Running head: IMPUTATION FOR FACTOR ANALYSIS

A comparison of imputation methods for Bayesian factor analysis models

Edgar C. Merkle

Department of Psychology

Wichita State University

Wichita, KS 67260

316.978.3823

edgar.merkle@wichita.edu

Abstract

Imputation methods are popular for the handling of missing data in psychology. The methods generally consist of predicting missing data based on observed data, yielding a complete dataset that is amiable to standard statistical analyses. In the context of Bayesian factor analysis, this paper compares imputation under an unrestricted multivariate normal model (*Multiple Imputation*) to imputation under the statistical model of interest (*Data Augmentation*). The former method is popular in applied research, but the latter method is more straightforward from a Bayesian perspective. Simulations demonstrate that Data Augmentation yields less-biased parameter estimates for moderate sample sizes and high missingness proportions. Multiple Imputation, on the other hand, yields less-biased parameter estimates for large sample sizes with misspecified models. The incorporation of auxiliary variables in Data Augmentation is also addressed, and BUGS code is provided.

Keywords: Missing data, factor analysis, data augmentation, multiple imputation, BUGS

A comparison of imputation methods for Bayesian factor analysis models

Incomplete data are ubiquitous among most fields of research. Experimental participants fail to respond to every item, recording instruments fail to function, and research assistants fail to save files. Because many common statistical analyses are designed with complete data in mind, these missing data become a problem: researchers do not want to spend their resources on data that they cannot analyze. Furthermore, even if researchers can analyze the data, missingness may render the results misleading.

Incomplete data are especially problematic for factor analysis and structural equation models. These models generally require large sample sizes for adequate power (e.g., MacCallum, Widaman, Zhang, & Hong, 1999), so the need to use all collected data is increasingly important. In response to such needs, many general methods for "completing" incomplete data have been utilized. These methods include Mean Imputation and Hot-Deck Imputation. In Mean Imputation, missing data are replaced with the observed mean of the corresponding variable. In Hot-Deck Imputation, missing data are replaced with observed data on the corresponding variable. Sinharay, Stern, and Russell (2001) provide more information on these techniques.

While the above imputation methods are intuitive and relatively easy to implement, they generally yield biased parameter estimates (e.g., Schafer & Graham, 2002). A basic problem is that, in replacing missing data with observed data, the uncertainty associated with the missing data is ignored. One solution to this problem is Multiple Imputation (MI; Rubin, 1987; Schafer, 1997). MI generally consists of replacing ("imputing") missing data with multiple predictions that are based on observed data. We are left with multiple complete datasets, each of which may be analyzed via complete-data methods. Final results are obtained by averaging over results for each completed dataset.

Traditional MI methods involve first creating a small number (e.g., 3–10) of complete datasets, then analyzing each dataset via a Maximum Likelihood procedure. Many researchers (e.g., Meng, 1994; Schafer, 1997) have noted that traditional MI methods have a *separation* advantage; that is, they separate the handling of missing data from the issue of model estimation. Thus, a missing data expert could create the imputed datasets, and applied researchers could then use the datasets for many types of analyses.

As an alternative to MI, researchers can handle the missing data and model estimation steps simulateneously. For example, the Data Augmentation (DA) algorithm (Little & Rubin, 2002; Tanner & Wong, 1987) involves sequentially imputing missing data and sampling from a complete-data Bayesian model via Markov chain Monte Carlo (MCMC). Final results are obtained by summarizing sampled parameters from the posterior distribution. While DA cannot be used to separate imputation from model estimation, the separation advantage is lost on many situations where factor analysis is employed. In these situations, psychologists collect data primarily for scale development; the data are often not intended for general use across many analyses or for public consumption. Therefore, researchers are often required to handle both the imputation and analysis steps themselves.

Instead of distinguishing traditional MI from DA based on separation, the more relevant difference for factor analysis models is what Meng (1994) calls *congeniality*. Congeniality is generally defined as the correspondence between the imputation model and the data analysis model. In traditional MI for continuous data, the imputation model is usually based on a multivariate normal model with an unrestricted covariance matrix (i.e., a saturated multivariate normal model). The statistical model of interest may then be, for example, a factor analysis model or a regression model. In contrast, DA utilizes the factor analysis model for both imputation and data analysis. Meng examines the asymptotic properties (number of imputations $\rightarrow \infty$) of uncongenial imputation methods for

regression, and he recommends that the imputation model be more general than the analysis model. However, he does not study the issue for finite numbers of imputations, as are used in applied research. Furthermore, the models that he studies are generally less complex than factor analysis models.

The goal of this paper is to compare MI with DA in the context of factor analysis. In using MI methods, many researchers overlook congeniality issues and the potential negative impact they may have on statistical analyses. On the other hand, because the multivariate normal model used in MI is more general than factor analysis models, MI may be able to provide better parameter estimates when the factor analysis model is misspecified. In comparing MI methods with DA, it is possible to demonstrate the extent to which congeniality and generality trade off with one another and to make recommendations for applied researchers. I focus explicitly on MI and DA because they are both imputation methods that involve Bayesian estimation, with the main difference involving model congeniality. Full Information Maximum Likelihood (FIML; e.g., Wothke, 2000) is also popular and useful for the handling of missing data in factor analysis models, but it is not studied here. Assuming non-informative prior distributions, DA often yields similar parameter estimates to FIML, with the main difference being that the DA standard error estimates are not asymptotic.

I first briefly review the assumptions that are employed in the handling of missing data. Next, I describe MI and DA, along with the use of BUGS[1] to estimate factor analysis models. Next, DA and MI are compared with one another in a series of simulations containing various missingness proportions, sample sizes, and model (mis)specifications. Finally, I show how auxiliary variables related to missingness can be incorporated into estimation via DA. This utilizes a model developed by Graham (2003) for FIML estimation; to my knowledge, incorporation of auxiliary variables in DA has not been previously addressed.

Missing Data Assumptions

Rubin (1976) initially developed conditions under which we may ignore the missing data process, which eases model estimation. Little and Rubin (1987) extended the conditions and introduced terminology that is now in common use. A brief review of these conditions appears below; other descriptions of the missingness assumptions have been presented in, e.g., Collins, Schafer, and Kam (2001), Schafer (1997), and Sinharay, Stern, and Russell (2001). The notation used here is consistent with that of Schafer (1997).

Missingness Conditions

Let $\boldsymbol{Y}$ be a data matrix of n individuals measured on p variables. Label the missing part of $\boldsymbol{Y}$ as $\boldsymbol{Y_{mis}}$ and the observed part of $\boldsymbol{Y}$ as $\boldsymbol{Y_{obs}}$. Intuitively, the data in $\boldsymbol{Y}$ are *missing at random* (MAR) if the probability that an observation is missing depends on $\boldsymbol{Y_{obs}}$ but not on $\boldsymbol{Y_{mis}}$.

Formally, let $\boldsymbol{M}$ be an $n \times p$ matrix whose elements equal 1 if the corresponding element in $\boldsymbol{Y}$ is observed and 0 otherwise. Let $\boldsymbol{\xi}$ be a vector of parameters on which the missing data mechanism depends. The data in $\boldsymbol{Y}$ are MAR if:

$$P(\boldsymbol{M} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}, \boldsymbol{\xi}) = P(\boldsymbol{M} \mid \boldsymbol{Y_{obs}}, \boldsymbol{\xi}) \tag{1}$$

A special case of the "missing at random" condition is *missing completely at random* (MCAR). Under the MCAR condition, the probability of missingness depends on neither the values of the missing data nor the values of the observed data. Formally, data are MCAR if:

$$P(\boldsymbol{M} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}, \boldsymbol{\xi}) = P(\boldsymbol{M} \mid \boldsymbol{\xi}) \tag{2}$$

Finally, data are *missing not at random* (MNAR) if the probability of missingness does depend on the missing values. There are no general techniques for dealing with

MNAR data because there are an infinite number of possible MNAR mechanisms. Specific models have been proposed on a case-by-case basis with varying degrees of success (Little & Rubin, 2002). Denoting parameters from the model of interest by $\boldsymbol{\theta}$, MNAR data essentially require one to work with the joint distribution $f(\boldsymbol{Y_{obs}}, \boldsymbol{M} \mid \boldsymbol{\theta}, \boldsymbol{\xi})$ instead of the more convenient $f(\boldsymbol{Y_{obs}} \mid \boldsymbol{\theta})$, which can be used with MAR data.

Use of MAR procedures with MNAR data may yield biased parameter estimates or inflated standard error estimates. In practice, however, the MAR assumption is often a good one to make. If one does not make the MAR assumption, one must know enough about the missingness mechanism to form a model of it. If the proposed MNAR model is different from the true missingness mechanism, then the results can be worse than results obtained by assuming MAR data (Little & Rubin, 2002). The missing data methods in this paper all assume that the data are MAR. These methods are further described below.

Multiple Imputation

As described in the introduction, the general idea of Multiple Imputation (Schafer, 1997; Sinharay et al., 2001) is straightforward. Starting with incomplete data, m complete data sets (*imputations*) are created, each with different predicted values of $\boldsymbol{Y_{mis}}$. Each of these m data sets is then analyzed via standard, complete-data procedures. Finally, results arising from the m complete-data analyses are combined to yield a single estimate and standard error for each parameter of interest.

To generate values for the missing data in $\boldsymbol{Y_{mis}}$, we ultimately wish to sample from the distribution $P(\boldsymbol{Y_{mis}} \mid \boldsymbol{Y_{obs}})$. In order to do this, a data model for $\boldsymbol{Y} = (\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}})$ must be assumed. Use of the data model, along with a prior distribution for the parameters of the data model, allows for sampling from the desired distribution. While several types of data models have been used previously, the most common for continuous data is the multivariate normal. Use of the multivariate normal model is convenient and

often yields accurate predictions even when the data in $\boldsymbol{Y}$ are not multivariate normal (Schafer, 1997). Depending on the type of data, other possible data models include loglinear models and multinomial models. Noninformative prior distributions are commonly used for the parameters of the data model but are not required.

MI actually uses DA (described in detail below) to sample from the distribution $P(\boldsymbol{Y_{mis}} \mid \boldsymbol{Y_{obs}})$. In MI, however, DA is used in concert with a saturated multivariate normal model to impute data. MI thus results in uncongenial models: one model is used to impute the data, while a second, different model is used to analyze the data. This might lead one to question the selection of an imputation model (e.g., "If you believe that a factor analysis model best describes the data, why use a saturated multivariate normal model to impute the data?"), and it could plausibly lead to imprecise parameter estimates in practice. This issue is contrasted with the general DA algorithm presented below.

Data Augmentation

The DA algorithm (Tanner & Wong, 1987; Little & Rubin, 2002) is a general Bayesian method, based on Gibbs sampling[2] (Casella & George, 1992; Gelfand & Smith, 1990), that is designed to make posterior simulation easier. The algorithm's generality means that it can also ease model estimation in situations with complete data, but I will focus here on its application to missing data. For a given incomplete data set, we generally wish to sample from the observed-data posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{Y_{obs}})$. This can be accomplished via DA in the following two steps. Given $\boldsymbol{\theta}^{(t)}$, a sampled value of $\boldsymbol{\theta}$ at iteration t, the algorithm:

1. Draws $\boldsymbol{Y_{mis}}^{(t+1)}$ from $p(\boldsymbol{Y_{mis}} \mid \boldsymbol{Y_{obs}}, \boldsymbol{\theta}^{(t)})$;
2. Draws $\boldsymbol{\theta}^{(t+1)}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}^{(t+1)})$.

Step 1 above is sometimes called the imputation step, and Step 2 above is sometimes called the posterior step. Values of $\boldsymbol{\theta}$ sampled in this manner converge to the

observed-data posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{Y_{obs}})$. The utility of this algorithm comes from the fact that the distributions listed in the two steps above (a regression-like distribution and a complete-data posterior distribution, respectively) are usually of a simpler form than the observed-data posterior distribution.

The DA method is congenial; that is, the distribution from which $\boldsymbol{Y_{mis}}$ is sampled is related to the distribution from which $\boldsymbol{\theta}$ is sampled. For example, in a factor-analytic context, we would sample $\boldsymbol{Y_{mis}}$ based on the factor analysis model and previously-sampled model parameters. This is different from multiple imputation, where one model is used to sample missing data points and a different model is typically used to analyze the data.

BUGS Implementation

It is relatively straightforward to develop DA methods for factor analysis based on complete-data Bayesian methods. Factor analysis involves a multivariate normal distribution, and it is easy to obtain the conditional distribution $p(\boldsymbol{Y_{mis}} \mid \boldsymbol{Y_{obs}}, \boldsymbol{\theta}^{(t)})$ based on properties of the multivariate normal distribution. An imputation step based on this conditional distribution can be inserted into a method for estimating complete-data Bayesian models, such as that of Scheines et al. (1999). Alternatively, as discussed below, it is straightforward to implement DA for factor analysis using BUGS. This is likely to be preferable for applied researchers, so it is used in the simulations in this paper.

BUGS is a general program for estimating Bayesian models via Markov chain Monte Carlo. Its popularity lies in the fact that the model specification is very flexible, allowing researchers to estimate a wide variety of Bayesian models. Complete-data factor analysis models have been previously estimated in BUGS (e.g., Lee, 2007; Zhang, McArdle, Wang, & Hamagami, 2008), and it is straightforward to estimate these models with missing data. BUGS treats missing data as parameters to be estimated, resulting in a DA algorithm. Missing observations are specified with "NA" values in the data file, with this feature

being used in the simulations below. Examples of BUGS code for factor analysis appears in the appendix.

Simulations

To examine the impact of congeniality in missing data methods for factor analysis, I now describe a series of simulation studies comparing MI with DA. The simulations are designed to examine the impact of congeniality across varying missingness proportions, sample sizes, and data-generating models.

Data

The simulations are based on classic data from Holzinger and Swineford (1939), which consist of 73 female high school students measured on 6 aptitude tests. Three of the tests are verbal in nature, and three of the tests are spatial in nature. Thus, the confirmatory factor analysis model pictured in Figure 1 provides a good description of the data. The model is also known to statistically fit the data well, as judged by χ^2 statistics, RMSEA statistics, and examination of residuals. Specific data deletion techniques are used to examine the congeniality issue.

In the simulations below, data are generated from a known mechanism that is designed to mimic the Holzinger & Swineford data; population parameters were set equal to the values in Figure 1. This is similar to a procedure used by Jennrich (2006) to study the signed permutation problem in exploratory factor analysis. The procedure is advantageous over the use of generic data because the researcher can control the specific parameter values used for data generation, while the reader maintains some level of familiarity with the data. The generation of many datasets ensures that there are no peculiarities about a particular dataset that have an adverse impact on results.

Simulation 1: Well-Specified Model

As outlined above, both DA and MI assume data that are MAR (of which MCAR data are a special case). Thus, a "clean" (though unrealistic) comparison of the methods involves simulations with a well-specified model and MCAR data.

Procedure

Five-hundred incomplete datasets were generated from the confirmatory factor analysis model in Figure 1. The MCAR deletion mechanism involved random deletion of five manifest variables (MVs) to yield specific missingness proportions. The sixth MV (VISPERC in Figure 1) was always observed so that all cases would have one observation in common; "fragmented" missing data can cause problems for multiple imputation methods (Schafer, 1997). The five MVs were deleted to yield 10%, 25%, and 40% expected overall missingness proportions; that is, data were randomly deleted with probabilities expected to lead to these missingness proportions. The actual missingness proportions for each dataset vary around the expected proportions.

After obtaining the 500 incomplete datasets, a Bayesian confirmatory factor analysis model was fit to each dataset via both DA and MI. For each incomplete dataset estimated with the DA algorithm, three chains were sampled for 1,000 iterations each after burn-in, with burn-in time varying depending on how quickly the chains converged.[3] For each incomplete dataset estimated with MI, ten imputed datasets were created. The factor analysis model was fit to each dataset via BUGS, and results were combined using Rubin's (1987) rules. This procedure was carried out for datasets of size 100 and 500, with the previously-noted expected missingness proportions of 10%, 25%, and 40%.

The simulations offer a direct examination of the congeniality issue. Because DA is congenial, I expected it to generally yield parameter estimates that were closer to the true parameter values. To examine this expectation, I define some measures that make use of

the fact that DA and MI were employed on the same samples of data. First, we can examine the proportion of samples for which the MI estimates are further from the true parameter values (i.e., the θ_i) than the DA estimates. If MI and DA perform equivalently, then this proportion should be .5. If DA outperforms MI, then the proportion should be larger than .5. Second, we can compare the distances from the MI parameter estimates to the true parameter values with the distances from the DA parameter estimates to the true parameter values. Assuming we have generated 500 datasets, this leads to a relative bias measure for each θ_i that can be written as:

$$RB(\theta_i) = (1/500) \sum_{j=1}^{500} |\hat{\theta}_{ij}^{MI} - \theta_i| - |\hat{\theta}_{ij}^{DA} - \theta_i|, \tag{3}$$

where j indexes generated datasets. When $RB(\theta_i)$ is greater than 0, MI estimates tend to be further from the true parameter value than do DA estimates. The magnitude of the above measure is difficult to interpret, however, because each parameter has its own magnitude and variability. Thus, I define a standardized relative bias measure as:

$$SRB(\theta_i) = RB(\theta_i)/\widehat{SE(\theta_i)}, \tag{4}$$

where $\widehat{SE(\theta_i)}$ is a standard error estimate of θ_i. For the results in this paper, I used the average posterior standard deviation of $\hat{\theta}_i$ resulting from DA. The standardized relative bias can be roughly interpreted as the number of standard errors that MI estimates stray from the true values, as compared to DA estimates. This is similar to the standardized bias measure defined by Collins et al. (2001), except it makes use of the fact that DA and MI were used in concert with the same datasets.

Results

SRB measures and proportions of samples for which MI estimates are further from the true θ are listed in Table 1 for each of the three levels of missingness. Within the table, proportions greater than .5 and positive standardized biases reflect instances where DA estimates tended to be closer to the truth than MI estimates.

At $N = 100$, DA tends to yield parameter estimates closer to the true parameters than does MI. This advantage of DA over MI is small at 10% missingness and generally increases with missingness proportion. At $N = 500$, the DA advantage generally remains but is less pronounced; at 10% missingness, some MI parameter estimates are slightly less biased than DA parameter estimates. Across the two sample sizes, the DA advantage tends to be larger for unique variance estimates than for factor loadings.

Discussion

In comparing the performance of DA and MI in estimating factor analysis models with MCAR data, results showed that DA parameter estimates tend to be closer to the true parameter values. This finding was most evident for unique variances at smaller sample sizes and with larger missinginess proportions. Similar simulations were conducted with smaller numbers of imputations in MI, and results (not shown) were similar. In particular, the ability of DA to yield better unique variance estimates was inflated when only three imputations were used in MI. This highlights another difference between MI and DA: MI uses a small number of imputed datasets to account for uncertainty in the missing data, while DA imputes missing data at every iteration of the MCMC algorithm, resulting in thousands of imputed datasets. Thus, part of DA's advantage may stem from the fact that it better accounts for the uncertainty in the missing data. With this in mind, a more accurate comparison of DA and MI would involve the use of thousands of imputed datasets for MI, with the factor analysis model fit to each imputed dataset and results

then combined across the thousands of datasets. MI is not used this way in practice, however, so such a comparison is primarily of theoretical interest. Additionally, a simulation involving MI with thousands of imputed datasets and Bayesian estimation is very computationally intensive.

From a practical point of view, the differences between DA and MI were not large in these simulations: DA standardized biases were usually less than 10% better than MI standardized biases (in standard errors away from true parameter values), and MI was nearly equivalent to DA at $N = 500$ and 10% missingness. Thus, these simulations do not yield strong evidence for consistently preferring DA over MI. In fact, the saturated multivariate normal distribution used in MI may be preferable if the factor analysis model is misspecified. This scenario is examined in Simulation 2.

Simulation 2: Misspecified Model

The results from Simulation 1 showed that DA has a notable advantage over MI that is highlighted at small sample sizes and high missingness proportions. It may be the case, however, that MI is more robust to model misspecification. That is, if the estimated model does not match the true model, then DA's handling of missing data within the estimated model may be faulty. It may be more desirable to impute data via a saturated multivariate normal model, as is done in MI. This is the focus of Simulation 2.

Procedure

The procedure for Simulation 2 was similar to the procedure for Simulation 1: 500 incomplete datasets of sample sizes 100 and 500 were generated from a known model, and the same confirmatory factor analysis model was estimated by both DA and MI with Bayesian estimation. Data were also deleted in the same way as they were in Simulation

1. The major difference in Simulation 2 was that the true model (i.e., the data generation model) did not match the estimated model. In particular, to generate data, three new paths from latent variables to manifest variables (i.e., nonzero factor loadings) were added to the original model. Parameter values for these new paths were chosen to match the analogous parameter values from the other latent variable; the new paths (with parameter values) were: "Verbal" to "Lozenges" (5.96), "Spatial" to "Paragraph" (3.24), and "Spatial" to "Sentence" (4.32; see Figure 1 for MV names). The additional paths make the estimated model more restrictive than the true model. The saturated multivariate normal model used in MI is less restrictive than the true model, however, giving MI a potential advantage here.

Results

SRB measures and proportions of samples for which MI estimates are further from the true θ are listed in Table 2 for each of the three levels of missingness. Within the table, proportions greater than .5 and positive standardized biases reflect instances where DA estimates tend to be closer to the truth than MI estimates.

Examining Table 2, DA again offers a slight advantage at the smaller sample size ($N = 100$). The advantage is most apparent for unique variances and the factor correlation at high rates of missingness. At $N = 500$, MI provides better estimates of some parameters than does DA. The parameters for which MI provided better estimates include $\{\lambda_{31}, \lambda_{42}, \lambda_{52}, \psi_{33}, \psi_{44}, \psi_{55}\}$, which are also the parameters most impacted by the model misspecification (because they impact the manifest variables with added paths). DA provides better estimates of the factor correlation at $N = 500$, just as it does at $N = 100$.

Discussion

Simulation 2 compared DA and MI estimates in a situation when the factor analysis model was misspecified. While results were similar for the two methods at the smaller sample size ($N = 100$), MI generally yielded better estimates of the parameters most impacted by the model misspecification at $N = 500$. These results stem from the fact that, for MI, the imputation model was more general than the true model. In contrast, the imputation model for DA (the misspecified factor analysis model) was more restrictive than the true model. However, the smaller sample size and larger missingness proportions removed MI's "generality" advantage.

To summarize, if one expects the model to be misspecified and has a large sample size, MI may be preferable. However, this situation raises the question of why one is reporting results of a model that they expect to be misspecified. If one has a small sample size or believes that his or her model offers a reasonable description of the data, then DA appears to be preferable.

Including Auxiliary Variables

MI has an additional advantage that is not addressed in the above simulations: auxiliary variables can be easily included in the imputation step. If these auxiliary variables are related to the causes of missingness, then the MAR assumption is more tenable and parameter estimates may be less biased (e.g., Collins et al., 2001). As a result, model parameters estimated via MI with auxiliary variables may be less biased than those estimated via DA in the absence of auxiliary variables.

While the ability to include auxiliary variables has traditionally been restricted to

MI, Graham (2003) showed how auxiliary variables may be included in structural equation models estimated via FIML, so that their impact on other model parameters is minimal. While he focused on regression models in his paper, his *extra DV model* may be extended to the factor-analytic situations considered in this paper. In the context of factor analysis, the extra DV model requires extra paths from each latent variable to the auxiliary variable(s). For each auxiliary variable, this adds $(q + 2)$ parameters to the model, where q is the number of factors (a path from each factor to the auxiliary variable(s), along with a mean and unique variance for the auxiliary variable). Graham showed that this model can yield less-biased parameter estimates when the auxiliary variable is related to missingness, although it does have an impact on some measures of model fit.[4]

Code to estimate the extra DV model in BUGS is provided in the appendix. As a brief illustration of the model's performance, a sample of size 500 was generated from a model similar to the one in Figure 1, with the only difference being that the factors were not correlated. Observations on MVs 4–6 (PARAGRAPH, SENTENCE, WORDMEAN) were deleted based on values of PARAGRAPH in two steps: first, if the value of PARAGRAPH was less than its third quartile, then the three MVs were jointly accepted as candidates for deletion with probability .95. Each candidate observation was then individually deleted with probability .975. This results in MVs 4–6 being missing together most of the time, with a single one of these variables being occasionally observed (similar to the deletion scheme in Graham 2003). An always-observed auxiliary variable was also generated, which was correlated .95 with the factor giving rise to MVs 4–6 (VERBAL).

The above deletion scheme weakens the correlations between manifest variables 4-6, which in turn biases downward the factor loadings corresponding to these MVs.

Incorporating the auxiliary variable should make up for this bias, however, because the auxiliary variable is highly related to the missing data. Three estimation methods were used to fit the factor analysis model to a single dataset of size 500: the extra DV model was estimated via DA, the standard factor analysis model was estimated via MI with an auxiliary variable, and the standard factor analysis model was estimated via DA with no auxiliary variable. Results are presented in Table 3, with the true parameter values also listed for comparison. A main focus of comparison lies in the factor loadings corresponding to the second latent variable ($\lambda_{42}, \lambda_{52}, \lambda_{62}$): the models including the auxiliary variable yield estimates somewhat closer to the true values. Further, examining the unique variances corresponding to MVs 4-6, the extra DV model results in better estimates than does MI with the auxiliary variable. There are also differences in the standard errors associated with unique variances; specifically, MI's standard errors resemble those of DA with no auxiliary variable. These results are generally of a small magnitude, and full simulations would be needed to describe the exact impact of auxiliary variables here. However, the example here shows that: (1) auxiliary variables can be incorporated into DA; and (2) such a model can yield better parameter estimates than the analogous model that utilizes MI with the auxiliary variable.

General Discussion

Model congeniality is an often-overlooked consideration underlying imputation methods for incomplete data. Traditional MI methods result in uncongenial models: one statistical model is used to impute the data, and a second statistical model is used to analyze the data. Alternatively, the DA algorithm can simultaneously handle missing data

and estimate a statistical model of interest. The goal of this paper was to examine the extent to which congeniality plays a role in accommodating missing data within factor analysis models.

In this paper, I first described the general use of DA for handling missing data. MI and DA were then applied to the estimation of Bayesian, confirmatory factor analysis models. Examining the methods' performances across incomplete datasets of varying sizes and missingness proportions, I found DA to be advantageous at a smaller sample size ($N = 100$) and higher missingness proportion. This advantage was most pronounced for unique variance parameters and factor correlations. MI and DA were nearly equivalent at the larger sample size of 500. Next, I examined the methods' performance when the factor analysis model was misspecified (specifically, when the estimated model was more restrictive than the true model). DA still showed a mild advantage at the smaller sample size of $N = 100$, while, at $N = 500$, MI showed an advantage for parameters most impacted by the misspecification.

In general, lack of congeniality had the largest impact on unique variances. When the uncongenial models involve a saturated multivariate normal for imputation and factor analysis for estimation, it is likely that the relationships between variables are preserved: this is because the factor analysis model is nested within the saturated multivariate normal model. On the other hand, the variability among manifest variables may not be preserved. "Outlying" imputations may be taken from the saturated multivariate normal model that contain more variability than is expected under a factor analysis model. When the number of imputations is relatively small, this outlying imputation will have influence the variance estimates within the model. Thus, the unique variance parameters are less

precise while the factor loading parameters (representing relationships between variables) are relatively accurate. These findings bear some similarity to those of Kim and Fuller (2004), who found that a specific form of Hot-Deck Imputation (*Fractional Hot-Deck Imputation*) far outperformed MI in estimating variances. Their statistical models were simpler than the factor analysis models in the current paper, and congeniality was not a consideration in their comparison of the missing data methods. Sinharay et al. (2001, p. 328) also discuss a situation where MI provides biased estimates of a variance parameter.

Practical Implications

The results in this paper provide evidence that DA is to be preferred over MI at smaller sample sizes and at higher missingness proportions. In contrast, MI may be preferred when one expects model misspecification and has a large sample size. The advantages that one gains from either method may not be large, and, as always, care must be taken in translating simulation results to novel research situations.

While use of DA in the social sciences is not as popular as other model estimation procedures (though see Aitkin & Aitkin, 2005; Johnson & Junker, 2003; Lanza, Collins, Schafer, & Flaherty, 2005), it is straightforward to implement for many models via BUGS (see appendix). If one does not elect to use BUGS, it is still straightforward to implement DA for many types of linear models. The most difficult part of the general implementation may be the derivation of a model-based regression distribution: if the model of interest does not have nice mathematical properties, it will be more difficult to obtain the distribution of missing data conditioned on observed data. If the form of the distribution cannot be obtained analytically, however, it may be possible to at least sample from the

regression distribution via MCMC.

Summary

In summary, model uncongeniality has an adverse impact on factor analysis parameter estimates. The congeniality issue is one that researchers often overlook, and they should be aware of it in choosing a missing data method. To this end, the Data Augmentation algorithm is a general-purpose imputation method that utilizes congenial models. This algorithm should receive more attention as it is straightforwardly implemented in BUGS, removing much of the programming necessary for implementation.

References

Aitkin, M., & Aitkin, I. (2005). Bayesian inference for factor scores. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (p. 207-222). Mahwah, NJ: Lawrence Erlbaum Associates.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167-174.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330-351.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–511.

Genz, A., Bretz, F., & Hothorn, T. (2006). *mvtnorm: Multivariate normal and T distribution.* (R package version 0.7-5)

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100.

Holzinger, K. J., & Swineford, F. A. (1939). *A study of factor analysis: The stability of a bi-factor solution* (No. 48). Chicago: University of Chicago Press.

Jennrich, R. I. (2006). What do factor analysis loadings and their standard errors

estimate: The signed permutation problem. Presented at the 71st annual meeting of the Psychometric Society, Montreal, Quebec.

Johnson, M. S., & Junker, B. W. (2003). Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding response models. *Journal of Educational and Behavioral Statistics*, *28*, 195-230.

Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, *91*, 559-578.

Lanza, S. T., Collins, L. M., Schafer, J. L., & Flaherty, B. P. (2005). Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods*, *10*, 84-100.

Lee, S.-Y. (2007). *Structural equation modelling: A Bayesian approach.* Chichester: Wiley.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS– A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84-99.

Martin, A. D., & Quinn, K. M. (2006). *MCMCpack: Markov chain Monte Carlo (MCMC) package.* (R package version 0.7-2)

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*, 538-558.

Novo, A. A., & Schafer, J. L. (2002). *norm: Analysis of multivariate normal datasets with missing values.* (R package version 1.0-9)

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). *coda: Output analysis and diagnostics for MCMC.* (R package version 0.10-7)

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37-52.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*, 317-329.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software*, *12*, 1–16.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*, *6*, 12–17.

van Buuren, S., & Groothuis-Oudshoorn, K. (forthcoming). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zhang, Z., McArdle, J. J., Wang, L., & Hamagami, F. (2008). A SAS interface for Bayesian analysis with WinBUGS. *Structural Equation Modeling*, *15*, 705–728.

Appendix

BUGS Code for Confirmatory Factor Analysis

This appendix provides code for the confirmatory factor analysis models estimated in this paper. As described in the paper, missing data are automatically handled in BUGS by specifying "NA" values in the data file. The code for the first model (standard factor analysis model) is based on similar code described in Lee (2007); the code for the second model (extra DV model) expands on the first code. To avoid issues with parameter identifiability, I make use of the `I(0,)` syntax for factor loadings. This ensures that the sampled parameters are positive, which is reasonable for the data used here. The mildly-informative prior distributions on the mean parameters aid in convergence of parameter chains.

```
model{
  for (i in 1:N){
    for (t in 1:3){
      y[i,t] ~ dnorm(condmn[i,t], invsig2[t])
      condmn[i,t] <- mu[t] + fload[t]*fscore[i,1]
    }
    for (t in 4:6){
      y[i,t] ~ dnorm(condmn[i,t], invsig2[t])
      condmn[i,t] <- mu[t] + fload[t]*fscore[i,2]
    }
    fscore[i,1:2] ~ dmnorm(mn.fs[],siginv.fs[,])
  }

  mn.fs[1] <- 0
  mn.fs[2] <- 0
  sig.fs[1,1] <- 1
  sig.fs[2,2] <- 1
  sig.fs[1,2] <- phi
  sig.fs[2,1] <- phi

  # Prior distributions:
```

```
  phi ~ dunif(-1,1)
  for (t in 1:6){
    fload[t] ~ dnorm(0, 1.0E-3)I(0,)

    invsig2[t] <- 1/psi[t]
    psi[t] ~ dunif(0,400)

    mu[t] ~ dnorm(20,.05)
  }
}
```

The extra DV model is then a simple modification of the above code:

```
model{
  for (i in 1:N){
    for (t in 1:3){
      y[i,t] ~ dnorm(condmn[i,t], invsig2[t])
      condmn[i,t] <- mu[t] + fload[t]*fscore[i,1]
    }
    for (t in 4:6){
      y[i,t] ~ dnorm(condmn[i,t], invsig2[t])
      condmn[i,t] <- mu[t] + fload[t]*fscore[i,2]
    }
    y[i,7] ~ dnorm(condmn[i,7], invsig[7])
    condmn[i,7] <- mu[7] + fload[7]*fscore[i,1] + fload[8]*fscore[i,2]

    fscore[i,1:2] ~ dmnorm(mn.fs[],siginv.fs[,])
  }

  mn.fs[1] <- 0
  mn.fs[2] <- 0
  sig.fs[1,1] <- 1
  sig.fs[2,2] <- 1
  sig.fs[1,2] <- phi
  sig.fs[2,1] <- phi

  # Prior distributions:
  phi ~ dunif(-1,1)
  for (t in 1:6){
    fload[t] ~ dnorm(0, 1.0E-3)I(0,)
  }
  fload[7] ~ dnorm(0, 1.0E-3)I(0,)
  fload[8] ~ dnorm(0, 1.0E-3)I(0,)

  for (t in 1:7){
```

```
    invsig2[t] <- 1/psi[t]
    psi[t] ~ dunif(0,400)

    mu[t] ~ dnorm(20,.05)
  }
}
```

Author Note

This research was partially supported by National Science Foundation Grant SES-0214574. The author thanks Robert Cudeck, Mario Peruggia, Trisha Van Zandt, two anonymous referees, and the editor for helpful comments. Any remaining errors are solely due to the author.

Portions of this work are based on the author's doctoral dissertation at The Ohio State University, and portions of this work were presented at the 71st meeting of the Psychometric Society, Montreal, Quebec. The work benefited from the R packages of coda (Plummer, Best, Cowles, & Vines, 2006), MCMCpack (Martin & Quinn, 2006), MICE (van Buuren & Groothuis-Oudshoorn, forthcoming), mvtnorm (Genz, Bretz, & Hothorn, 2006), norm (Novo & Schafer, 2002), and R2WinBUGS (Sturtz, Ligges, & Gelman, 2005).

Footnotes

[1]For the purposes of this paper, I make no distinction between OpenBUGS (Thomas, O'Hara, Ligges, & Sturtz, 2006) and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). The term "BUGS" is generally used to describe this family of software.

[2]Gibbs sampling is a method for drawing samples from a distribution of interest. Many posterior distributions are intractable, meaning that we cannot easily obtain point estimates of model parameters. Instead, methods such as Gibbs sampling allow us to draw many samples from the posterior distribution. These samples can then be summarized to obtain, e.g., point estimates and associated standard errors.

[3]Convergence of all chains throughout the paper was monitored using the Gelman-Rubin statistic (Gelman & Rubin, 1992), with chains being judged to have converged if all values were less than 1.2.

[4]Graham presented a second model (the *saturated correlates model*) that has no impact on model fit measures. This model is also potentially applicable to factor analysis, but it is not easily estimated via BUGS or other MCMC methods for factor analysis. In particular, the model's error covariance is not diagonal, forcing the use of multivariate instead of univariate normal distributions.

Table 1
Proportion of datasets for which MI yields parameter estimates with larger bias and standardized relative biases for MCAR datasets of $N = 100$ and $N = 500$. DA = Data Augmentation, MI = Multiple Imputation with Bayesian analysis, Prop = P(MI estimates exhibit more bias), SRB = Standardized Relative Bias.

$N = 100$	% Missing	λ_{11}	λ_{21}	λ_{31}	λ_{42}	λ_{52}	λ_{62}	ϕ_{12}
Prop	10	.52	.57	.52	.53	.53	.52	.47
	25	.57	.55	.55	.59	.57	.53	.56
	40	.61	.56	.57	.63	.58	.56	.59
SRB	10	.01	.02	.01	.01	.01	.01	.00
	25	.02	.03	.03	.04	.03	.03	.04
	40	.07	.07	.07	.12	.07	.07	.06

	% Missing	ψ_{11}	ψ_{22}	ψ_{33}	ψ_{44}	ψ_{55}	ψ_{66}
Prop	10	.53	.56	.52	.58	.54	.53
	25	.57	.59	.55	.63	.59	.65
	40	.61	.64	.66	.70	.66	.64
SRB	10	.01	.02	.02	.03	.03	.02
	25	.03	.05	.03	.11	.06	.08
	40	.09	.13	.12	.21	.15	.16

$N = 500$	% Missing	λ_{11}	λ_{21}	λ_{31}	λ_{42}	λ_{52}	λ_{62}	ϕ_{12}
Prop	10	.53	.49	.49	.55	.54	.52	.49
	25	.51	.57	.52	.53	.52	.51	.55
	40	.53	.55	.55	.52	.54	.53	.55
SRB	10	.01	-.00	.00	.01	.01	.01	.00
	25	.01	.03	.02	.02	.02	-.00	.02
	40	.04	.03	.05	.02	.02	.03	.05

	% Missing	ψ_{11}	ψ_{22}	ψ_{33}	ψ_{44}	ψ_{55}	ψ_{66}
Prop	10	.49	.52	.52	.53	.47	.51
	25	.53	.58	.56	.56	.53	.54
	40	.61	.54	.52	.56	.53	.56
SRB	10	.00	.00	.01	.01	-.01	.01
	25	.00	.04	.03	.03	.03	.03
	40	.07	.04	.04	.05	.03	.06

Table 2
Proportion of datasets for which MI yields parameter estimates with larger bias and standardized relative biases for MCAR datasets of $N = 100$ and $N = 500$, with the estimated model being misspecified. DA = Data Augmentation, MI = Multiple Imputation with Bayesian analysis, Prop = P(MI estimates exhibit more bias), SRB = Standardized Relative Bias.

$N = 100$	% Missing	λ_{11}	λ_{21}	λ_{31}	λ_{42}	λ_{52}	λ_{62}	ϕ_{12}
Prop	10	.55	.52	.48	.49	.46	.50	.68
	25	.62	.50	.51	.52	.46	.43	.75
	40	.59	.55	.62	.59	.50	.50	.79
SRB	10	.01	.01	-.00	-.01	-.01	-.01	.09
	25	.03	.00	.00	.01	-.02	-.03	.21
	40	.05	.05	.09	.04	.00	-.02	.30

	% Missing	ψ_{11}	ψ_{22}	ψ_{33}	ψ_{44}	ψ_{55}	ψ_{66}
Prop	10	.53	.59	.46	.54	.53	.55
	25	.62	.67	.51	.67	.64	.56
	40	.69	.72	.56	.69	.69	.63
SRB	10	.01	.04	-.01	.03	.02	.02
	25	.03	.12	.01	.12	.13	.04
	40	.09	.23	.07	.26	.23	.10

$N = 500$	% Missing	λ_{11}	λ_{21}	λ_{31}	λ_{42}	λ_{52}	λ_{62}	ϕ_{12}
Prop	10	.54	.54	.44	.46	.47	.38	.68
	25	.50	.56	.40	.51	.54	.30	.79
	40	.62	.64	.38	.48	.51	.25	.83
SRB	10	.01	.02	-.02	-.01	-.00	-.04	.14
	25	.00	.03	-.05	.00	.01	-.12	.30
	40	.04	.12	-.07	-.00	.01	-.20	.41

	% Missing	ψ_{11}	ψ_{22}	ψ_{33}	ψ_{44}	ψ_{55}	ψ_{66}
Prop	10	.53	.63	.37	.46	.58	.39
	25	.50	.70	.30	.49	.54	.23
	40	.59	.77	.31	.55	.55	.19
SRB	10	.01	.04	-.07	-.01	.03	-.04
	25	.00	.10	-.16	-.01	.03	-.14
	40	.04	.22	-.19	.06	.07	-.24

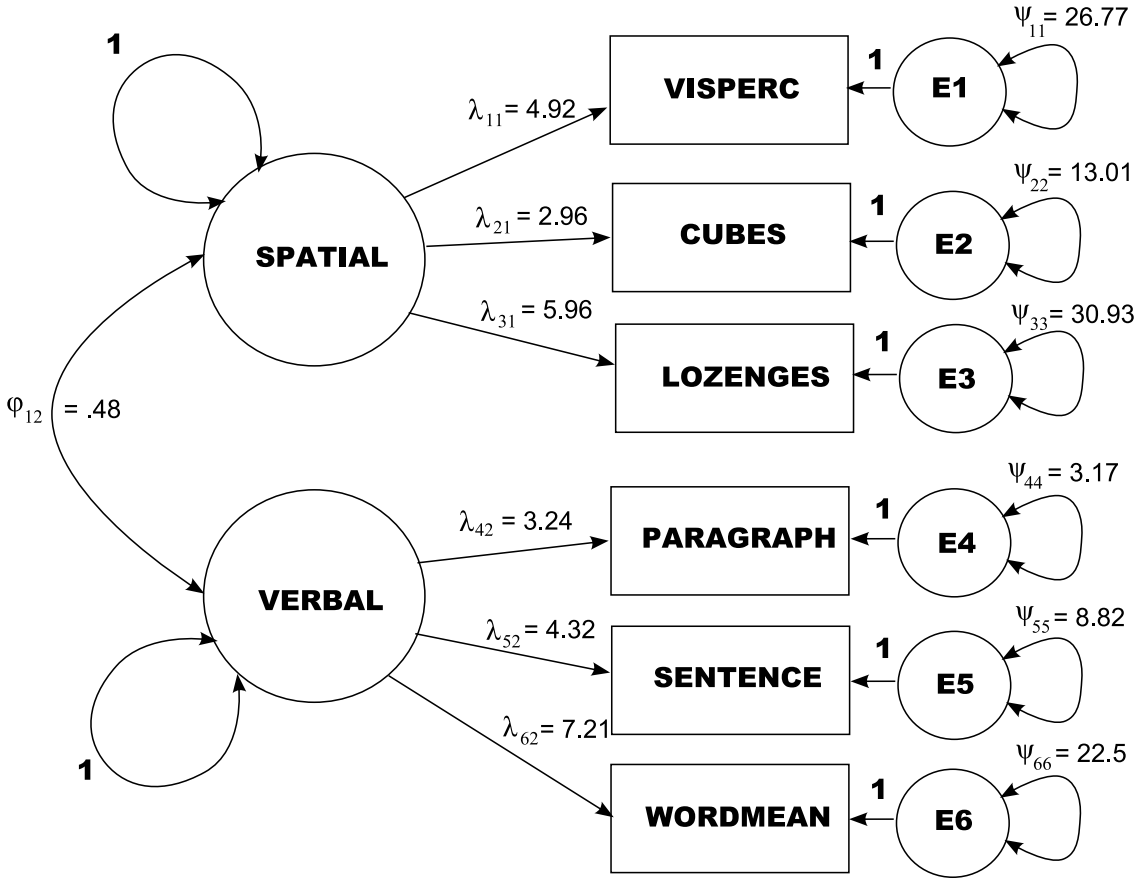
Table 3
Parameter estimates (standard errors in parentheses) from missing data methods that incorporate auxiliary variables. True=true parameter values; XDV=extra DV model estimated via DA; MI=Bayesian estimation, with data imputed via MI with the auxiliary variable; DA=standard estimation via DA with no auxiliary variable.

Method	λ_{11}	λ_{21}	λ_{31}	λ_{42}	λ_{52}	λ_{62}	ϕ_{12}
True	4.92	2.96	5.96	3.24	4.32	7.21	0
XDV	5.7(.35)	2.9(.23)	6.0(.41)	2.4(.19)	4.1(.30)	7.1(.52)	-.20(.06)
MI	5.7(.36)	2.9(.22)	6.0(.41)	2.4(.21)	4.3(.27)	7.2(.55)	-.17(.06)
DA	5.7(.36)	2.9(.22)	6.0(.40)	2.1(.20)	3.9(.34)	6.5(.58)	-.23(.08)

Method	ψ_{11}	ψ_{22}	ψ_{33}	ψ_{44}	ψ_{55}	ψ_{66}
True	26.77	13.01	30.93	3.17	8.82	22.5
XDV	21.0(2.89)	13.7(1.13)	34.6(3.43)	3.8(.45)	7.9(1.10)	26.4(3.28)
MI	21.2(2.98)	13.7(1.11)	34.3(3.73)	4.3(.60)	6.8(1.33)	28.9(5.00)
DA	21.3(2.92)	13.7(1.11)	34.2(3.66)	4.1(.57)	7.3(1.49)	27.1(4.62)

Figure Captions

Figure 1. Path diagram of the confirmatory factor analysis model applied to the Holzinger & Swineford (1939) data.


1
SPATIAL
VERBAL
1
φ12 = .48
λ11 = 4.92
λ21 = 2.96
λ31 = 5.96
λ42 = 3.24
λ52 = 4.32
λ62 = 7.21
VISPERC
CUBES
LOZENGES
PARAGRAPH
SENTENCE
WORDMEAN
1
E1
E2
E3
E4
E5
E6
ψ11 = 26.77
ψ22 = 13.01
ψ33 = 30.93
ψ44 = 3.17
ψ55 = 8.82
ψ66 = 22.5