MWSUG 2018 - Paper AA120

Handling Missing Data in Exploratory Factor Analysis Using SAS

Min Chen, Cook Research Incorporated, West Lafayette, IN

ABSTRACT

Exploratory factor analysis (EFA) is a statistical technique to reduce the dimension of multivariate data and to explore the latent structure within the data. Missing data is almost inevitable while conducting EFA. By default, the FACTOR procedure will only include the complete cases which most of the time it is not the first choice of researchers. Given EFA could be performed on individual-level data, correlation or covariance matrix, different formats of data could be fed into SAS and different missing data handling techniques could be applied. This article will demonstrate the above with SAS examples, and comment briefly on how this is generally handled in other statistical software.

INTRODUCTION

Missing data are almost always present in our daily analysis work, and statistical techniques handling missing data have been extensively studied and widely reported. However, missing data in exploratory factor analysis (EFA) represents a unique scenario. EFA aims to perform data reduction and explore the underlying constructs of multivariate data, and the input for EFA could be the raw data, correlation or covariance matrix. EFA procedures are available in most statistical software, but the options for missing data handling varies. The SPSS FACTOR offers listwise deletion, pairwise deletion, and mean substitution options for dealing with missing data. Mplus applies full information maximum likelihood (FIML) to missing data by default. R packages are also available for EFA, with flexibility on handing missing values.

In SAS, FACTOR is the procedure for EFA. By default, it omits the cases with missing values of any analyzed variables from analysis and there is no second option from my own knowledge. This article is not intended to serve as an introduction to EFA or missing data handling techniques but will focus on how to deal with missing values in EFA dataset.

SOURCE DATA

For demonstration purpose, I'm borrowing a dataset from UCLA statistical consulting group website (see the link in the references). Variables, item13 to item24, from the dataset were selected for the analyses below. The subset contains 1428 observations, with 63 having at least one missing item value. The possible item values range from 1 to 5.

LISTWISE DELETION

Under missing data free environment, the EFA results should be reproduced by feeding a correlation or covariance matrix to the FACTOR procedure, if the raw data is not available. However, the story begins when missing values are present. The FACTOR procedure by default takes only the complete cases into the analysis. By running the codes below, SAS logs will include a warning message saying 63 of 1428 observations in data set WORK.DS omitted due to missing values. However, the EFA results would still be generated.

```
proc factor data = ds nfactors = 2;
    var item13 - item24;
run;
```

PAIREWISE DELETION

It is not unreasonable to perform complete case analysis as the first steps. However, most researchers will not be satisfied as observations with incomplete data also carries informative information. One solution is to request the correlation or covariance matrix using CORR procedure, which follows the pairwise deletion rule. Pairwise deletion uses all the information observed and thus preserves more info than the listwise deletion. In CORR procedure, OUTP creates an output data set containing Person correlation statistics. To request for covariance matrix, COV option is needed. While feeding the output dataset to FACTOR procedure, the type of matrix needs to be specify with TYPE=CORR or TYPE=COV. However, the procedure recognize TYPE=CORR by default, and the statement could be dropped.

```
proc corr data = ds outp = pairwise_corr;
    var item13 - item24;
run;
proc corr data = ds cov nocorr out = pairwise_cov(type=cov);
    var item13-item24;
run;


proc factor data = pairwise_corr nfactors = 2;
    var item13 - item24;
run;
proc factor data = pairwise_cov(type=cov) nfactors = 2;
    var item13 - item24;
run;
```

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Pairwise deletion includes different cases for calculations and are only unbiased with missing complete at random (MCAR) data. Substituting a plausible value for missing data point (e.g., mean or regression substitution), which are not covered here, has some clear shortcomings as well. Maximum likelihood approaches have demonstrated some clear advantages, with more unbiased and efficient estimates most of the time (Schlomer et al., 2010).

One great feature of MI procedure is the possibility of using algorithm to obtain expectation maximization (EM) estimation, one of MLEs, of covariance matrix (as indicated below), which could be directly fed to the FACTOR procedure (Graham, 2012). Standard errors and confidence intervals are not provided with EM method. However, this is not an issue for EFA which does not require hypothesis testing.

```
proc mi data=ds nimpute=0;
    em outem=EM_cov;
    var item13 - item24;
run;
```

An alternative way to request MLE of covariance matrix is through CALIS procedure, which is widely known for confirmatory factor analysis and structural equation modeling. It offers the flexibility to choose MLE approaches. For example, full information maximum likelihood (FIML) method in the example below. To request the covariance matrix, MSTRUCT modeling language with variables of interest are specified. The output covariance matrix can then be the input data for FACTOR procedure. Given the exploratory nature of EFA, I would not recommend more sophisticated methods (e.g., multiple imputation).

```
Ods output predcov=FIML_cov;
proc calis data=ds pcorr method=fiml outstat=FIML_cov;
   mstruct var=item13-item24;
run;
```

CONCLUSION

Exploratory factor analysis using the FACTOR procedure in SAS only include complete cases. However, by applying additional SAS procedures it offers greater flexibility to handle missing data in EFA in SAS.

REFERENCES

Factor Analysis | SAS Annotated Output. UCLA: Statistical Consulting Group. From

https://stats.idre.ucla.edu/sas/output/factor-analysis/ (accessed August 1, 2018).

Graham, JW, Missing Data: Analysis and Design, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4614-4018-5_7, © Springer Science+Business Media New York 2012.

Schlomer, GL., Bauman, S., and Card, NA. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57*(1), 1-10.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Min Chen
Cook Research Inc.
1 Geddes Way
West Lafayette, IN 47906
Min.Chen@CookMedical.com