*Behavior Research Methods, Instruments, & Computers*
*1998, 30 (3), 506-516*

# CPCA: A program for principal component analysis with external information on subjects and variables

MICHAEL A. HUNTER
*University of Victoria, Victoria, British Columbia, Canada*

and

YOSHIO TAKANE
*McGill University, Montreal, Quebec, Canada*

A program is described for principal component analysis with external information on subjects and variables. This method is called *constrained principal component analysis* (CPCA), in which regression analysis and principal component analysis are combined into a unified framework that allows a full exploration of data structures both within and outside known information on subjects and variables. Many existing methods are special cases of CPCA, and the program can be used for multivariate multiple regression, redundancy analysis, double redundancy analysis, dual scaling with external criteria, vector preference models, and GMANOVA (growth curve models).

Regression analysis and principal component analysis (PCA) typically proceed from opposite ends of the data analysis spectrum. Regression methods are usually considered explanatory techniques in which the variability in a set of dependent variables is partitioned into two orthogonal components: variability accounted for by known external information (e.g., one or more predictor variables), and variability that is independent of external information (i.e., error variability). PCA, on the other hand, is an exploratory technique used to investigate structures in data when no external information is available. In this paper, we describe a program for constrained principal component analysis (CPCA), in which features of regression analysis and PCA are combined into a unified framework that captures advantages of both.

Takane and Shibayama (1991) have described in detail the rationale and methodology underlying CPCA. Briefly, the analysis can be conceptualized as proceeding in two distinct stages: (1) an external (regression) analysis in which a data set is partitioned into predictable and error variation, and (2) an internal analysis in which PCA is applied to the results of the external analysis. The external analysis decomposes an $N$ subject $\times$ $n$ variable matrix of dependent variables $\mathbf{Z}$, into variation predictable from external information on subjects and variables and variation unrelated to the external information. Assuming the information on subjects (e.g., age, gender, IQ, treatment condition, etc.) is coded into an $N \times p$ ($\ll N$) matrix, $\mathbf{G}$, and the information on variables (e.g., occasions, design matrix for pair comparisons, characteristics such as order, familiarity, and complexity of stimuli presented in a memory task, etc.) into an $n \times q$ ($\ll n$) matrix, $\mathbf{H}$, the model can be written as

$$\mathbf{Z} = \mathbf{GMH}' + \mathbf{BH}' + \mathbf{GC} + \mathbf{E}, \qquad (1)$$

where $\mathbf{M}$ ($p \times q$), $\mathbf{B}$ ($N \times q$), and $\mathbf{C}$ ($p \times n$) are matrices of coefficients to be estimated, and $\mathbf{E}$ ($N \times n$) is a matrix of error components. The four terms on the right side of the equation explain unique (orthogonal) portions of the variation in $\mathbf{Z}$. The first term pertains to variation that can be explained uniquely by information on subjects and variables combined, the second term by information on variables alone, and the third term by information on subjects alone. The fourth term includes variation in $\mathbf{Z}$ that is independent of information on subjects and variables.

At this stage, the analysis is related to multivariate multiple regression. In fact, if only the main data $\mathbf{Z}$, and information on subjects, $\mathbf{G}$, were available, then only the third and fourth terms in Equation 1 would be estimated, and the parameters in $\mathbf{C}$ would equal the regression coefficients obtained from a standard multivariate multiple regression analysis (assuming both $\mathbf{G}$ and $\mathbf{Z}$ contained continuous variables). Thus, the Stage 1 full model can be viewed as a doubly multivariate multiple regression analysis with external information not only on subjects but also on variables. Of course, the full model may not always be desired, or other submodels may be of substantive interest along with the full model. In such cases, it is possible to obtain $\mathbf{Z} = \mathbf{BH}' + \mathbf{E}$ (by simply omitting $\mathbf{G}$ from the analysis), or $\mathbf{Z} = \mathbf{GC} + \mathbf{E}$ (by omitting $\mathbf{H}$ from the analysis).

The second stage, the internal analysis, applies PCA to each of the decomposed submatrices from the Stage 1

Correspondence should be addressed to M. A. Hunter, Department of Psychology, University of Victoria, P. O. Box 3050, Victoria, BC, V8W 3P5 Canada (e-mail: mhunter@uvic.ca).

**Table 1**
**Summary of Analyses**

| Number | Data | External Analysis | Internal Analysis |
|---|---|---|---|
| 1 | **Z** | none | PCA of **Z** |
| 2 | **G**,**Z** | Regress **Z** on **G** to obtain ${}_{G}\mathbf{Z}'$ | PCA of ${}_{G}\mathbf{Z}'$ |
| 3 | **G**,**Z** | Regress **Z** on **G** to obtain ${}_{G}\mathbf{Z}'$ | PCA of $\mathbf{Z} - {}_{G}\mathbf{Z}'$ |
| 4 | **Z**,**H** | Regress **Z** on **H** to obtain $\mathbf{Z}_{H}'$ | PCA of $\mathbf{Z}_{H}'$ |
| 5 | **Z**,**H** | Regress **Z** on **H** to obtain $\mathbf{Z}_{H}'$ | PCA of $\mathbf{Z} - \mathbf{Z}_{H}'$ |
| 6 | **G**,**Z**,**H** | Regress **Z** on **G** and and **H** to obtain ${}_{G}\mathbf{Z}_{H}'$ | PCA of ${}_{G}\mathbf{Z}_{H}'$ |
| 7 | **G**,**Z**,**H** | Regress **Z** on **H** to obtain $\mathbf{Z}_{H}'$ and on **G** and **H** to obtain ${}_{G}\mathbf{Z}_{H}'$ | PCA of $\mathbf{Z}_{H}' - {}_{G}\mathbf{Z}_{H}'$ |
| 8 | **G**,**Z**,**H** | Regress **Z** on **G** to obtain ${}_{G}\mathbf{Z}'$ and on **G** and **H** to obtain ${}_{G}\mathbf{Z}_{H}'$ | PCA of ${}_{G}\mathbf{Z}' - {}_{G}\mathbf{Z}_{H}'$ |
| 9 | **G**,**Z**,**H** | Regress to obtain ${}_{G}\mathbf{Z}'$, $\mathbf{Z}_{H}'$ and ${}_{G}\mathbf{Z}_{H}'$ | PCA of $\mathbf{Z} - {}_{G}\mathbf{Z}' - \mathbf{Z}_{H}' - {}_{G}\mathbf{Z}_{H}'$ |

analysis. Alternatively, submatrices can be recombined for PCA if desired. For example, the first and second terms can be combined, amounting to PCA of that part of **Z** that is predictable overall from external information on variables (including that which is unique to **H** and that part which **H** shares with **G**). Similarly, the first and third terms can be combined, and the part of **Z** that is predictable from **G**, both uniquely and in common with **H**, can be component analyzed.

In all, nine analyses are possible allowing a full exploration of main data structures both inside and outside the known external information (see Table 1, and the Example Analyses section). The nine analyses include: (1) PCA of **Z** unconstrained by **G** or **H**. This is a straightforward principal component analysis of **Z**. (2) PCA of **Z** constrained by **G** and (3) PCA of **Z** independent of **G**. Analysis 2 is a component analysis of the part of **Z** that is predictable overall from subject information, whereas Analysis 3 component analyzes the part of **Z** from which subject information has been partialed. In these analyses, external information on variables is ignored, which amounts to combining the first and third terms from the full model presented above (obviously, these would be the analyses of choice if no stimulus information were available). Notice, however, that Analysis 3 is not an analysis of **E** from the full model. The error component will differ depending on whether **H** is included or ignored. (4) PCA of **Z** constrained by **H**, and (5) PCA of **Z** independent of **H**. These analyses are equivalent to Analyses 2 and 3 above, except that now information on variables is incorporated and subject constraints are ignored (or are unavailable). (6–9) PCA of the four terms in the full model.

## RELATIONS TO OTHER METHODS

CPCA is a very general method for finding a reduced rank representation of the relationship among sets of variables, and it subsumes several existing methods as special cases. For example, if only **G** and **Z** were available and both contained continuous variables, CPCA would reduce to what has been called *principal components of instrumental variables* (Rao, 1964), *reduced rank regression* (Anderson, 1951), and *redundancy analysis* (Van den Wollenberg, 1977). The unique aspect of CPCA is that it extends redundancy analysis to incorporate external information on variables and on subjects, suggesting as a possible synonym for CPCA the term *double redundancy analysis* (or if only **Z** and **H** are available, we refer to the analysis as *redundancy analysis of structured variables*). Additionally, unlike standard redundancy analysis, CPCA focuses not only on structuring the predictable parts of **Z** but also on structuring

**Table 2**
**Analysis of Main Data Matrix (Z)**

ANALYSIS 1: Analysis of **Z** ❶
SS & %SS for the external analysis ❷
SS= 484.000 %SS= 100.000
SS & %SS for the internal analysis ❸
SS= 356.148 %SS (per term SS)= 73.584 %SS (per total SS)= 73.584

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 253.493 | 102.654 |
| %SS (per term SS) | 52.375 | 21.210 |
| %SS (per total SS) | 52.375 | 21.210 |

Matrix A of Component Loadings ❹

| | | |
|---|---|---|
| 1 | .761 | −.464 |
| 2 | .763 | .357 |
| 3 | −.592 | .573 |
| 4 | −.903 | .335 |
| 5 | .810 | .321 |
| 6 | −.678 | −.487 |
| 7 | −.574 | −.549 |
| 8 | .883 | .246 |
| 9 | −.450 | −.619 |
| 10 | .638 | −.378 |
| 11 | −.774 | .566 |

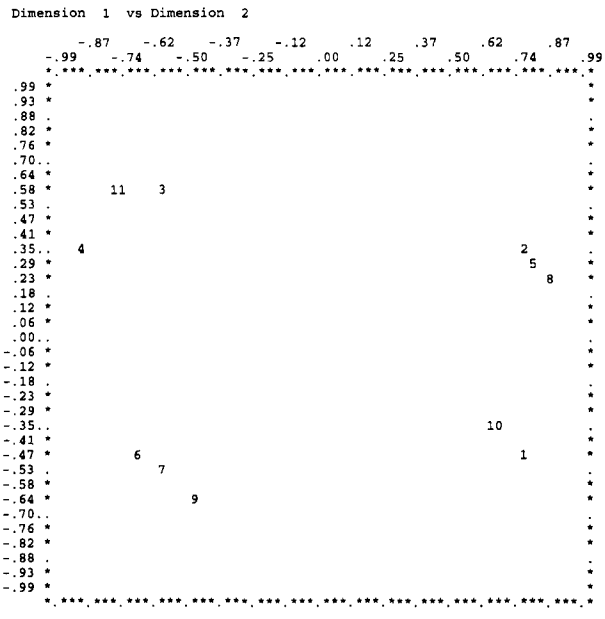


**Figure 1. Plot of loading matrix for Analysis 1.**

residuals in **Z** after variation accounted for by external information has been removed.

Other special cases of CPCA include growth curve models (GMANOVA; see Khatri, 1966; Rao, 1965) and two-way CANDELINC (Carroll, Pruzansky, & Kruskal, 1980), which both analyze only the first term in Decomposition 1. If **Z** is a matrix of pairwise preference data, **G** is a vector of ones, and **H** is a design matrix for paired comparison, then three vector models for pairwise preference data (e.g., De Soete & Carroll, 1983; Heiser & de Leeuw,1981; Takane and Shibayama,1988) are all special cases of CPCA that differ according to which term in Decomposition 1 is analyzed. (See Takane & Shibayama, 1991, for details and for a discussion of how CPCA relates to yet other methods. Example applications are also included.)

A commonly used method for investigating the reduced rank relationship between two sets of variables (e.g., **Z** and **G**) is canonical correlation. It obtains linear combinations (components) of each set of variables under the constraint that corresponding pairs of components are maximally correlated. The problem with canonical correlation is that components that maximally correlate might account for only a small portion of the variation in their respective sets. As a result, a component in one set (e.g., **G**) might correlate highly with a component in a second set (e.g., **Z**) but might explain only a small portion of the overall variation in Set **Z** (and vice versa). This overall portion of variation in one set that is explained by a component of another set is called *redundancy*, and it is this quantity that CPCA maximizes(for a full description of the difference between canonical correlation and redundancy analysis, see Van den Wollenberg, 1977). CPCA also differs from canonical correlation by accommodating information on subjects and on variables. A method that does incorporate both subject and variable constraints is GENFOLD2 (DeSarbo & Rao, 1984). However, GENFOLD2 does constrained unfolding analysis by fitting a distance model using an iterative gradient procedure, whereas CPCA does constrained principal component analysis by fitting a bilinear model.

## THE CPCA PROGRAM

### Program Description

CPCA is written in Fortran for the IBM PC XT, AT, PS/2, or compatibles under MS-DOS/PC-DOS 3.0 or higher. A math coprocessor is supported. The program can analyze problems with up to 1,000 subjects and 40 variables in each of **Z**, **G**, and **H**. At this point in development, the program does not accommodate missing data. At least 550 K of available memory is required. The program is very efficient. For example, using a 486DX2/66

**Table 3**
**Analysis of Z Regressed on G (the Subjects Information Matrix)**

ANALYSIS 2: Analysis of P(**G**)**Z** ❶
SS & %SS for the external analysis ❷
SS= 346.747 %SS= 71.642

Matrix **C** transposed (in Model **Z** = **GC** + **E**) ❸

| | | | | |
|---|---|---|---|---|
| 1 | .422 | −1.267 | 1.202 | −.357 |
| 2 | 1.484 | −.399 | −.152 | −.933 |
| 3 | −.608 | 1.145 | −.652 | .115 |
| 4 | −.984 | 1.293 | −.895 | .586 |
| 5 | 1.466 | −.538 | −.014 | −.914 |
| 6 | −.712 | −.157 | −.498 | 1.367 |
| 7 | −.667 | −.125 | −.306 | 1.098 |
| 8 | 1.355 | −.547 | .294 | −1.102 |
| 9 | −.600 | −.258 | −.039 | .898 |
| 10 | −.068 | −.731 | 1.478 | −.679 |
| 11 | −.735 | 1.451 | −.893 | .177 |

SS & %SS for the internal analysis ❹
SS= 319.763 %SS (per term SS)= 92.218 %SS (per total SS)= 66.067

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 243.727 | 76.036 |
| %SS (per term SS) | 70.290 | 21.929 |
| %SS (per total SS) | 50.357 | 15.710 |

Matrix **A** of Component Loadings ❺

| | | |
|---|---|---|
| 1 | .764 | −.500 |
| 2 | .741 | .456 |
| 3 | −.619 | .344 |
| 4 | −.933 | .257 |
| 5 | .792 | .373 |
| 6 | −.613 | −.394 |
| 7 | −.502 | −.360 |
| 8 | .870 | .313 |
| 9 | −.346 | −.40 |
| 10 | .621 | −.440 |
| 11 | −.796 | .453 |

Matrix of correlations between **G** and **F** ❻

| | | |
|---|---|---|
| 1 | .686 | .480 |
| 2 | −.567 | .667 |
| 3 | .458 | −.591 |
| 4 | −.576 | −.555 |

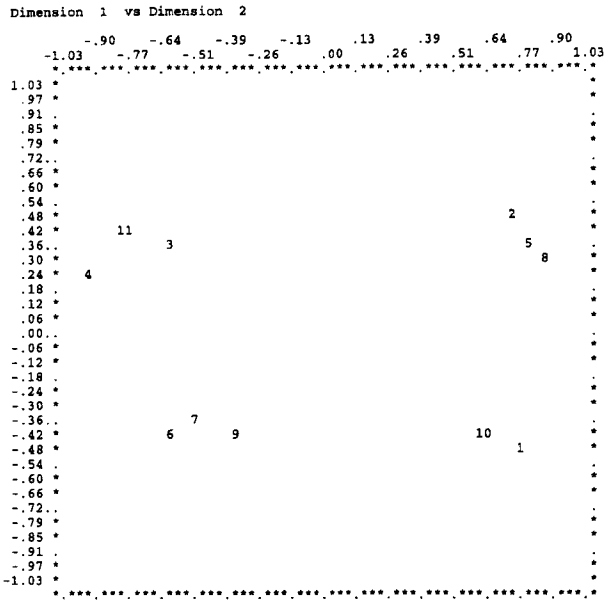


**Figure 2. Plot of loading matrix for Analysis 2.**

computer, an analysis of 200 subjects with 11 variables in **Z**, 4 variables in **G**, and 3 variables in **H** required 24 sec of CPU time to perform all nine analyses.

## Input

Input to the program consists of job parameter information on number of subjects, number of variables in **Z**, number of variables in **G**, number of variables in **H**, data transformations (raw data, column centered, or standardized), number of components, input/output options, and input data format. This information along with the data for *N* subjects (both **Z** and **G** data), a design matrix for **H**, and optional target matrices for procrustes rotations are all placed in an input file using any standard word-processing program or editor. The program is run by typing cpca-t followed by an input filename and an optional output filename. The syntax is

cpca-t infile.ext outfile.ext

where infile.ext is the file containing job parameter information and the data, and outfile.ext is an optional file that receives the results of the analysis.

## Output

For Analysis 1, the program output includes total sums of squares (SS) for the main data, SS and percent sums of squares (%SS) for each of the components retained, component loadings, and component scores. Loadings and scores can be plotted for each pair of components, and, to aid interpretation, optional varimax, promax, and procrustes rotations are available. When rotations are requested, SS and %SS of squares for rotated components are printed, and, in the cases of promax rotation, correlations among components are given.

For all analyses in which the main data are constrained by external information (e.g., Analyses 2, 4, 6, 7, and 8 above), the program prints out (1) total SS for the main data, (2) SS and %SS accounted for by the external information, (3) matrices of parameter estimates (i.e., matrix **M**, **B**, or **C**), (4) SS and %SS for each of the components retained in the internal analysis, (5) loadings of the main variables on components constrained by external information, and scores on those components (optional plots are available), (6) weights and loadings of the external variables on the externally defined components (e.g., these would be redundancy variate weights and loadings for the case in which only **G** and **Z** are analyzed), and (7) all of the above for optional varimax, promax, and procrustes rotations, as well as correlations among factors for promax rotated components.

For all analyses of error terms (e.g., Analyses 3, 5, and 9), the program printout includes total SS for the main data, SS and %SS for the part of the main data that is independent of the external information, SS and %SS for each of the components retained in the internal analysis, and component loadings and scores. Again, optional plots and rotations are available.

## Availability

A copy of the program and the program user's manual may be obtained from the first author or can be downloaded from http://castle.uvic.ca/psyc/files.html

## EXAMPLE CPCA ANALYSES

This example analyzes scores on four groups, each containing 11 psychiatric patients (G1 = manic-depressive [depressed]; G2 = manic-depressive [manic]; G3 = simple schizophrenia; G4 = paranoid schizophrenia) on 11 psychopathological items from the Brief Psychiatric Rat-

**Table 4**
**Analysis of Residuals After Regressing Z on G**

ANALYSIS 3: Analysis of Q(G)Z ❶
SS & %SS for the external analysis ❷
SS= 137.253 %SS= 28.358
SS & %SS for the internal analysis ❸
SS= 73.887 %SS (per term SS)= 53.832 %SS (per total SS)= 15.266

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 45.523 | 28.364 |
| %SS (per term SS) | 33.167 | 20.665 |
| %SS (per total SS) | 9.406 | 5.860 |

Matrix A of Component Loadings ❹

| | | |
|---|---|---|
| 1 | .761 | −.464 |
| 1 | .139 | −.026 |
| 2 | .035 | −.099 |
| 3 | .432 | .403 |
| 4 | .054 | −.063 |
| 5 | .119 | .117 |
| 6 | −.317 | .140 |
| 7 | −.572 | −.354 |
| 8 | .067 | −.059 |
| 9 | −.594 | .551 |
| 10 | .089 | .006 |
| 11 | .132 | 2.035 |

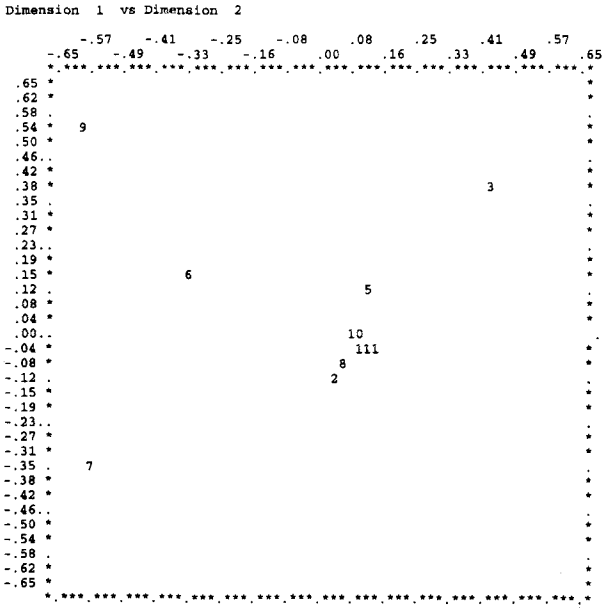


**Figure 3. Plot of loading matrix for Analysis 3.**

ing Scale. The 11 items are (1) emotional withdrawal, (2) guilt feelings, (3) tension, (4) grandiosity, (5) depressive mood, (6) suspiciousness, (7) hallucinatory behavior, (8) motor retardation, (9) unusual thought content, (10) blunted affect, and (11) excitement. For purposes of the analyses presented below, **Z** is a 44 × 11 matrix containing scores on the 11 items for all 44 patients, **G** is a 44 × 4 matrix of dummy coded variables indicating the four patient-type groups, and **H** is an 11 × 4 matrix of dummy coded variables indicating item-type classifications. For this example, we classified together items that might be expected to be symptomatic of depression (guilt, depressive mood, and motor retardation), manic state (tension, grandiosity, and excitement), simple schizophrenia (emotional withdrawal and blunted affect), and paranoid schizophrenia (suspiciousness, hallucinatory behavior, and unusual thought content). In what follows, we show the results of each of the nine analyses provided by the CPCA program, along with explanatory comments. Note that optional output, such as printed factor scores and graphs of factor scores, and the results of various rotations are excluded due to space constraints.

Table 2 and Figure 1 show the output for Analysis 1. The highlighted numbers below correspond to those shown in Table 2.

❶ Analysis 1 is a straightforward principal components analysis of **Z**.

❷ The SS and %SS are given for the external (or regression) analysis. In Analysis 1, there are no external constraints on **Z**; therefore, in this case, the values of 484 and 100% reflect the total variability in **Z**.

❸ Summary results are given for the internal (or PCA) analysis. The SS term reflects the amount of variability accounted for by the number of components retained. In

**Table 5**
**Analysis of Z Regressed on H (the Variables Information Matrix)**

ANALYSIS 4: Analysis of **ZP(H)** ❶
SS & %SS for the external analysis ❷
SS= 400.371 %SS= 82.721
Matrix **B** in Model **Z** = **BH** (transpose) + **E** ❸

| | | | | |
|---|---|---|---|---|
| 1 | −.038 | 1.629 | −.464 | −.883 |
| 2 | .638 | 1.465 | −1.028 | −1.341 |
| 3 | −.279 | 1.163 | −1.039 | .252 |
| 4 | −.714 | 1.722 | −.574 | −.231 |
| 5 | −.123 | 1.350 | −1.109 | −.046 |
| 6 | .273 | 1.683 | −.741 | −.909 |
| 7 | −.026 | 1.381 | −.332 | −.523 |
| 8 | .178 | 1.663 | −.919 | −1.174 |
| 9 | .604 | 1.502 | −.838 | −.445 |
| 10 | .705 | .996 | −.565 | −.759 |
| 11 | .730 | 1.227 | −.922 | −1.198 |
| 12 | −1.036 | −.480 | 1.171 | .336 |
| 13 | −1.006 | −.538 | 1.093 | .807 |
| 14 | −.897 | −.552 | 1.486 | −.176 |
| 15 | −.822 | −.332 | 1.767 | −.206 |
| 16 | −.780 | −.424 | 1.273 | −1.228 |
| 17 | −1.174 | −.412 | 1.257 | −.529 |
| 18 | −1.006 | −.343 | 1.363 | −.897 |
| 19 | −.897 | −.379 | 1.773 | −.704 |
| 20 | −1.006 | −.671 | .915 | .966 |
| 21 | −1.187 | −1.093 | .884 | .268 |
| 22 | −1.174 | −.214 | 1.279 | −.617 |
| 23 | 1.584 | −.292 | −.558 | −.588 |
| 24 | .990 | .607 | −.910 | −.363 |
| 25 | 1.259 | −.349 | −1.180 | .032 |
| 26 | 1.888 | −.350 | −.767 | −.619 |
| 27 | .865 | .145 | −1.071 | .442 |
| 28 | 1.424 | −.063 | −1.139 | −.301 |
| 29 | 1.317 | .487 | −.072 | −1.414 |
| 30 | 1.670 | .081 | −.956 | −.028 |
| 31 | 1.024 | −.305 | −.920 | .605 |
| 32 | 1.482 | .445 | −.834 | −.730 |
| 33 | 1.233 | .064 | −.539 | −.129 |
| 34 | −.285 | −.995 | −.111 | 1.552 |
| 35 | −.945 | −.809 | .443 | 1.070 |
| 36 | −.621 | −1.066 | .099 | 1.020 |
| 37 | −.678 | −.459 | −.304 | 1.620 |
| 38 | −.396 | −.971 | .658 | .810 |
| 39 | −.519 | −1.071 | .007 | 1.330 |
| 40 | −.860 | −.942 | .902 | .655 |
| 41 | −.480 | −1.022 | .389 | .737 |
| 42 | −.234 | −1.369 | .497 | 1.084 |
| 43 | −.956 | −.780 | .219 | 1.342 |
| 44 | .277 | −1.331 | .420 | 1.111 |

SS & %SS for the internal analysis ❸
SS= 350.107 %SS (per term SS)= 87.446 %SS (per total SS)= 72.336

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 249.789 | 100.318 |
| %SS (per term SS) | 62.389 | 25.056 |
| %SS (per total SS) | 51.609 | 20.727 |

Matrix **A** of Component Loadings ❺

| | | |
|---|---|---|
| 1 | .696 | −.426 |
| 2 | .819 | .313 |
| 3 | −.760 | .488 |
| 4 | −.760 | .488 |
| 5 | .819 | .313 |
| 6 | −.566 | −.550 |
| 7 | −.566 | −.550 |
| 8 | .819 | .313 |
| 9 | −.566 | −.550 |
| 10 | .696 | −.426 |
| 11 | −.760 | .488 |

Matrix of correlations between **H** and **A**

| | | |
|---|---|---|
| 1 | .708 | .434 |
| 2 | −.638 | .668 |
| 3 | .464 | −.432 |
| 4 | −.473 | −.728 |

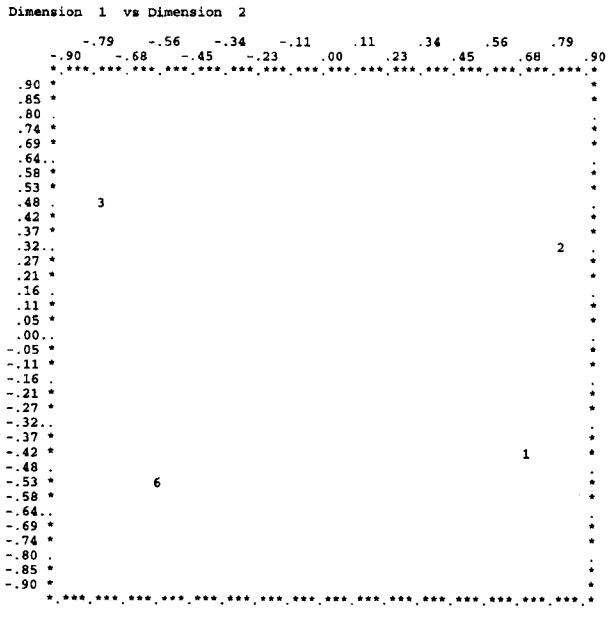


**Figure 4. Plot of loading matrix for Analysis 4.**

this example, the amount of variability accounted for by the first two principal components was 356.148. The %SS (per term SS) reexpresses this variability into a percentage of the variance in **Z** that is related to the external constraints. In contrast, the %SS (per total SS) reflects the percentage of total variability accounted for by the number of components retained. In Analysis 1, there are no external constraints; therefore, these two values are the same. The example results show that the first two principal components account for 73.584% of the total variation in **Z**. The dimensionwise SS and %SS give the same information, but for each component considered separately. For example, the current analysis indicates that the amount of variability accounted for by the first component is 253.493, which accounts for 52.375% of the total variance in **Z**.

❶ Matrix **A** contains the (unrotated) loadings of the variables in **Z** on the first two principal components, and the loadings are plotted in Figure 1 (if more than two components are retained, pairwise plots are produced). The plot indicates that the relations among the items are fairly clear: Items characteristic of depression (Items 2, 5 and 8) are located in the third quadrant, those characteristic of manic state (Items 3, 4, and 11) are in the fourth quadrant, simple schizophrenic symptoms (Items 1 and 10) are in the second quadrant, and those related to paranoid schizophrenia (Items 6,7, and 9) are in the first quadrant.

The results for Analysis 2 are shown in Table 3 and Figure 2. The highlighted numbers below correspond to those shown in Table 3.

❶ Analysis 2 is an analysis of the part of **Z** that is predictable from **G**—that is, it is an analysis of ${}_{G}\mathbf{Z}'$, where ${}_{G}\mathbf{Z}' = \mathbf{GC}$, and is equivalent to redundancy analysis. In the current example, it is an analysis of between-groups variation.

❷ The values in this section show that the between-groups variation is 346.747, which accounts for 71.642% of the total variation. This total percent of variation in **Z** that can be accounted for by **G** is called the total redundancy.

❸ Matrix **C** is a matrix of parameter estimates for predicting **Z** from **G**. In general, this matrix contains the regression coefficients that would be produced via a multivariate multiple regression of **Z** on **G**. In the current example, **G** is a matrix of dummy variables coding group membership, so the values in matrix **C** are the means for each group on each of the 11 variables.

❹ This section indicates that when PCA is applied to the between-groups variation, ${}_{G}\mathbf{Z}'$, the variation accounted for by a two-component solution is 319.76, which represents 92.218% of the between-groups variation and 66.067% of the total variation. In the language of redundancy analysis, we would say that the first two predictor (redundancy) variates account for 66.067% of the overall variance in **Z**, which represents 92.218% of the total redundancy. The dimensionwise SS and %SS give corresponding values separately for the first and second components. Thus, the first two redundancy variates account for 50.357% and 15.710%, respectively, of the total variance in **Z**.

Matrix **A** shows the loading matrix that arises when PCA is applied to ${}_{G}\mathbf{Z}'$. These loadings are plotted in Figure 2 and are strikingly similar to the ones presented earlier for the unconstrained solution. Overall, these results suggest that most of the variation in **Z** (and the factor structure arising out of it), is due to differences among patient groups. In redundancy analysis, Matrix **A** would be referred to as the matrix of cross-loading of the criterion variables (**Z**) on the predictor (redundancy) variates.

**Table 6**
**Analysis of Residuals After Regressing Z on H**

ANALYSIS 5: Analysis of ZQ(H) ❶
SS & %SS for the external analysis ❷
SS= 83.629 %SS= 17.279
SS & %SS for the internal analysis ❸
SS= 47.968 %SS (per term SS)= 57.358 %SS (per total SS)= 9.911

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 28.680 | 19.288 |
| %SS (per term SS) | 34.294 | 23.064 |
| %SS (per total SS) | 5.926 | 3.985 |

Matrix A of Component Loadings ❹

| | | |
|---|---|---|
| 1 | .016 | −.226 |
| 2 | .119 | −.057 |
| 3 | −.372 | −.017 |
| 4 | .218 | −.068 |
| 5 | −.124 | −.083 |
| 6 | −.001 | −.443 |
| 7 | .454 | .219 |
| 8 | .005 | .141 |
| 9 | −.453 | .224 |
| 10 | −.016 | .226 |
| 11 | .154 | .086 |

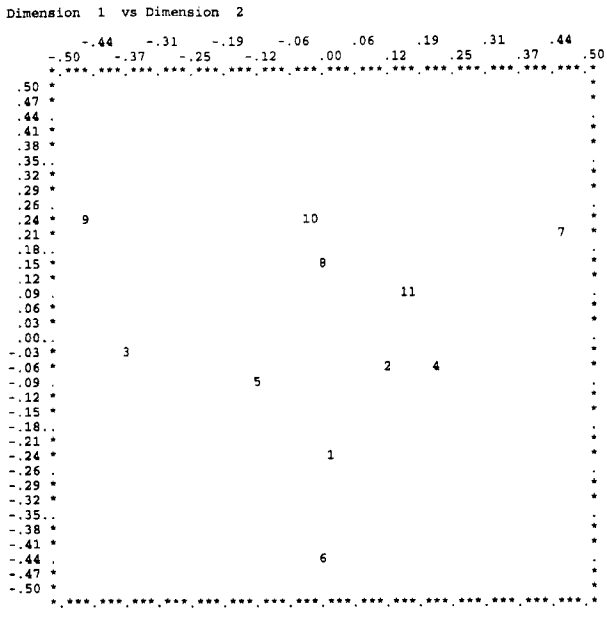


**Figure 5. Plot of loading matrix for Analysis 5.**

❻ The matrix of correlations between **G** and **F** is a loading matrix that reflects the structure of the predictor (redundancy) variates. The sign and size of the loadings in the current example indicate that the difference between Groups 1 and 3 (depressed and simple schizophrenic patients) and Groups 2 and 4 (manic and paranoid schizophrenics) accounts for most of the variability in **Z** (50.357% as seen above). The second redundancy variate, accounting for 15.71% of the variance in **Z** contrasts depressed and manic patients with the two groups of schizophrenics.

The results for Analysis 3 are shown in Table 4 and Figure 3. The highlighted numbers below correspond to those in Table 4.

❶ Analysis 3 is an analysis of the part of **Z** that is not predictable from **G**—that is, it is an analysis of $(\mathbf{Z} - {}_{\mathbf{G}}\mathbf{Z}')$, where ${}_{\mathbf{G}}\mathbf{Z}' = \mathbf{GC}$. In the current example, it is an analysis of within-groups variation.

❷ The values in this section show that the within-groups variation is 137.253, which represents 28.358% of the total variation.

❸ This section indicates that when PCA is applied to the within-groups variation $(\mathbf{Z} - \mathbf{GZ}')$, the variance accounted for by a two-component solution is 73.887, which represents 53.832% of the within-groups variation but only 15.266% of the total variation. The dimensionwise SS and %SS give corresponding values separately for the first and second within-groups components.

❹ The loading matrix given in this section and illustrated in Figure 3 shows that the within-groups structure differs substantially from the total structure.

The results for Analysis 4 are shown in Table 5 and Figure 4. The highlighted numbers below correspond to those shown in Table 5.

❶ Analysis 4 is an analysis of the part of **Z** that is predictable from **H**—that is, it is an analysis of $\mathbf{Z}'_{\mathbf{H}}$, where $\mathbf{Z}'_{\mathbf{H}} = \mathbf{BH}'$. We refer to this analysis as redundancy analysis of structured variables. In the current example, it is an analysis of between-item-type variation, in which items are classified into types as described above.

❷ The values in this section show that the between-item-type variation is 400.371, which represents 82.721% of the total variation in **Z**. Again, this percent represents redundancy in **Z**, but with respect to item types rather than patient types.

❸ Matrix **B** is a matrix of parameter estimates for predicting **Z** from **H**. In general, this matrix contains the regression coefficients that would be produced if **Z** were transposed to produce a $p$ (variables) $\times$ $N$ (subjects) matrix, which was then subjected to a multivariate multiple regression on **H**. In the current example, **H** is a matrix of dummy variables coding item types; the values in matrix **B** are the means for each subject on each of the four types of items. For example, items classified as characteristic of depression include guilt, depressive mood, and motor

**Table 7**
**Analysis of Z Regressed on G and H**

ANALYSIS 6: Analysis of P(G)ZP(H) ❶
SS & %SS for the external analysis ❷
SS= 335.453 %SS= 69.309

Matrix **M** in the full model ❸

| | | | | |
|---|---|---|---|---|
| 1 | .177 | 1.435 | −.776 | −.660 |
| 2 | −.999 | −.494 | 1.296 | −.180 |
| 3 | 1.340 | .043 | −.813 | −.281 |
| 4 | −.518 | −.983 | .293 | 1.121 |

SS & %SS for the internal analysis ❹
SS= 313.775 %SS (per term SS)= 93.538 %SS (per total SS)= 64.830

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 239.164 | 74.611 |
| %SS (per term SS) | 71.296 | 22.242 |
| %SS (per total SS) | 49.414 | 15.416 |

Matrix **A** of Component Loadings ❺

| | | |
|---|---|---|
| 1 | .691 | −.472 |
| 2 | .802 | .381 |
| 3 | −.783 | .351 |
| 4 | −.783 | .351 |
| 5 | .802 | .381 |
| 6 | −.486 | −.385 |
| 7 | −.486 | −.385 |
| 8 | .802 | .381 |
| 9 | −.486 | −.385 |
| 10 | .691 | −.472 |
| 11 | −.783 | .351 |

Matrix of correlations between **H** and **A** ❻

| | | |
|---|---|---|
| 1 | .701 | .580 |
| 2 | −.681 | .534 |
| 3 | .464 | −.578 |
| 4 | −.422 | −.614 |

Matrix of correlations between **G** and **F** ❼

| | | |
|---|---|---|
| 1 | .689 | .482 |
| 2 | −.569 | .665 |
| 3 | .454 | −.596 |
| 4 | −.574 | −.551 |

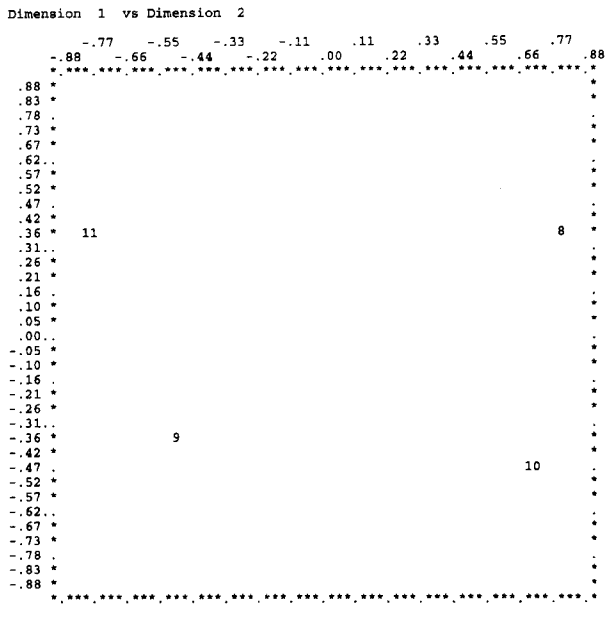


**Figure 6. Plot of loading matrix for Analysis 6.**

retardation; the parameters in **B** associated with this classification would equal the subjects' means on these three items.

❶ This section indicates that when PCA is applied to the between-item-type variation, the variance accounted for by a two-component solution is 350.107, which represents 87.446% of the between-item-type variation and 72.336% of the total variation (i.e., the item-type redundancy is 72.336). The dimensionwise SS and %SS indicate that the first two item-type redundancy variates account for 51.609% and 20.727%, respectively, of the total variance in **Z**.

❺ Matrix **A** shows the loading matrix that arises when PCA is applied to the $\mathbf{Z}'_{\mathbf{H}}$ obtained from $\mathbf{Z}'_{\mathbf{H}} = \mathbf{BH}'$. These loadings are plotted in Figure 4. These results show that dummy coding the items into types has the effect of constraining the loadings for items within a type to be equal. Moreover, the figure clearly indicates that these constraints capture the essence of the structure of **Z** (only one item from each item type appears in the figure because same-type items overlap).

❻ The matrix of correlations between **H** and **A** indicates that the first item-type redundancy variate is defined by the difference between Item Types 1 and 3 (depressed and simple schizophrenic) and Item Types 2 and 4 (manic and paranoid schizophrenic). The second item-type redundancy variate contrasts the combined manic and depressive item types with item types characteristic of schizophrenia (simple and paranoid item types combined).

The results of Analysis 5 are shown in Table 6 and Figure 5. The highlighted numbers below correspond to those shown in Table 6.

❶ Analysis 5 is an analysis of the part of **Z** that is not predictable from **H**—that is, it is an analysis of $(\mathbf{Z} - \mathbf{Z}'_{\mathbf{H}})$,

**Table 8**
**Analysis of the Part of Z That Is Predictable From H But Unrelated to G**

ANALYSIS 7: Analysis of Q(G)ZP(H) ❶
SS & %SS for the external analysis ❷
SS= 64.917 %SS= 13.413

Matrix **B** in the full model ❸

| | | | | |
|---|---|---|---|---|
| 1 | −.215 | .194 | .311 | −.223 |
| 2 | .461 | .031 | −.253 | −.681 |
| 3 | −.456 | −.272 | −.263 | .912 |
| 4 | −.891 | .287 | .202 | .428 |
| 5 | −.300 | −.085 | −.334 | .614 |
| 6 | .096 | .249 | .035 | −.249 |
| 7 | −.203 | −.053 | .443 | .136 |
| 8 | .001 | .229 | −.144 | −.514 |
| 9 | .427 | .067 | −.062 | .214 |
| 10 | .528 | −.438 | .211 | −.099 |
| 11 | .553 | −.208 | −.146 | −.538 |
| 12 | −.038 | .014 | −.126 | .516 |
| 13 | −.008 | −.044 | −.204 | .987 |
| 14 | .102 | −.057 | .189 | .004 |
| 15 | .177 | .162 | .471 | −.026 |
| 16 | .219 | .071 | −.023 | −1.048 |
| 17 | −.175 | .083 | −.040 | −.349 |
| 18 | −.008 | .152 | .067 | −.717 |
| 19 | .102 | .115 | .477 | −.524 |
| 20 | −.008 | −.177 | −.381 | 1.146 |
| 21 | −.188 | −.599 | −.413 | .448 |
| 22 | −.175 | .280 | −.018 | −.437 |
| 23 | .244 | −.334 | .256 | −.307 |
| 24 | −.349 | .565 | −.096 | −.082 |
| 25 | −.081 | −.392 | −.366 | .313 |
| 26 | .548 | −.393 | .046 | −.337 |
| 27 | −.474 | .102 | −.258 | .723 |
| 28 | .084 | −.106 | −.326 | −.019 |
| 29 | −.022 | .445 | .741 | −1.133 |
| 30 | .330 | .038 | −.143 | .254 |
| 31 | −.316 | −.348 | −.107 | .887 |
| 32 | .143 | .403 | −.020 | −.449 |
| 33 | −.106 | .021 | .274 | .152 |
| 34 | .233 | −.011 | −.404 | .431 |
| 35 | −.427 | .174 | .150 | −.051 |
| 36 | −.103 | −.083 | −.194 | −.101 |
| 37 | −.160 | .525 | −.597 | .499 |
| 38 | .122 | .012 | .365 | −.311 |
| 39 | −.001 | −.088 | −.285 | .209 |
| 40 | −.342 | .041 | .610 | −.466 |
| 41 | .038 | −.039 | .097 | −.384 |
| 42 | .284 | −.386 | .205 | −.037 |
| 43 | −.438 | .203 | −.074 | .221 |
| 44 | .795 | −.348 | .127 | −.010 |

SS & %SS for the internal analysis ❹
SS= 53.715 %SS (per term SS)= 82.744 %SS (per total SS)= 11.098

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 41.157 | 12.558 |
| %SS (per term SS) | 63.400 | 19.344 |
| %SS (per total SS) | 8.504 | 2.595 |

Matrix **A** of Component Loadings ❺

| | | |
|---|---|---|
| 1 | .120 | −.270 |
| 2 | .081 | .205 |
| 3 | .187 | .066 |
| 4 | .187 | .066 |
| 5 | .081 | .205 |
| 6 | −.511 | .015 |
| 7 | −.511 | .015 |
| 8 | .081 | .205 |
| 9 | −.511 | .015 |
| 10 | .120 | −.270 |
| 11 | .187 | .066 |

Matrix of correlations between **H** and **A** ❻

| | | |
|---|---|---|
| 1 | .266 | .680 |
| 2 | .492 | .144 |
| 3 | .268 | −.888 |
| 4 | −.991 | −.055 |

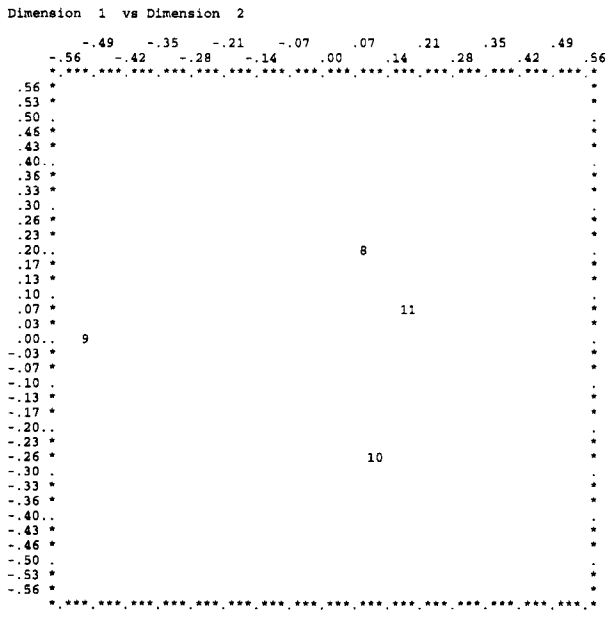


**Figure 7. Plot of loading matrix for Analysis 7.**

where $\mathbf{Z}'_{\mathbf{H}} = \mathbf{BH}'$. In the current example, it is an analysis of within-item-type variation.

❷ The values in this section show that the variation within item types is 83.629, which is only 17.279% of the total variation in **Z**.

❸ This section indicates that when PCA is applied to the within-item-type variation ($\mathbf{Z} - \mathbf{Z}'_{\mathbf{H}}$), the variance accounted for by a two-component solution is 47.968, which represents 57.358% of the within-item-type variation, but only 9.911% of the total variation. The dimensionwise SS and %SS give corresponding values separately for the first and second within-item-type components.

❹ The loading matrix given in this section and illustrated in Figure 5 shows that the within-item-type structure differs substantially from the total structure.

The results for Analysis 6 are shown in Table 7 and Figure 6. The highlighted numbers below correspond to those shown in Table 7.

❶ Analysis 6 analyzes the part of **Z** that is predictable from both **G** and **H**. It is an analysis of ${}_{\mathbf{G}}\mathbf{Z}'_{\mathbf{H}}$, where ${}_{\mathbf{G}}\mathbf{Z}'_{\mathbf{H}} = \mathbf{GMH}'$. We refer to this analysis as double redundancy analysis. In the current example, it is an analysis of between-groups variation in between-item-type variation (i.e., it evaluates whether the item-type profiles vary over groups).

❷ The values in this section show that combined, **G** and **H** account for 69.309% of the total variation in **Z**. This percent represents redundancy in **Z** as a function of patient-type variation in item types.

❸ Matrix **M** is a matrix of parameter estimates for predicting **Z** from **G** and **H**. In the current example, because both **G** and **H** are matrices of dummy variables, the values in matrix **M** are the means for each of the four patient types on each of the four item types.

❹ This section indicates that when PCA is applied to the between-patient-type variation in item-type variation, the variance accounted for by a two-component solution is 313.775, which represents 93.538% of the explained variation and 64.83% of the total variation. The dimensionwise SS and %SS give corresponding values separately for each of the components.

❺ Matrix **A** shows the loading matrix that arises when PCA is applied to the ${}_{\mathbf{G}}\mathbf{Z}'_{\mathbf{H}}$ obtained from ${}_{\mathbf{G}}\mathbf{Z}'_{\mathbf{H}} = \mathbf{GMH}'$. These loadings are plotted in Figure 6. Again, these results show that the structure of the explained variation captures the essence of the total variation in **Z**.

❻ and ❼ These sections show the matrix of correlations between H and A and between G and F, respectively. These matrices have have the same interpretation as redundancy variates, as described previously. In this case, they indicate that the first overall redundancy variate is defined by

**Table 9**
**Analysis of the Part of Z That Is Predictable From G But Unrelated to H**

ANALYSIS 8: Analysis of P(G)ZQ(H) ❶
SS & %SS for the external analysis ❷
SS= 11.293 %SS= 2.333

Matrix **C** (transposed) in the full model ❸

| | | | | |
|---|---|---|---|---|
| 1 | .245 | −.268 | −.138 | .161 |
| 2 | .049 | .096 | −.195 | .050 |
| 3 | .168 | −.151 | .161 | −.178 |
| 4 | −.209 | −.003 | −.081 | .294 |
| 5 | .031 | −.043 | −.057 | .069 |
| 6 | −.052 | .023 | −.217 | .246 |
| 7 | −.007 | .055 | −.025 | −.023 |
| 8 | −.080 | −.052 | .252 | −.119 |
| 9 | .060 | −.078 | .242 | −.223 |
| 10 | −.245 | .268 | .138 | −.161 |
| 11 | .041 | .155 | −.080 | −.116 |

SS & %SS for the internal analysis ❹
SS= 9.909 %SS (per term SS)= 87.750 %SS (per total SS)= 2.047

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 5.962 | 3.947 |
| %SS (per term SS) | 52.799 | 34.951 |
| %SS (per total SS) | 1.232 | .815 |

Matrix **A** of Component Loadings ❺

| | | |
|---|---|---|
| 1 | .123 | −.171 |
| 2 | .081 | .025 |
| 3 | −.109 | −.124 |
| 4 | .130 | .084 |
| 5 | .047 | −.022 |
| 6 | .161 | .042 |
| 7 | −.002 | .022 |
| 8 | −.128 | −.003 |
| 9 | −.159 | −.064 |
| 10 | −.123 | .171 |
| 11 | −.021 | .03 |

Matrix of correlations between **G** and **F** ❻

| | | |
|---|---|---|
| 1 | .050 | −.810 |
| 2 | −.105 | .815 |
| 3 | −.784 | −.080 |
| 4 | .839 | .075 |

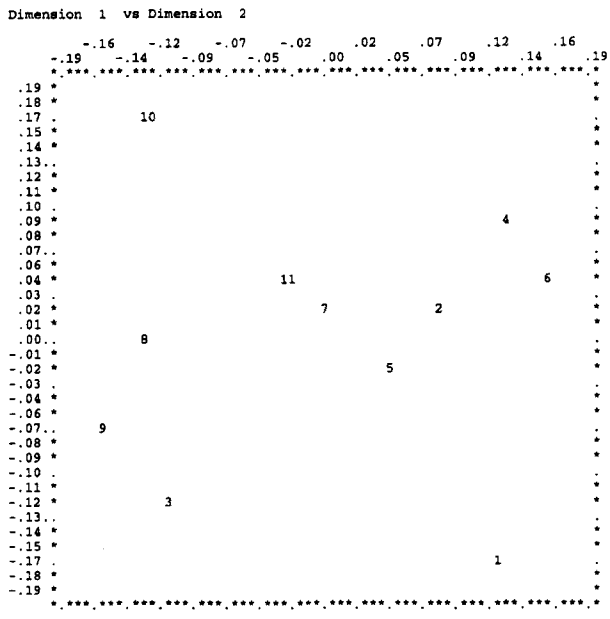


**Figure 8. Plot of loading matrix for Analysis 8.**

**Table 10**
**Analysis of Residuals After Z Has Been Regressed on G and H**

ANALYSIS 9: Analysis of Q(G)ZQ(H) ❶
SS & %SS for the external analysis ❷
SS= 72.336 %SS= 14.946
SS & %SS for the internal analysis ❸
SS= 42.139 %SS (per term SS)= 58.255 %SS (per total SS)= 8.706

| Dimensionwise SS & %SS | 1 | 2 |
|---|---|---|
| SS= | 26.916 | 15.224 |
| %SS (per term SS) | 37.209 | 21.045 |
| %SS (per total SS) | 5.561 | 3.145 |

Matrix A of Component Loadings ❹

| | | |
|---|---|---|
| 1 | .006 | −.159 |
| 2 | .091 | −.038 |
| 3 | −.329 | −.010 |
| 4 | .171 | −.039 |
| 5 | −.144 | −.015 |
| 6 | −.098 | −.429 |
| 7 | .498 | .140 |
| 8 | .052 | .053 |
| 9 | −.400 | .289 |
| 10 | −.006 | .159 |
| 11 | .159 | .049 |

the difference between patients who are depressed or simple schizophrenic and patients who are manic or paranoid schizophrenic in the item types hypothesized to be symptomatic of such patients. The second redundancy variate captures differences between schizophrenics overall and manic depressives overall in corresponding item types.

The results for Analysis 7 are shown in Table 8 and Figure 7. The highlighted numbers below correspond to those shown in Table 8.

❶ Analysis 7 is an analysis of the part of **Z** that is uniquely predictable from **H**, independent of **G**. It analyzes $(\mathbf{Z}'_{\mathbf{H}} - {}_{\mathbf{G}}\mathbf{Z}'_{\mathbf{H}})$ and can be conceptualized as regressing **Z** on **G**, obtaining the residuals from that analysis, and then applying item-type constraints to those residuals. In the current example, it is an analysis of within-patient-type variation in between-item-type variation.

❷, ❸, ❹, ❺, and ❻ These sections and Figure 7 report results identical in interpretation to similar results reported previously, with the proviso that they pertain to between-item-type variation after between-patient-type variation has been partialed out. Overall, the results indicate there is very little variation in **Z** that is uniquely attributable to **H**, and that the structure of this variation is quite different from the structure of the total variation. There is, however, some indication that, independent of group membership, items classified as characteristic of paranoid schizophrenics are responded to differently than other item types.

The results of Analysis 8 are shown in Table 9 and Figure 8. The highlighted numbers below correspond to those shown in Table 9.

❶ Analysis 8 is an analysis of the part of **Z** that is uniquely predictable from **G**, independent of **H**. It analyzes $({}_{\mathbf{G}}\mathbf{Z}' - {}_{\mathbf{G}}\mathbf{Z}'_{\mathbf{H}})$ and can be conceived as regressing **Z** (transposed) on **H**, obtaining the residuals from that analysis, and then applying subject constraints to those residuals. In the current example, it is an analysis of between-patient-type variation in within-item-type variation.

❷, ❸, ❹, ❺, and ❻ These sections and Figure 8 show results identical in interpretation to similar results reported previously, this time with the proviso that they pertain to between-patient-type variation after between-item-type variation has been partialed out. The results indicate no independent effect of patient types.

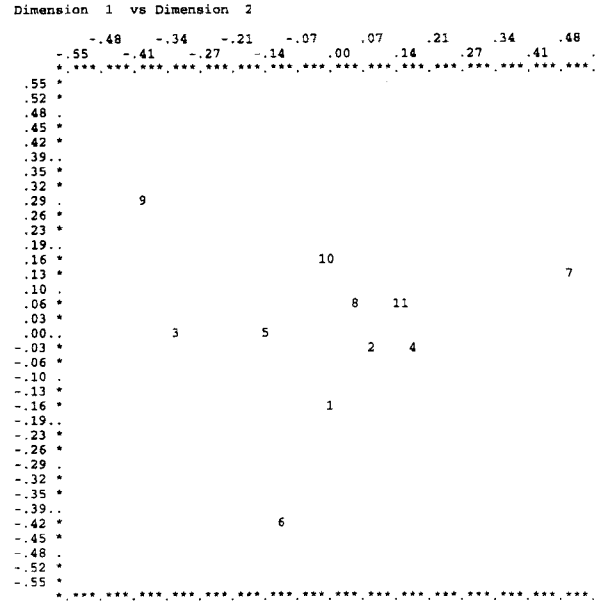


**Figure 9. Plot of loading matrix for Analysis 9.**

The results of Analysis 9 are shown in Table 10 and Figure 9. The highlighted numbers below correspond to those shown in Table 10.

❶ Analysis 9 is an analysis of the part of **Z** that is not predictable from **G** and **H**. In the current example, it is an analysis of within-patient-type variation in within-item-type variation.

❷, ❸, and ❹ These sections and Figure 9 show that very little variation in **Z** remains that is independent of **G** and **H**, and that the structure of the overall residuals does not resemble the total structure at all.

## REFERENCES


ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, **22**, 327-351.

CARROLL, J. D., PRUZANSKY, S., & KRUSKAL, J. B. (1980). CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika*, **45**, 3-24.

DESARBO, W. S., & RAO, V. R. (1984). GENFOLD2: A set of models and algorithms for the GENeral UnFOLDing analysis of preference/dominance data. *Journal of Classification*, **1**, 147-186.

DE SOETE, G., & CARROLL, J. D. (1983). A maximum likelihood method for fitting the wandering vector model. *Psychometrika*, **48**, 553-566.

HEISER, W. J., & DE LEEUW, J. (1981). Multidimensional mapping of preference data. *Mathematique et sciences humaines*, **19**, 39-96.

KHATRI, C. G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of the Institute of Statistical Mathematics*, **18**, 75-86.

RAO, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya*, **26**, 329-358.

RAO, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447-458.

TAKANE, Y., & SHIBAYAMA, T. (1988). Three vector models of pairwise preference ratings and their generalizations. In S. Kashiwagi (Ed.), *Proceedings of the 16th Annual Meeting of the Behaviormetric Society* (pp. 131-132). Tokyo: Behaviormetric Society of Japan.

TAKANE, Y., & SHIBAYAMA, T. (1991). Principal component analysis with external information on both subjects and variables. *Psychometrika*, **56**, 97-120.

VAN DEN WOLLENBERG, A. L. (1977). Redundancy analysis: An alternative for canonical correlation analysis. *Psychometrika*, **42**, 207-219.


(Manuscript received June 13, 1996;
revision accepted for publication January 22, 1997.)