SPSS Syntax for Two-Way Imputation of Missing Test Data Using Factor Loadings

Joost R. van Ginkel
Leiden University,
L. Andries van der Ark
Tilburg University

August 24, 2008

1 Files in this zip-file

- `tw-flmanual.pdf`: this file
- `original.sav`: A simulated data file in SPSS format containing the responses of 300 'respondents' to 41 items, denoted `V1`, ..., `V41`. Variables `V1` to `V40` have ordered answer categories ranging from 0 to 4, variable `V41` is a dichotomous variable with values 1 and 2. There are no missing data. The dataset comes from a simulation study by Van Ginkel, Van der Ark & Sijtsma (2007, Appendix). This file is only included to allow comparing the original data and the completed data.
- `incomplete2.sav`: The SPSS data file (`original.sav`) with 5% of the responses missing. A missing value is indicated by a comma.
- `tw-fl.sps`: SPSS syntax file. This is a read-only file.
- `runtw-fl.sps`: SPSS syntax file. This file may be modified to suit your needs (see Options).

2 About the SPSS Syntax

2.1 The Purpose

The SPSS syntax allows to transform a data file with missing values (an *incomplete* data file) into a data file without missing values (a *completed* data

file). The researcher can use the completed data file for further analysis. It may be noted that using the standard missing data options in SPSS (task bar: `Analyze`, `Missing Value Analysis`) a completed data file can be obtained, where the missing values are replaced by (real-valued) EM-estimates.

2.2 The Method

Applying the SPSS syntax will perform a *two-way imputation with factor loadings* (TW-FL; Van Ginkel, Van der Ark, & Sijtsma, 2007) of the missing values. Method TW-FL estimates the missing score of respondent i on item j in two steps: Item scores are imputed first using method two-way with normally distributed errors (TW-E; see, Bernaards & Sijtsma, 2000). Method TW-E estimates the missing values as follows: let PM_i be the person mean of across all observed scores of person i, let IM_j be the item mean of all observed scores on item j, let OM be the overall mean, and let e_{ij} be a random variable from a normal distribution with mean 0 and a residual variance. The missing value is estimated as

$$X_{ij} = PM_i + IM_j - OM + e_{ij}.$$

Next, a principal components analysis (PCA) with varimax rotation is applied to the completed data set. The factor loadings from this PCA are used for a new imputation procedure. Suppose person i has a missing score on item j, and item j loads highest on rotated factor k with factor loading a_{jk}. Let $PM^*_{i(k)}$ be the person mean of across all observed scores of person i, weighted with the loadings on factor k, let $IM^*_{j(k)}$ be the weighted item mean of item j, and $OM^*_{(k)}$ be the weighted overall mean. The estimate of the missing value of person i on item j equals

$$X_{ij} = PM^*_{i(k)} + IM^*_{j(k)} - OM^*_{(k)} + e_{ij(k)}.$$

The final result is an integer by default. For computational details, see the Appendix. The idea of method TW-FL was to extend method TW-E to multidimensional test data by taking into account the factorial structure of the test.

A simulation study (Van Ginkel, Van der Ark & Sijtsma, 2007) shows that method TW-FL works better for multidimensional test-data than method TW-E. Another advantage is that unlike method TW-E, method TW-FL automatically corrects imputed scores for contra-indicatively worded items. Thus, no items have to be recoded. Biographic variables such as *gender*, and *socio-economic status* are usually not part of a test or questionnaire and

we advocate not to complete missing values of such variables using method TW-FL.

It may be noted that in the investigation of method TW-FL it was assumed that all items had the same number of ordered answer categories.

2.3 Disclaimer and Bugs

It should be emphasized that this SPSS syntax is distributed without any warranty on the part of the authors. Although the SPSS syntax has been tested thoroughly, one can never fully exclude the possibility of errors. The authors appreciate suggestions and reports of detected errors (please enclose SPSS data file). All correspondence can be sent to

Joost R. van Ginkel,
Leiden University,
Faculty of Social and Behavioural Sciences,
Data Theory Group,
PO Box 9555,
2300 RB Leiden,
The Netherlands
jginkel@fsw.leidenuniv.nl

2.4 Known Bugs

1. Procedure `tw-fl.sps` cannot handle variables that are coded as strings, even if they are not included in the analysis. This means that, for example, the categories of variable Gender should be coded numerically (e.g., *1* and *2*) and should not be coded *male* and *female*. `Detected: October 14 2003`.

2. Procedure `tw-fl.sps` cannot be applied to a (sub)set of item scores if one or more respondents have missing values on all item scores in the (sub)set. Cases with all item scores missing should be removed manually. SPSS-command `select cases` does not work here. `Detected: October 14 2003`.

3. For variables with names longer than 8 characters, only the first 8 characters are preserved in the new file, and variable properties (e.g. value labels, formats) are lost. This can be remedied by temporarily renaming these variables, and renaming them back afterwards. `Detected: March 13 2006`.

3 Using the SPSS Syntax

3.1 Preparing Your SPSS Data File

- Remove respondents with many (e.g., 60%) missing values
  There are no general rules available to decide how many missing values a respondent may have before he or she is deleted from the analysis. The idea is that respondents with many missing values provide little useful information. In the file `incomplete2.sav` the largest percentage of missing values is 7.5% (3 out of the 40 responses missing) for respondents 49, and 125.

- Missing values should be defined as `system missing` or `user missing` in SPSS.
  In the incomplete data file `incomplete2.sav` all missing values are system missing. The SPSS-command `recode` can be used to recode numerical values into system missing.

- Rename variables with names longer than 8 characters, and recode string variables.
  As already noted, procedure `tw-fl.sps` has trouble handling string variables and variable names longer than 8 characters. Therefore, string variables must be recoded, and variable names longer than 8 characters must be renamed using the syntax command `rename variables`.

- Determine the items you want to use for imputation.

3.2 Imputation of Missing Data

From now on we assume that `incomplete2.sav` is the incomplete data file and we assume that its variables (`V1`, ..., `V41`) measure different constructs.

1. Copy the SPSS syntax files `tw-fl.sps` and `runtw-fl.sps` to the desired directory.

2. Open SPSS

3. Open the SPSS syntax file `runtw-fl.sps` (task bar: `File`, `Open`, `Syntax`). The SPSS syntax file `runtw-fl.sps` looks like

```
INCLUDE "{path}tw-fl.sps" .
TWOWAYFL FILE = '{path + filename}' .
```

4. Specify the paths of the files `tw-fl.sps` and `incomplete2.sav`. For example, if `tw-fl.sps` is located in

   `C:\Program Files\SPSS\tw-fl.sps`

   and the incomplete data file `incomplete2.sav` is located in

   `C:\mydatasets\incomplete2.sav`

   this should be specified as

   ```
   INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
   TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav' .
   ```

5. Run `runtw-fl.sps` (task bar: `run`, `all`).

6. The completed data file is now in your directory. The name of the completed data file is `incomplete2_imp.sav`. In this completed dataset, an extra variable is added, called `imputation_#`. This variable is only important when doing *multiple* imputation (to be discussed later on). For this example, this variable may be ignored.
   Note that if you rerun the SPSS syntax `incomplete2_imp.sav` will be overwritten. Also note that if your incomplete data file has another name (e.g., `mydata.sav`) then the second syntax line in Step 4 should be changed into

   `TWOWAYFL FILE = 'C:\mydatasets\mydata.sav' .`

   The resulting completed data file will be `mydata_imp.sav`.

3.3 Options

- **Applying imputation to different item subsets.**
  This option is recommended when some variables are not part of the test or questionnaire
  For example if variables `V1`, ..., `V40` measure different aspects of schizotypal personality disorder, and `V41` represents gender.

  Missing values on `V1`, ..., `V40` can be completed by modifying `runtw-fl.sps` into

```
INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav'
       /SELECT = V1 TO V40 .
```

Missing values on `V41` cannot be completed using method TW-FL.

Note that the period '.' that was originally placed at the end of the second line (`TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav'`), has now been moved to the end of the last line. Changing the file `runtw-fl.sps` into

```
INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav' .
         /SELECT = V1 TO V40 .
```

will ignore the `/SELECT = V1 TO V40` subcommand and SPSS will generate an error message.

- **Number of components extracted by PCA**.
  By default, all components with eigenvalues greater than 1 are extracted by PCA. The user can also specify the desired number of components. For example, if the user expects four different dimensions in the data, this may be specified as

```
INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav'
       /NCOMP = 4 .
```

- **Screeplot**.
  When a researcher has no prior knowledge about the number of underlying components in the data, it is common in PCA to inspect the screeplot first. Procedure `tw-fl.sps` can provide the screeplot of the factor solution that is used for imputation. This can be achieved by means of

```
INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav'
       /SCREE = YES .
```

  After inspecting the screeplot, the user can rerun the imputation procedure with the desired number of components.

- **Minimum and maximum scores**.
  These options are not recommended
  SPSS reads the minimum and maximum item score from the incomplete data file. The user can also specify the minimum and maximum item scores. For example, if the required item score range is $1, 2, 3$ this may be specified as

```
INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav'
       /MAX = 3
       /MIN = 1 .
```

- **Multiple imputation**.
  A more sophisticated way of dealing with missing values is called multiple imputation, which creates m (usually $m = 5$) completed datasets that are analyzed using standard statistical methods. Then, the m results are combined into one overall result. Changing the syntax file `runtw-fl.sps` into

```
INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav'
       /m = 5 .
```

  will append five different complete versions of the incomplete dataset after another in the new file `incomplete2_imp.sav`. The indicator variable `imputation_#` contains the dataset number. The interested reader is referred to Schafer (1997).

- **Using several options at a time**.
  One may want to use more options at a time. For example, apply `tw-fl.sps` to variables `V1`, ..., `V40` with four components, and do this $m = 5$ times.

```
INCLUDE "C:\Program Files\SPSS\tw-fl.sps" .
TWOWAYFL FILE = 'C:\mydatasets\incomplete2.sav'
       /SELECT = V1 TO V40
       /NCOMP = 4
       /m = 5 .
```

References

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35*, 321-364.

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

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology, 60*, 315-337.

Appendix: Computation weighted person means, item means, and overall mean

Let $obs(i)$ be the set of all observed scores of person i, let $obs(j)$ be the set of all observed scores on item j, and obs be the set of all observed scores in the data. Furthermore, let x_{mid} be the middle answer category of item j (e.g., for an item with 5 answer categories ranging from 0 to 4, $x_{mid} = 2$). The weighted person mean across all observed scores of person i is obtained as

$$PM^*_{i(k)} = \frac{\sum\limits_{j \in obs(i)} a_{jk} \times \ (X_{ij} - x_{mid})}{\sum\limits_{j \in obs(i)} |a_{jk}|} + x_{mid}.$$

The weighted item mean of item j is computed as

$$IM^*_{j(k)} = \frac{\sum\limits_{i \in obs(j)} a_{jk} \times \ (X_{ij} - x_{mid})}{\sum\limits_{i \in obs(j)} |a_{jk}|} + x_{mid}.$$

Finally, the weighted overall mean equals

$$OM^*_{(k)} = \frac{\sum\sum\limits_{i,j \in obs} a_{jk} \times \ (X_{ij} - x_{mid})}{\sum\sum\limits_{i,j \in obs} |a_{jk}|} + x_{mid}.$$