maximization (EM) covariance matrix estimator under the normal model

The EM method is an indirect way to deal with missing data problems because it is a method that results in an alternative (but efficient) estimate of the covariance matrix S that serves as input to a SEM program. The algorithm handles missing data in an iterative way. Each iteration consists in two steps. In the first step the missing data are replaced by estimated values (such as the complete cases estimates). Next, the parameters are estimated and in the third step the missing values are re-estimated assuming that the parameter estimates from step 2 are correct. In the fourth step the parameters are re-estimated again and so forth. This goes on until the process converges. One of the advantages of EM is the fact that under general conditions, the loglikelihood is increased in each iteration. In other words, it converges reliably. On the other hand, the convergence of EM can be slow if a lot of data are missing. Dempster, Laird and Rubin (1977) show that its convergence is linear with rate proportional to the fraction of missing information defined in terms of the eigenvalues of the information matrix. Of course we need an initial set for the parameters. If we have at least $K+1$ complete observations, the complete cases method provides efficient estimates if we have MCAR missings (which is not so realistic). A second method requiring none of the two conditions and yielding good starting values implies the use of the available information for the univariate parameters and setting the covariances to zero (see Little and Rubin, 1987, p.143).

The FORTRAN 77 routine that computes ML estimates under the normal modelof the covariances with EM, is made available for this study by Donald B. Rubin and written by Chuanhai Liu.

3.2.2 Full information estimation in SEM

Allison (1987) described an alternative modelling approach applicable in LISREL which yields ML estimates that are consistent, asymptotically efficient and asymptotically normally distributed in the presence of MCAR or MAR missing data. The approach also produces consistent estimates of the standard errors of the parameter estimates. A similar approach for confirmatory factor analysis was proposed by Werts, Rock and Grandy (1979) and for LISREL models by Baker and Fulker (1983). Muthén et al. (1987) described a similar technique to be implemented with EQS. Lee (1986) also described a direct ML estimation method for structural equation models with missing data. The basic idea presented in the paper by Allison is to use data from sub-samples with different sets of observed variables. One of these sets may be a fully observed set. For a three variable case with one variable partially missing, say X_1, two sets are required. The first set is fully observed and in the second set the information on X_1 is missing. If two variables, say X_1 and X_2 are partially missing and there is no simple missing values pattern, 4 sets are required: a completely observed set, a set where X_1 is missing, a set where X_2 is missing and a set where both X_1 and X_2 are missing. The missing variables are then treated as latent variables. For all samples simultaneously the model is then estimated with appropriate equality constraints imposed across the sub-samples. This is possible using a combination of the structured means method and the multiple group approach, both present in later versions of the LISREL program.

An interesting property of this ML approach is that it uses all the available information. This ML approach takes into account the overidentifying restrictions, present in most confirmatory factor analyses and simultaneous equations models, as opposed to methods where an alternative covariance estimator is used as in pairwise and listwise procedures and other ML estimators such as EM. This is a requirement for an efficient ML estimation of the model parameters (Allison , 1987).

As described by Allison (1987, p.74) "the method is primarily useful when the number of sub-samples with distinct sets of variables present is relatively small and when the number of cases in each sub-sample is large". In theory the method yields valid estimates for general missing data problems with numerous

sub-samples. In practice there are two drawbacks. First, the number of missing data patterns must be limited, which is not the case in many practical problems. Second, the implementation of the procedure needs a very high level of expertise with SEM "programming". These two drawbacks could however be minimized if a LISREL (or other SEM program that does multiple group analyses) preprocessor would have been developed that writes the program code lines needed to implement the approach upon reading the data and program code for the essential model (the code that would have been written if there were no missing data). This has never been done (Paul D. Allison, Kenneth A. Bollen in personal communications). Of course, even if such a preprocessor existed, many data sets would not satisfy the need for large sub-samples. Arbuckle (1995a, 1995b) incorporated these idea's in a generalization of the ML estimation in confirmatory factor analysis with incomplete data by Finkbeiner (1979). This method works with almost every pattern of missing data (Finkbeiner, 1979).

Let μ_i and Σ_i be the population mean and covariance matrix for the variables that are observed for case i. Each μ_i can be obtained by deleting elements of μ, and each Σ_i can be obtained by deleting rows and columns of Σ.

If we further assume multivariate normality, the loglikelihood of the i -th case is

$$\ln L_i = A_i - \tfrac{1}{2}\ln|\Sigma_i| - \tfrac{1}{2}(x_i - \mu_i)\Sigma_i^{-1}(x_i - \mu_i)'$$

where A_i is a constant that depends only on K_i. The loglikelihood of the total sample is then

$$\ln L(\mu, \Sigma) = \sum_{i=1}^{N} \ln L_i .$$

Given a structural equation model that specifies $\mu = \mu(\omega)$ and $\Sigma = \Sigma(\omega)$ as a function of some parameter vector ω, ML estimates of ω are obtained by maximizing

$$\ln L(\mu(\omega), \Sigma(\omega)),$$

or by minimizing

$$C(\omega) = -2\ln L(\mu(\omega), \Sigma(\omega)) + 2\sum_{i=1}^{N} A_i = \sum_{i=1}^{N} \ln|\Sigma_i| + \sum_{i=1}^{N} (x_i - \mu_i)\Sigma_i^{-1}(x_i - \mu_i)' .$$

The structural equation programs AMOS (Arbuckle, 1993) and Mx (Neale, 1994) use this approach to handle missing values. The full information method is a direct approach because the model parameters are estimated in the presence of missing data.

3.2.3 The multiple imputation approach

Next to the complete cases method and the available information method, various imputation methods are often used to handle the missing data problem. Imputation methods replace each missing value by a value making the data matrix free of missing data.

The basic problem with one imputation for each missing value is that they are treated as if they were known. Even if we perfectly understood the missing data process, the estimates tend to be too sharp because the extra variability due to the unknown missing values is not taken into account. If the missing

data process is less understood, single imputation does not control for the extra uncertainty due to the unknown missing data process (Rubin, 1987). In multiple imputation we replace each missing value by M values. After multiple imputing a data set containing missing data, we have M completed data sets. The two main advantages of single imputation remain : as we have complete data sets, numerous types of statistical techniques can be applied. Secondly, the know-how of the database administrator can be incorporated in the way we impute the missings. Compared to single imputation, multiple imputation implies that we have to analyse M data sets in an identical way which means more effort to do a statistical analysis. In general, the final results are obtained by bringing these M estimates together in a proper way (see Little and Rubin, 1987, p.257). In fact, multiple imputation has two important advantages compared to single imputation (Rubin, 1987): First of all, when imputations are randomly drawn in an attempt to represent the distribution of the data, multiple imputation increases the efficiency of estimation: as $M \rightarrow \infty$, $\hat{\omega} \rightarrow \omega$ with correct standard errors. This happens because the uncertainty about the real values of the missing data are now considered by using M values instead of one. In practice we can use $M = 3$. Secondly it is possible to apply M different models to impute values to reflect the uncertainty about the process that causes the missing data.

During the last decade quite a number of algorithms have been developed that perform multiple imputations under a specific model for the missing data (e.g. Tanner and Wong, 1987 and Rubin and Schaffer, 1987). According to Rubin (1987), the most fruitful approach is a Bayesian method which yields inferences with good frequentist properties. In this approach a prior is added to the likelihood and the inferences are based on the posterior distribution. Because of the complexity of the likelihood function it does not allow for explicit expressions for marginal posterior distributions of the parameters. These distributions can be approximated by simulation. For monotone patterns (the variables can be arranged so that for $j = 1,2,\ldots,K-1, X_j$ is observed whenever X_{J+1} is observed) we first draw μ and Σ from their posterior distribution. In the second step we draw the missing values from their posterior distribution conditionally given the drawn values of the parameters (for more details see Rubin (1987)). The first step is important because if we do not draw the parameters and use the observed parameters, we act as if the respondents' distribution of X values were exactly the same as the population distribution of X values. In this way we underestimate the variability. With ignorable non-response, the respondents and the non-respondents share the same parameters, but the sample mean and the sample variance for respondents are not perfect estimates of these parameters, and our imputations must reflect this uncertainty to be proper.

An efficient statistical algorithm that imputes properly for ignorable missing data under the condition that the observations of the K variables are independently identically distributed with a normal distribution $N_K(\mu, \Sigma)$ is presented in two research papers by Liu Chuanhai (1992a, 1992b).

In practice the M multiple imputed data sets can be analyzed in SEM programs using the multiple group approach. In this indirect method, each data set is then assigned to a group and the parameter estimates are estimated with equality constraints over the M groups (see Jöreskog and Sörbom, 1989). The FORTRAN 77 routine that does the multiple imputation is made available for this study by Donald B. Rubin and written by Chuanhai Liu.

3.3 Study Design

This study can be considered a factorial design performance study with 5 factors. The factors being:

- the five missing data solutions already mentionned,
- the type of SEM sub-model : a measurement model versus a full model,
- the percentage of missing data for each variable,
- the kind of process creating the missing data : MCAR versus MAR,
- multivariate normal versus non-normal data.

The steps that will allow us to study the performances are :

First, using the SIMCHI approach (Verleye, 1996), draw 300 data sets with a fixed covariance matrix that goes with the model of interest and distribution type. The measurement model is the 4CM model used by Boomsma (1983) in his robustness study. It is a 2 Correlated factors model with 4 indicators for each latent variable and Medium sized factor loadings. The full SEM model used in this study is the "Peer influence on aspiration model" by Duncan, Haller and Portes (1968). It is selected for the same reason we decided to use the 4CM model : this model is definite a reference model. It is also discussed in Jöreskog and Sörbom (1989), Boomsma (1983)and Fox (1984). When we fix all model parameters, this implies a covariance matrix. So we know the exact values of the model parameters, fit indices and given the sample sizes, the standard errors. These are the reference values. One half of all the data sets will be drawn under a multivariate normal model while the other halve will be non-normal data. In this study the non-normal distributed variables are always χ_3^2 variables. Since we want to exclude potential sample size effects on estimation, we decided to use large data sets and fix N at 1000 observations. According to the terminology used by Boomsma (1983), we work with large samples thereby avoiding ML convergence problems and sample size effects on the sampling distributions of the SEM parameter estimates and the goodness of fit statistic. Next to the question of sample size we address the question of the number of repetitions NR in each cell of our factorial design. It is well known that in this type of Monte Carlo research where a model is used to generate data, generally called distribution sampling (Kleijnen, 1974), the error in the Monte Carlo estimation of statistics from the sampling distributions decreases with a factor $NR^{-\frac{1}{2}}$. Following Kendall and Stuart (1969), it is not easy to fix NR at a value that is safe. Most often, values below 100 are suspect. Following Boomsma (1983) we took NR=300 because this value is far beyond the 100 threshold and the standard errors are decreased by a factor of almost 2 in going from NR=100 to NR=300. This step is related to the type of SEM sub-model factor and to the multivariate normal versus non-normal data factor.

In step two the factors percentage of missing data and the missing data process are considered. Values will be deleted by means of two processes : a MCAR and a MAR process. In the MCAR case, the values are deleted completely at random. In order to obtain a missing data process that results in MAR and not OAR missing data, an adaptation of a process described by Rubin (1976, p.583) was used. As an example, take 1000 one variable observations. If the sum of the first N_1 values exceeds some predefined value, then all values that come after N_1 are made missing.

Three types of data sets will be created: each variable has 5%, 15% or 25% missing values. We are aware that 25% missing data is perhaps more than what is encountered in research. However it may be fruitful to incorporate such a high level since this may highlight differences between the performance of the various potential methods to handle missing data that are not so explicit with lower degrees of missing data. In their 1975 paper Gleason and Staelin (1975) advise to "go beyond what is missing in empirical work" for the same reason.

The missing data will be treated with the 5 techniques, the models will be estimated and subsequently the results will be compared to the reference values according to a number of criteria. This is the remaining factor.

In total, 36000 output files (300 repetions x 2 distribution types x 2 missing data processes x 3 levels of missing data x 5 approaches to handle the missing data x 2 models) are to be processed.

To evaluate the performance of the five techniques two sets of criteria are applied. We decided to combine two approaches that together form a rather exhaustive schedule:

1. numerical indicators of (1) non-convergence and improper solutions, (2) bias of parameter estimates, (3) bias of estimates for standard errors, (4) confidence intervals for parameter estimates, (5) confidence intervals for the mean of standardized parameter estimates, (6) the chi-square statistic for goodness of fit, (7) dependencies between parameter estimates and their corresponding standard error, (8) normality tests for the standardized parameter estimates.
2. graphical analysis of the standardized parameter estimates and the goodness of fit statistic: compare the theoretical sampling distribution with the empirical sampling distribution.

4. RESEARCH HYPOTHESES

In this section a number of hypotheses are presented. These statements are motivated by means of the special features that characterize the 5 different methods tested in this study. The hypotheses are:

1. Given the redundancy (due to the non-zero correlations) in the 2 covariance matrices that are used in this analysis, the EM maximum likelihood solution and the multiple imputation procedure should be more efficient compared to listwise and pairwise deletion. One can verify that the 2 quick methods do not use the redundancy in the data while the two ML methods estimate parameters using the information in the non-zero correlations between the variables.
2. The higher efficiency of ML solutions should be more pronounced in the MAR case than in the MCAR case. ML methods are still efficient under the weaker MAR condition, while the 2 quick methods require MCAR data.
3. The full information approach implemented in AMOS should be superior to EM estimation of the covariance matrix and the multiple imputation method. The main reason for this is that AMOS uses a direct estimation procedure. A more fundamental reason is that the EM covariance estimates and the multiple imputation estimates of the missing values do not take into account the identification of both SEM models. In fact both models are overidentified. According to Allison (1987, p.79), "in order to have efficient parameter estimates in the presence of missing data, the overidentifying restrictions should be incorporated in the ML estimation procedure".
4. The 5 procedures do not perform differently for a measurement model and a full model. None of the 5 methods should be superior for one of both models. All 5 methods do not use techniques that favour a factor model or a full structural model.
5. In comparing the relative efficiency of the techniques, better results for the ML methods are expected in the presence of normal data. The reason for this are the normality assumptions that underlie the three ML methods. However, since the SEM ML procedures are developed under the multivariate normal distribution these discrepancies should be analyzed with caution. Some results (in the cases where raw data files are used as input to SEM programs: liswise deletion, multiple imputation and the available information method) may very well be partially explained because SEM estimates may be sensitive to departures from the normality assumption. Research by Boomsma (1983) shows that although departures from normality do not seem to influence the parameter estimates, inaccurate standard errors may be obtained.
6. The 2 quick methods, pairwise and listwise deletion should work equally well under normal and non-normal distributional conditions because no distributional assumptions are present in their heuristics.
7. In the presence of few missing data (5%), smaller differences are expected between the performance of the 2 quick methods and the three ML methods compared to the situations with moderate (15%) and higher (25%) fractions of missing data. This effect should be noticeable because the two quick methods are both characterized by an absence of effort to do something about the missing data problem. The complete data procedure even throws away observed information as seen earlier. The pairwise method tries to use the maximum information available but leaves the missing information as such. This is in contrast to the two indirect ML procedures that try to use the redundancy in the data. The direct method in AMOS is also tuned to handle missing data in a more active approach as seen before. We therefore expect increasing discrepancies between the quick methods and the ML approaches as the amount of missing data increases.

5. RESULTS

In this section the conclusions based on the analyses of the results for both the measurement model and the full structural model are reviewed. Given this information, the question raises whether the research hypotheses presented stand the test. The relationship to previous research efforts are also highlighted.

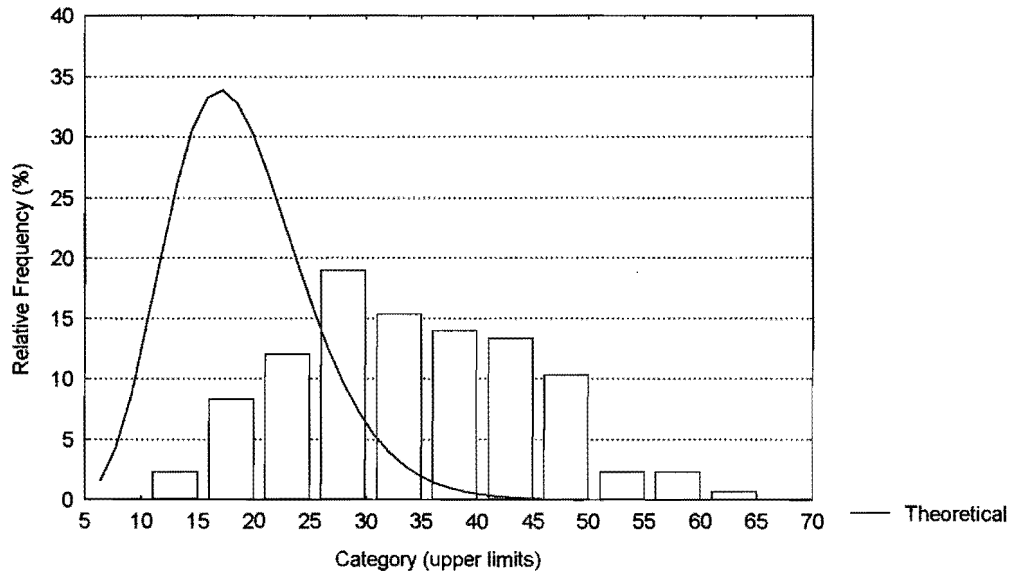
5.1 The research hypotheses

According to the first hypothesis, the EM maximum likelihood technique and the multiple imputation method should outperform the two quick procedures (listwise and pairwise deletion) because there is redundancy in the covariance matrices. This redundancy is used in these two indirect ML methods. From our analysis of the results it is clear that this hypothesis is confirmed:

- The two ML methods always yield convergence and there are no improper solutions. The performance of the listwise deletion method is clearly worse. However, the pairwise deletion method leads neither to such convergence nor to improper solution problems.
- The bias of the parameter estimates after treating the missing data problem with an indirect ML method is smaller than the bias obtained with the two quick methods. Pairwise deletion shows better results (smaller bias for the parameter estimates) than the listwise deletion method. However, the pairwise method leads too frequently to rejection of a correct model. An example is this can be seen in figure 1.
- If the information in both the bias of the parameter estimates and the standard errors is combined, the higher efficiency of the two indirect ML techniques is confirmed.

Figure 1.

Theoretical versus sampling distribution of the Chi2 GFI.



The second hypothesis stated that the two quick methods should perform worse under MAR versus MCAR conditions. The hypothesis also said that the three ML techniques should have equal performance levels under both MCAR and MAR conditions. From the analysis of the tables, it is clear that ML methods work equally well under both MCAR and MAR conditions. This is not the case for the listwise method: this procedure performs worse under MAR conditions than under MCAR conditions. In terms of bias for the parameter estimates, the pairwise deletion method yields results that are similar for MCAR and MAR.

This conclusion can also be drawn if combinations of bias in parameter estimates and the standard errors are jointly studied. The only performance indicator for the pairwise deletion method that is influenced by the nature of the missing data process is the chi-square goodness of fit. The pairwise deletion method shows less model rejections under the MCAR condition than under MAR condition.

Following the third hypothesis, the full information procedure present in AMOS should yield better results than EM and multiple imputation because it is a direct estimation method. A number of findings support this hypothesis. In the case of the measurement model the bias of the parameter estimates (for the 5% missing values condition) is better with AMOS compared to the four other missing data methods. Although AMOS slightly overestimates the standard errors when the fraction of missing data is moderate or high, the results from the analysis of the confidence intervals for the parameter estimates consistently indicate that direct ML yields the best estimates. This full information method shows smaller correlations between the parameter estimates and the standard errors compared to the two indirect ML methods. The standardized parameter estimates computed with AMOS are standard normal distributed. With the two indirect methods, the variances of these standardized parameter estimates are larger than 1. One further advantage of AMOS compared to the two ML methods is that there is no need to pre-process the data with customized software. Both EM and multiple imputation require such a step. One can easily spend one hour preparing thedata followed by adaption and subsequent application of a single FORTRAN program to obtain the EM covariance matrix or imputed data sets.

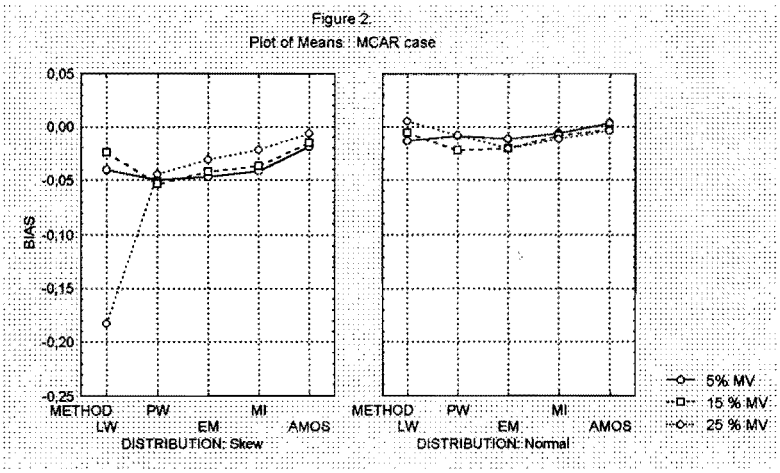
According to hypothesis 4, the results obtained for the measurement model should be equal to those obtained for the full structural model. None of the five techniques dealing with missing data problems yields consistently better or worse results for one of the two models. Analysis of the tables and figures supports this hypothesis. Omitting one of the models would certainly not lead to different conclusions. There are however small differences:

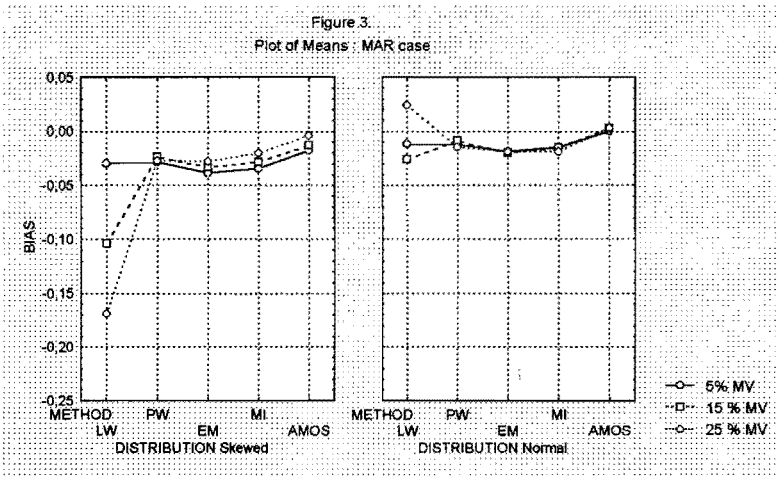
- In the case of the full structural model, the listwise deletion procedure yields non-convergence and improper solutions under both MCAR and MAR conditions. For the measurement model, this only occurred under the less restrictive MAR condition.
- The underestimation of the parameter estimates is for the listwise deletion procedure a little more pronounced in the full structural model case compared to the measurement model results.
- The overestimation of the standard errors computed by AMOS for the full structural model holds for every fraction of missing data, while in the case of the measurement model this clear overestimation only occurred for moderate and high fractions of missing data. Both this and the two previous findings are maybe related to the fact that the full structural model is more complex than the measurement model.
- The measurement model shows some dependencies between the parameter estimates and the corresponding standard errors to be unstable for the pairwise solution. Such discrepancies are not seen in the results of the full structural model.

Because the three ML techniques are developed under the normal model, following hypothesis 5, we expected better results under normality in the data. In contrast to that, no impact of the distribution of the data is expected for the two quick methods because these techniques do not require distributional assumptions. This was in fact hypothesis six. As for the three ML methods, no improvement in parameter estimation due to normality is observed. The only result is that the differences between the parameter estimates for the ML methods become smaller when the data are normally distributed. No single best method exists however. The bias for the standard errors is not influenced by the presence of normality for the three ML methods. In the case of listwise deletion, the bias in the parameter estimates in the normal data case is smaller compared to this bias in the presence of skewed data. As seen before, this may be due to LISREL which is known to yield better parameter estimates under normality conditions in small samples. No relationship between distributional characteristics and the quality of the parameter estimates is noticed for the pairwise deletion procedure. For the two quick methods, the bias of the standard errors is not influenced by the skewness of the data.

The last hypothesis stated that larger differences between the two quick methods and the ML procedures could be expected as the amount of unobserved data increases. This hypothesis is not rejected and the effect is especially observable for the bias in the parameter estimates. Large differences are present for the bias of the parameter estimates with 25% missing data. The differences are smaller with 5% missing data.

A general picture of the results can be seen in figures 2 and 3. The data are the lambda parameter values from the measurement model.





5.2 Previous research

Our results from the performance study clearly line up with findings from the other sources. Yet our study tried to test them in one overall research design. We will first present the literature findings that match with our results.

- Haitovski (1968) concluded that in the presence of a lot of missing data, the complete cases method (listwise deletion) is worse than the available information method (pairwise deletion).
- According to Beale and Little (1975) ML outperforms the listwise deletion method. They also found that different ML approaches yield not very different results.
- Kim and Curry (1977) concluded that the pairwise method is better than the listwise procedure.
- Finkbeiner (1979) concluded that direct ML is best. Another finding is that the pairwise deletion method yields estimates that are close to ML estimates while the listwise procedure is much worse. This listwise procedure was found to improve as less data are missing.

- Brown (1983) found the pairwise method more efficient than the listwise procedure. The EM technique outperforms the pairwise method. All ML procedures yield similar high quality estimates that are better than the pairwise deletion results.
- Malhotra (1987) concludes that the EM method outperforms the listwise deletion method. He also finds that as the fraction of missing data decreases, the differences between the methods decrease.
- Brown (1994) states that the listwise deletion method is wasteful. There is a tendency for error variances and structural disturbances (Φ, Ψ, Θ_δ and Θ_ε to be more biased than other parameters. Underestimation of structural parameters γ is consistent. Listwise deletion yields overestimated standard errors. These are larger compared to other methods. The pairwise deletion method yields good estimates for the standard errors.
- Arbuckle (1995a, 1995b) indicates that both the pairwise method and the direct ML procedure provide very good estimates of the parameter values under the MCAR condition. The ML estimator is more efficient and is normal in shape. The pairwise method yields results that are in between the results of listwise deletion and direct ML. Arbuckle concludes that under MCAR conditions, the ML method is superior to the pairwise and listwise results. Under MAR conditions, the direct ML method outperforms the pairwise and listwise method. With normal distributed data, the direct ML method performs better than under non-normal data conditions.

In a few circumstances however, our results do not match.

- Timm (1970) found that the listwise deletion procedure does not yield non positive-definite covariance matrices.
- According to Finkbeiner (1979), the pairwise deletion method can result in non positive-definite covariance matrices.
- Brown (1994) finds that with large samples the listwise deletion method shows no convergence problems. Another conclusion says that the pairwise deletion method shows a constant low level of model rejections and that imputation methods show stable rejection rates across the MCAR levels.

A possible explanation for these findings is that they may be related to the model used in the Monte Carlo study. Each model uses different covariances, so in some cases obtaining positive definite matrices may be more likely than in other. For the model rejection, Brown (1994) used not the same test as we did. Our analysis uses the likelihood ratio test (L^2) while Brown (1994) applied the $\frac{L^2}{df}$ test. Both tests suffer from a sample size effect as already mentioned (Bollen, 1987, p.278). The sample size used by Brown (1994) is half the size of the samples we used. These can be reasons why different outcomes are observed.

6. DISCUSSION

In this section, three items are treated:

1. suggestions for the SEM user with missing data,
2. the limitations of the performance study,
3. suggestions for future research.

First of all, it is our belief that the missing data should be analyzed before any other statistical analysis takes place. In chapter three some suggestions are presented in order to accomplish this. The question is "Do we have MCAR, MAR or systematic missing data?". As seen in chapter three, missing data that is not MCAR is easy to trace. In many occasions it is not so obvious to decide that the missing data process is systematic and not MAR. This for two reasons: the observed data will be consistent with some MAR process and one does not have access to the real values of the missing data. "This is especially true in the case were studies are designed intentionally to have missing data. Such studies are rare, but they might become more common now that good methods of handling large amounts of missing data are available"

(James Arbuckle in a personal communication). References on "planned" missing data are Duncan and Duncan (1994), Graham et al. (1996), Kaplan (1995) and McArdle (1994).

In general, for data that is to be modelled with SEM, it is our belief that the two quick methods (listwise and pairwise deletion) are better not applied. First of all, it is highly unlikely (especially in attitude research) that every missing data process (for every variable that contains missings) is MCAR. Secondly, even if the process can be assumed MCAR, these two methods have their proper drawbacks: listwise deletion is wasteful and for both deletion methods, the standard errors are not exact. One could argue that the study showed that the pairwise deletion method yields acceptable results in many occasions. However, we prefer one approach that is generally applicable. In this case it is clear that the direct estimation procedure is to be preferred. If for some reason an indirect method is needed, we highly recommend to apply the EM maximum likelihood estimation of the covariance matrix. Although EM was not explicitly compared with the similar pattern approach available in PRELIS, the EM method most probably outperforms the similar pattern imputation. Similar pattern imputation is a single imputation technique that is not always applicable. More important, it shares with other simple single imputation methods a common characteristic: there is no uncertainty added to the imputed value yielding an analysis that treats imputed values just like observed values. Although multiple imputation does not suffer from this, in practice it requires more extensive data preparation together with elaborate SEM programming. There is also a problem with the data itself: in order to repeat a SEM analysis that was carried out by someone who prepared multiple imputed data sets, one needs those data sets or at least a very explicit description of the approach and models that were used to obtain the M imputed data sets. It is much more convenient to add a covariance matrix (EM) to a paper. Of course, one can argue that to repeat the direct estimation procedure, one also needs raw data. However, no complex imputation information is required. Hopefully, this kind of performance analyses will persuade researches to apply a more efficient procedure to handle their missing data problems. The direct estimation procedure and EM estimation of the covariance matrix are now at every one's disposal so this should be possible.

The nature of a Monte Carlo study implies limitation of the number of factors that can simultaneously be analyzed. This analysis explored the impact of 5 factors: the type of model, the missing data process, the fraction of missing data, the distributional characteristics of the data and finally the method that is used to deal with the missing data. Before considering the relevance of other factors, it is interesting to look at the limitations of each factor that was present in this analysis.

This performance analysis was repeated for two models: a measurement model and a full structural model. It was found that the conclusions for both models were very similar, except for some minor differences. One could however ask to what extent the results from this study can be generalized. Given that statistical models such as regression analysis, causal path analysis and many more as we saw previously, are special cases of the general LISREL model, we believe that the results from this study are applicable for all LISREL sub-models. In fact, the results should be applicable to every analysis which models a covariance structure because that is what SEM does. Of course, a "proof" of this statement implies the test for every sub-model which leads to unpractical proportions of the study. Given that the algorithms of the five procedures do not favour a specific model with covariance data, the applicability is broad.

The second factor considered is the missing data process. Two types were analyzed in this study. The conclusions for the MCAR process are without questioning applicable in every MCAR situation for the class of LISREL models. However, although the MAR method used in this study is a procedure that satisfies the criteria for MAR, other missing data processes that are MAR exist. The efficiency of the three ML methods is independent of the kind of MAR process, as long as it is MAR. Maybe the two quick methods yield different results for different MAR processes. The following example is a missing data process that is also MAR but different from the MAR process used in our study. In the analysis by Arbuckle (1995a, p.17) "SENTENCE and PARAGRAPH scores were deleted (i.e., made missing) for every examinee whose WORDMEAN score was 12 or less, and retained for examinees with

WORDMEAN scores greated than 12." SENTENCE, PARAGRAPH and WORDMEAN are the scores on three psychological tests.

Anyway, it is known that the two quick procedures are not suited for MAR problems (lack of consistency and efficiency) so we believe that it is not very worthwhile to analyze the potential influence of the kind of MAR process on the efficiency of the quick methods. Even if this kind of exercise would yield results, the importance of such findings can be questioned.

The three levels of the fraction of missing data (5%, 15% and 25%) were chosen to cover the range found in real research. Data sets with fractions of missing data beyond 25% for each variable are sparce. In such occasions, one should question the quality of the data and the relevance of advanced statistical procedures applied on poor data. The question however remains to what extent missing data solutions yield efficient results with even larger fractions of missing data. In other words, what are the limits for the missing data methods? How much observed data do they need do work properly? Again, this calls for a separate analysis.

Real data to be treated with missing data techniques and subsequently analyzed with covariance structure models will probably not be distributed normally. The non-normality of the data that was used in this study is not excessive. (In this study the non-normal distributed variables are always χ^2_3 variables.) The purpose of this factor was to take into account departure from normality because too many simulation studies are based on normal distributed data. A reason for this may be the fact that random generation of multivariate data with a fixed covariance matrix and controllable deviations from normality is not often used (not easy available) nor known. As part of the preparation for this simulation study we developed SIMCHI, a method that appears to work satisfactory. As seen before an alternative is to use real data where missing values are created. We did not use real data because we wanted to include the normality factor in the design. Our results are therefore only applicable to data sets where the data are distributed normally or with moderate deviations from normality. Perhaps one could assess the limits of the qualities of the ML missing data techniques in the presence of excessive non-normality: skewness as well as kurtosis.

This Monte Carlo performance analysis compares five missing data techniques. Even if the inclusion of the present five techniques was motivated , the inclusion of other or more methods could have been interesting. Simple methods, known to be unefficient (mean substitution, etc.) or only working under limited conditions (e.g. the principle components method) are uninteresting to re-analyze because of their known unefficiency for general application. An interesting future excercise could be to compare the similar response pattern method now available in LISREL with direct ML. This technique was not available at the time this analysis was designed. An other potential fruitfull appraoch is presented by Steinberg D. and Colla P. (1995) and Breiman L. et al. (1984) and implemented in the CART program .

Next to the factors that are present in this performance analysis, other factors might be included in a future excercise. This study uses two models that are correctly specified. In a following step, it might be interesting to assess the efficiency of missing data techniques that are applied to models that are not correctly specified. Our results obtained with correctly specified models suggest that the direct approach used in AMOS is a most promising approach. Extra research efforts are needed to illuminate if the results available now still hold for non-correctly specified models. Two ML methods used in this study (EM and multiple imputation) need redundancy in the variables (covariance), as previously treated. How large must the correlations minimally be so that these methods can yield results? One could argue that this question is not relevant since people will not model data with SEM in the absence of covariance between the variables. However, it is our belief that this could be an interesting question in order to obtain a full description of the characteristics of the two methods.

Another line of future research is in the domain of systematic missing data where the missing data process is non ignorable. The joint information in Arbuckle (1995a, 1995b), Brown (1994) and this report covers

the domain of MCAR and MAR for SEM analyses rather well. This is in sharp contrast to the limited information on the treatment of systematic missing data in SEM applications. Examples of interesting approaches that could be fruitfull can be found in Heckman (1979) and Muthén and Jöreskog (1983).

Further research on systematic missing data should be encouraged. Even if such methods require information on the exact nature of the missing data process in an analytical form that is difficult to obtain (analysis of previous research efforts might help), such methods are needed because one cannot always proceed assuming that MAR techniques can do a proper job when it is clear that the missing data is not MCAR nor MAR.

Bibliography

Afifi A.A. and Elashoff R.M., 1966, Missing observations in multivariate statistics I, Journal of the American Statistical Association, 61, 595-604

Afifi A.A. and Elashoff R.M., 1967, Missing observations in multivariate statistics II, Journal of the American Statistical Association, 62, 10-29

Allison P.D., 1987, Estimation of linear models with incomplete data, in C.C. Clogg (ed.) Sociological Methodology, San Francisco: Jossey-Bass, 71-103

Arbuckle J., 1993, ,AMOS: the Analysis of MOment Structures, computer program, dept. of psychology, Temple University, Pennsylvania

Arbuckle J., 1995, Full information estimation in the presence of incomplete data, working paper, dept. of psychology, Temple University, Pennsylvania

Arbuckle J., 1995, Advantages of model-based analysis of missing data over pairwise deletion,research paper, Temple University, Dept. of Psychology, Philadelphia, Pennsylvania 19122, USA

Baker L.A. and Fulker D.W., 1983, Incomplete covariance matrices and LISREL, Data Analyst, 1, 3-5

Beale E.M. and Little R.J.A., 1975, Missing values in multivariate analysis, Journal of the Royal Statistical Society, 37, 129-145

Bollen K.A., 1987, Structural equations with latent variables, John Wiley and Sons Inc., New York

Boomsma A., 1983, On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality, Unpublished doctoral dissertation, University of Groningen, Groningen, The Nederlands

Breiman L., Friedman J., Olshen R. and Stone C., 1984, Classification and regression trees. Pacific Grove: Wadsworth

Brown C.H., 1983, Asymptotic comparison of missing data procedures for estimating factor loadings, Psychometrika, 48, 269-291

Brown R.L., 1994, Efficacy of the indirect approach for estimating structural equation models with missing data: a comparison of five methods, Structural equation modelling, 4, 287-316

Chuanhai Liu, 1992a, Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data, Working Paper, Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

Chuanhai Liu, 1992b, Efficiently drawing the posterior mean and covariance for monotone normal ignorable missing data, Technical Report, Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

Dempster A.P., Laird N.M. and Rubin, D.B., 1977, Maximum likelihood estimation from incomplete data via the EM Algorithm, Journal of the Royal Statistical Society , Series B, 39, 1-38

Dijkstra T.K., 1981, Latent variables in linear stochastic models, Amsterdam, Sociometric

Research Foundation

Duncan O.D., Haller A.O. and A. Portes, 1968, Peer influences on aspirations : a reinterpretation., The American Journal of Sociology, 74, 119-137

Duncan S.C. and Duncan T.E., 1994, Modeling incomplete longitudinal substance use data using latent variable growth curve methodology, Multivariate Behavioral Research, 29, 313-338

Finkbeiner C., 1979, Estimation for the multiple factor model when the data are missing, Psychometrika, 44, 409-420

Fox J., 1984, Linear Statistical models and related methods, John Wiley and Sons, New-York

Glasser M., 1964, Linear regression analysis with missing observations among the independent variables, Journal of the American Statistical Association, 59, 834-844

Gleason T.C. and Staelin R.A., 1975, A proposal for handling missing data ,Psychometrika, 40, 229-252

Graham J.W., Hofer S.M. and MacKinnon D.P., 1996, Maximizing the usefulness of data obtained with planned missing value patterns : An application of maximum likelihood procedures, Multivariate Behavioral Research, 31, 197-218

Haitovsky Y., 1968, Missing data in regression analysis, Journal of the Royal Statistical Society, Series B, 30, 67-82

Hauspie R.C., Chrzastek-Spruch H., Verleye G., Kozlowska M.A. and Susanne S., 1996, Determinants of growth in body length from birth to 6 years of age: A longitudinal study of Lublin Children, American Journal of Human Biology, 8, 21-29

Heckman J. J., 1979, Sample selection bias as a specification error, Econometrika, 47, 153-161

Jöreskog K.G. and Dag Sörbom, 1989, LISREL 7 User's reference Guide, Scientific Software Inc., Mooresville

Jöreskog K.G. and Dag Sörbom, 1993, PRELIS version 2, a LISREL pre-processor, Scientific Software, Inc., Mooresville

Kaplan D., 1995, The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis, Journal of Educational and Behavioral statistics, 20, 69-82

Kendall M.G. and Stuart A., 1969, The advanced theory of statistics Vol 1, Butler and Tanner Ltd., London

Kim J. and Curry J., 1977, The treatment of missing data in multivariate analysis, Sociological Methods and Research, 6, 215-241

Kleijnen J.P.C., 1974, Statistical techniques in simulation, Marcel Dekker Inc., New York

Lee S., 1986, Estimation for structural equation models with missing data, Psychometrika, 51, 93-99

Little R.J.A. and Rubin D.B., 1987, Statistical analysis with missing data, John Wiley and Sons, New York

Little R.J.A. and Rubin D.B., 1989, The analysis of social sciences data with missing values, Sociological Methods and Research, 18, 292-326

Malhotra N.K., 1987, Analyzing market research data with incomplete information on the dependent variable, Journal of Marketing Research, 14, 74-84

Marini M., Olsen A.R. and Rubin D.B., 1987, Maximum-likelihood estimation in panel studies with missing data, in C.C. Clogg (ed.), Sociological Methodology, San Francisco: Jossey-Bass, 292-326

McArdle J.J., 1994, Structural factor analysis experiments with incomplete data, Multivariate Behavioral Research, 29, 409-454

Muthén B. and Jöreskog K.G., 1983, Selectivity problems in quasi-experimental studies, Evaluation Review, 7, 139-174

Muthén B., Kaplan D. and Hollis, M., 1987, On structural equation modelling with data that are not missing completely at random, Psychometrika, 52, 431-462

Neale M.C., 1994, Mx statistical modelling 2nd edition, Box 710 MCV, Richmond, VA 23298, Department of Psychiatry

Rovine M.J., 1994, Latent variable models and missing data, in Latent variables analysis: applications for developmental research, von Eye A. and Clogg C.C. eds., Sage Publications

Rubin D.B., 1976, Inference and missing data, Biometrika, 63, 581-592

Rubin D.B., 1987, Multiple imputation for nonresponse in surveys. New York : John Wiley and Sons Inc.

Rubin D.B. and Schafer J.L., 1987, Efficiently creating multiple imputations for incomplete multivariate data, proceedings of the statistical computing section of the American Statistical Association

Rudinger G., Andres J. and Rietz C., 1989, Structural equation models for studying intellectual development, in Problems and methods in longitudinal research, Magnusson D., Bergman L.R., Rudinger G. and Törestad B. (eds.), Cambridge University Press

Steinberg D. and Colla P., 1995, CART: Tree-structured Non-Parametric Data Analysis. San Diego, CA: Salford Systems

Tanner M.A. and Wong W.H., 1987, The calculation of posterior distributions by data augmentation, Journal of the American Statistical Association, 82, 528-550

Timm N.H., 1970, The estimation of variance-covariance and correlation matrices from incomplete data, Psychometrika, 35, 417-437

Verleye G., 1996, Missing at random data problems in attitude measurement using maximum likelihood structural equation modelling, unpublished Ph.D. dissertation, Vrije Universiteit Brussel, Centrum voor Statistiek en Wiskunde.

Werts C.E., Rock D.A. and Grandy J., 1979, Confirmatory factor analysis applications: missing data problems and comparison of path models between populations, Multivariate Behavioral Research, 14, 199-213