Conducting an 86-variable Factor Analysis on a Small Computer and Preserving the Mean Substitution Option

Irvin Sam Schonfeld
The City College of New York and
New York State Psychiatric Institute
Candace Erickson
Columbia University
College of Physicians and Surgeons

Abstract

This paper shows how we overcame limitations imposed on us by the memory capacity of the relatively small mainframe we used in conducting a factor analysis in which means are substituted for missing values. Insufficient memory did not permit us to employ SPSSX, with its mean substitution feature, in conducting a factor analysis of 86 variables reflecting ways in which parents cope with the hospitalization of their children. Instead, we employed a two-step solution: (1) we ran SPSSX Condescriptive to create z-score equivalents of the 86 variables and recoded the z variables' system missing values to zeros; (2) the output of the Condescriptive run constituted the input of a BMDP P4M factor analysis run.

Frequently researchers who choose to conduct factor analyses will take advantage of software available in the SPSSX (SPSSX) package. There are several advantages that the SPSSX package offers over previous releases. SPSSX can handle more variables and it can substitute means for missing values. The latter feature is helpful because with it a case is not deleted when a missing variable is encountered.

A disadvantage of SPSSX is that it uses a great of deal of memory. This disadvantage came home to us when we attempted to factor analyze a data set consisting of 86 variables and 271 cases[1]. The variables consisted of parents' responses to 86 of 173 questionnaire items describing behaviors adults use to cope with the problem of having a child in the hospital. Subjects' response choices ranged from "not at all" (0) to "very much" (3). Examples of coping questionnaire items are presented in Figure 1.

If we were to permit the program to delete cases with any missing values, our data set would have been reduced substantially. Of the 271 cases 137 subjects, or 51%, had no missing values; therefore, we would have lost 49% of our subjects. The loss of subjects would have been extremely wasteful since about 27% of the parents failed to complete only 1% of the questionnaire items; 4%, 2% of the items; and another 4%, 3% of the items. About 11% of the parents failed to complete between 4 and 14% of the items. We therefore elected to use the mean substitution option in the SPSSX Factor procedure in order to avoid subject loss.

Unfortunately the four megabyte IBM 4331 computer we used at New York State Psychiatric Institute did not provide sufficient memory to execute the job. The program listing returned the "insufficient storage" error message. We think our solution to the problem might be of interest to readers who face similar storage obstacles to running large factor analyses and other

statistical procedures on small systems. In order to deal effectively with this problem we linearly transformed our original values, and then submitted the new transformed values to a factor analysis program supplied by a software package that uses computer memory more economically than SPSSX.

The data originally resided in an SPSS system file (Nie et al., 1975). Since SPSSX reads SPSS system files, we wrote an SPSSX program to read the system file. The program invoked a series of procedures the first of which, the Condescriptive procedure, created a new set of 86 variables (ZV1 to ZV86). The 86 new (ZV) variables corresponded one-to-one to variables (V1 to V86) in the original data set. Each new variable was the equivalent to the z-score transformation of the corresponding variable in the original data set. The Condescriptive procedure assigns a system missing value to any new (ZV) variable when the corresponding old variable is missing. Thus a parent who did not respond to questionnaire item V30 would receive a system missing value for new variable ZV30. Immediately after the Condescriptive routine was invoked the Recode command was employed to convert all system missing values in the new (ZV) variables to zero. The Recode command in effect substituted means for missing values since zero is, necessarily, the mean of a set of z-scores. Next the Write Outfile procedure was called upon to write out all the new (ZV) variables into a raw data file. Figure 2 depicts the SPSSX program that operated upon the original 86 variables.

BMDP (Brown et al., 1983) provides the user an economical alternative to SPSSX. When the user runs a BMDP job, one program out of the BMDP library of programs is called up. By contrast, when SPSSX is run, the entire SPSSX library of programs is called up. The advantage inherent in the SPSSX approach is that multiple procedures can be invoked in a single run. The disadvantage is that a great deal of memory is required to store the program

Figure 1

Circle the number that corresponds to the response that best describes your experience in the last week. If your child has been in the hospital for less than a week, circle the number that corresponds to the response that best describes your experience since your child entered the hospital.

	very much	pretty much	just a little	not at all
1. I think the doctors have made a mistake and that my child doesn't really need to be in the hospital.....	3	2	1	0
2. I watch myself doing things, and it feels like I'm watching someone else.....	3	2	1	0
3. I want someone around to hold or comfort me.....	3	2	1	0
4. Something ironic or humorous usually breaks the tension.....	3	2	1	0

Figure 2

SPSSX program to output data

```
COMMENT     SPSSX PROGRAM TO OUTPUT DATA TO BE READ BY BMDP PROGRAM.
FILE HANDLE  SYSFILE/NAME='HOSP SYSFILE A'
FILE HANDLE  ZDATA/NAME='Z DATA A'
GET FILE     SYSFILE
COMMENT ********************************************************************
     THE PURPOSE OF THE NEXT 6 STATEMENTS IS TO INCLUDE ONLY THOSE
     SUBJECTS WHO HAVE FEWER THAN 20% MISSING VALUES ON ALL 173 VARIABLES.
  ********************************************************************************
DO REPEAT    A = V1 TO V173/B=CT1 TO CT173
COUNT        B = A (9)
END REPEAT
COMPUTE      TOT9 = SUM (CT1 TO CT173)
COMPUTE      TOT9PER = TOT9/173
SELECT IF    (TOT9PER LT .20)
COMMENT ********************************************************************
     THE PURPOSE OF OPTION 3 OF THE CONDESCRIPTIVE PROCEDURE IS TO CREATE A
     SET OF NEW VARIABLES, ZV1 TO ZV86, WHICH ARE Z-SCORE TRANSFORMATIONS OF
     OLD VARIABLES, V1 TO V86.  WHEN A SUBJECT RECEIVED A MISSING VALUE FOR
     ONE OF THE OLD VARIABLES, S/HE IS ASSIGNED A SYSTEM MISSING VALUE ON THE
     CORRESPONDING NEW VARIABLE.
  ********************************************************************************
CONDESCRIPTIVE                V1 TO V86
OPTIONS                       3
```

Figure 2, continued

```
COMMENT **********************************************************************
      THE PURPOSE OF THE RECODE STATEMENT IS TO CONVERT THE SYSTEM MISSING
      VALUES FOR THE NEW 'ZV' VARIABLES TO ZEROS.
 **********************************************************************
RECODE ZV1 TO ZV86 (MISSING = 0)
COMMENT **********************************************************************
      THE PURPOSE OF THE WRITE OUTFILE STATEMTNENT IS TO WRITE OUT A
      RECTANGULAR DATA FILE THAT CAN BE READ BY A BMDP PROGRAM.
 **********************************************************************
WRITE OUTFILE = ZDATA TABLE
 /ZV1 TO ZV6
 /ZV7 TO ZV12
 /ZV13 TO ZV18
 /ZV19 TO ZV24
 /ZV25 TO ZV30
 /ZV31 TO ZV36
 /ZV37 TO ZV42
 /ZV43 TO ZV48
 /ZV49 TO ZV54
 /ZV55 TO ZV60
 /ZV61 TO ZV66
 /ZV67 TO ZV72
 /ZV73 TO ZV78
 /ZV79 TO ZV84
 /ZV85 TO ZV86
EXECUTE
FINISH
```

library, rendering insufficient memory for jobs like ours that are conducted on small systems. We could not run the SPSSX driven factor analysis even when we created a two or a three megabyte virtual machine. We, therefore, elected to use the output of the SPSSX Write Outfile procedure, that is, the coping items rescaled as z-scores with zeros having replaced missing values, as the input for the BMPD Factor Analysis program, P4M. We successfully ran BMDP P4M with storage defined at 1.5 megabytes. Figure 3 shows the BMDP factor analysis program.

We thus overcame a disadvantage of the BMDP Factor Analysis program, namely, that P4M does not include a mean substitution option. The listing of the BMDP program provides a check on the adequacy of the procedure just employed. The listing included the means and standard deviations of each ZV variable. The listing showed that each of the ZV means was within rounding error of zero, and that each standard deviation attained a value of one or, as would be expected from the additional zero scores, values slightly less than one.

Figure 3

BMDP program to read output from SPSSX program and perform the factor analysis

```
COMMENT      BMDP PROGRAM TO BE RUN UNDER P4M.
/PROBLEM     TITLE IS 'HOSPITALIZATION STUDY'.
/INPUT       VARIABLES ARE 86.
                 FORMAT IS FREE.
                 CASE = 271.
/VARIABLE NAMES ARE ZV1 TO ZV86.
                 USE = 1 TO 86.
/FACTOR      NUMB = 10.
/END
----------DATA IS PLACED HERE----------
```

References

Brown, M.B., Engelman, L., Frane, J.W., Hill, M.A., Jennich, R.I., & Toporek, J.D. (1983). BMDP Statistical Software. Berkeley, CA: University of California Press.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. (1975). Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill.

SPSSX (1983). SPSSX User's Guide. New York: McGraw-Hill.

Footnote

[1] We recognize that it would have been desirable to have perhaps 130 additional subjects in conducting the factor analysis. Actually the factor analysis was not our primary vehicle for studying the ways parents coped with having children in the hospital. The factor analysis was conducted as an adjunct to and a check on a more important set of analyses we had performed earlier. In the earlier analyses we constructed a priori scales by combining items clinical experience suggested went together. Typically, the scales we constructed had satisfactory internal consistency reliabilities as measured by the coefficient alpha. Generally, the items factored in ways anticipated by our a priori scales.