# CATPCA

The CATPCA procedure quantifies categorical variables using optimal scaling, resulting in optimal principal components for the transformed variables. The variables can be given mixed optimal scaling levels and no distributional assumptions about the variables are made.

In CATPCA, dimensions correspond to components (that is, an analysis with two dimensions results in two components), and object scores correspond to component scores.

## Notation

The following notation is used throughout this chapter unless otherwise stated:

| | |
|---|---|
| $n$ | Number of analysis cases (objects) |
| $n_w$ | Weighted number of analysis cases: $\sum_{i=1}^{n} w_i$ |
| $n_{tot}$ | Total number of cases (analysis + supplementary) |
| $w_i$ | Weight of object $i$; $w_i = 1$ if cases are unweighted; $w_i = 0$ if object $i$ is supplementary. |
| $\mathbf{W}$ | Diagonal $n_{tot} \times n_{tot}$ matrix, with $w_i$ on the diagonal. |
| $m$ | Number of analysis variables |
| $m_w$ | Weighted umber of analysis variables ( $m_w = \sum_{j=1}^{m} v_j$ ) |
| $m_{tot}$ | Total number of variables (analysis + supplementary) |
| $m_1$ | Number of analysis variables with multiple nominal scaling level. |
| $m_2$ | Number of analysis variables with non-multiple scaling level. |
| $m_{w1}$ | Weighted number of analysis variables with multiple nominal scaling level. |
| $m_{w2}$ | Weighted number of analysis variables with non-multiple scaling level. |

| | |
|---|---|
| $J$ | Index set recording which variables have multiple nominal scaling level. |
| $\mathbf{H}$ | The data matrix (category indicators), of order $n_{tot} \times m_{tot}$, after discretization, imputation of missings , and listwise deletion, if applicable. |
| $p$ | Number of dimensions |

For variable $j$, $j = 1, \ldots, m_{tot}$

| | |
|---|---|
| $v_j$ | variable weight; $v_j = 1$ if weight for variable $j$ is not specified or if variable $j$ is supplementary |
| $k_j$ | Number of categories of variable $j$ (number of distinct values in $\mathbf{h}_j$, thus, including supplementary objects) |
| $\mathbf{G}_j$ | Indicator matrix for variable $j$, of order $n_{tot} \times k_j$ |

The elements of $\mathbf{G}_j$ are defined as $i = 1, \ldots, n_{tot}$; $r = 1, \ldots, k_j$

$$g_{(j)ir} = \begin{cases} 1 & \text{when the } i\text{th object is in the } r\text{th category of variable } j \\ 0 & \text{when the } i\text{th object is not in the } r\text{th category of variable } j \end{cases}$$

| | |
|---|---|
| $\mathbf{D}_j$ | Diagonal $k_j \times k_j$ matrix, containing the weighted univariate marginals; i.e., the weighted column sums of $\mathbf{G}_j$ ($\mathbf{D}_j = \mathbf{G}_j' \mathbf{W} \mathbf{G}_j$) |
| $\mathbf{M}_j$ | Diagonal $n_{tot} \times n_{tot}$ matrix, with diagonal elements defined as |

$$m_{(j)ii} = \begin{cases} 0 & \text{when the } i\text{th observation is missing and missing strategy variable } j \text{ is passive} \\ 0 & \text{when the } i\text{th object is in } r\text{th category of variable } j \text{ and } r\text{th category is only used by supplementary objects (i.e. when } \mathbf{D}_{(j)rr} = 0) \\ v_j & \text{otherwise} \end{cases}$$

| | |
|---|---|
| $\mathbf{M}_*$ | $\sum_j \mathbf{M}_j$ |
| $\mathbf{S}_j$ | I-spline basis for variable $j$, of order $k_j \times (s_j + t_j)$ (see Ramsay (1988) for details) |
| $\mathbf{b}_j$ | Spline coefficient vector, of order $s_j + t_j$ |
| $d_j$ | Spline intercept. |
| $s_j$ | Degree of polynomial |
| $t_j$ | Number of interior knots |

The quantification matrices and parameter vectors are:

| | |
|---|---|
| $\mathbf{X}$ | Object scores, of order $n_{tot} \times p$ |
| $\mathbf{X}_w$ | Weighted object scores ( $\mathbf{X}_w = \mathbf{WX}$ ) |
| $\mathbf{X}_\mathrm{n}$ | $\mathbf{X}$ normalized according to requested normalization option |
| $\mathbf{Y}_j$ | Centroid coordinates, of order $k_j \times p$. For variables with optimal scaling level multiple nominal, this are the category quantifications |
| $\mathbf{y}_j$ | Category quantifications for variables with non-multiple scaling level, of order $k_j$ |
| $\mathbf{a}_j$ | Component loadings for variables with non-multiple scaling level, of order $p$ |
| $\mathbf{a}_{\mathrm{n}_j}$ | $\mathbf{a}_j$ normalized according to requested normalization option |
| $\underline{\mathbf{Y}}$ | Collection of category quantifications (centroid coordinates) for variables with multiple nominal scaling level ( $\mathbf{Y}_j$ ), and vector coordinates for non-multiple scaling level ( $\mathbf{y}_j\mathbf{a}_j'$ ). |

*Note:* The matrices $\mathbf{W}$, $\mathbf{G}_j$, $\mathbf{M}_j$, $\mathbf{M}_*$, and $\mathbf{D}_j$ are exclusively notational devices; they are stored in reduced form, and the program fully profits from their sparseness by replacing matrix multiplications with selective accumulation.

# Discretization

Discretization is done on the unweighted data.

**Multiplying** First, the orginal variable is standardized. Then the standardized values are multiplied by 10 and rounded, and a value is added such that the lowest value is 1.

**Ranking** The original variable is ranked in ascending order, according to the alpanumerical value.

### Grouping into a specified number of categories with a normal distribution

First, the original variable is standardized. Then cases are assigned to categories using intervals as defined in Max (1960).

### Grouping into a specified number of categories with a unifrom distribution

First the target frequency is computed as $n$ divided by the number of specified categories, rounded. Then the original categories are assigned to grouped categories such that the frequencies of the grouped categories are as close to the target frequency as possible.

### Grouping equal intervals of specified size

First the intervals are defined as lowest value + interval size, lowest value + 2*interval size, etc. Then cases with values in the $k^{\text{th}}$ interval are assigned to category $k$.

# Imputation of Missing Values

When there are variables with missing values specified to be treated as active (impute mode or extra category), then first the $k_j$'s for these variables are computed before listwise deletion. Next the category indicator with the highest weighted frequency (mode; the smallest if multiple modes exist), or $k_j + 1$ (extra category) is imputed. Then listwise deletion is applied if applicable. And then the $k_j$'s are adjusted.

If an extra category is imputed for a variable with optimal scaling level Spline Nominal, Spline Ordinal, Ordinal or Numerical, the extra category is not included in the restriction according to the scaling level in the final phase (see step (2) next section).

# Configuration

CATPCA can read a configuration from a file, to be used as the initial configuration or as a fixed configuration in which to fit variables.

For an initial configuration see step 1 in the Optimization section.

A fixed configuration $\mathbf{X}$ is centered and orthonormalized as described in the optimization section in step 3 (with $\mathbf{X}$ in stead of $\mathbf{Z}$) and step 4 (except for the factor $n_w^{1/2}$), and the result is postmultiplied with $\mathbf{\Lambda}^{1/2}$ (this leaves the configuration unchanged if it is already centered and orthogonal). The analysis variables are set to supplementary and variable weights are set to one. Then CATPCA proceeds as described in the Supplementary Variables section.

# Objective Function Optimization

## Objective Function

The CATPCA objective is to find object scores $\mathbf{X}$ and a set of $\underline{\mathbf{Y}}_j$ (for $j = 1, \ldots, m$) — the underlining indicates that they may be restricted in various ways — so that the function

$$\sigma\left(\mathbf{X};\underline{\mathbf{Y}}\right) = n_w^{-1}\sum_j c^{-1}\mathrm{tr}\left(\left(\mathbf{X}-\mathbf{G}_j\underline{\mathbf{Y}}_j\right)'\mathbf{M}_j\mathbf{W}\left(\mathbf{X}-\mathbf{G}_j\underline{\mathbf{Y}}_j\right)\right), \text{ with}$$

$c$ is $p$ if $j \in J$ and $c$ is 1 if $j \notin J$,

is minimal, under the normalization restriction $\mathbf{X}'\mathbf{M}_*\mathbf{W}\mathbf{X} = n_w m_w \mathbf{I}$ ($\mathbf{I}$ is the $p \times p$ identity matrix). The inclusion of $\mathbf{M}_j$ in $\sigma\left(\mathbf{X};\underline{\mathbf{Y}}\right)$ ensures that there is no influence of passive missing values (missing values in variables that have missing option passive, or missing option not specified ). $\mathbf{M}_*$ contains the number of active data values for each object. The object scores are also centered; i.e. they satisfy $\mathbf{u}'\mathbf{M}_*\mathbf{W}\mathbf{X} = \mathbf{0}$ with $\mathbf{u}$ denoting an $n$-vector with ones.

## Optimal Scaling Levels

The following optimal scaling levels are distinguished in CATPCA:

**Multiple Nominal** $\underline{\mathbf{Y}}_j = \mathbf{Y}_j$ (equality restriction only).

**Nominal** $\underline{\mathbf{Y}}_j = \mathbf{y}_j \mathbf{a}'_j$ (equality and rank - one restrictions).

**Spline Nominal** $\underline{\mathbf{Y}}_j = \mathbf{y}_j \mathbf{a}'_j$ and $\mathbf{y}_j = d_j + \mathbf{S}_j \mathbf{b}_j$ (equality, rank – one, and spline restrictions).

**Spline Ordinal** $\underline{\mathbf{Y}}_j = \mathbf{y}_j \mathbf{a}'_j$ and $\mathbf{y}_j = d_j + \mathbf{S}_j \mathbf{b}_j$ (equality, rank – one, and monotonic spline restrictions),

with $\mathbf{b}_j$ restricted to contain nonnegative elements (to garantee monotonic I-splines).

**Ordinal** $\underline{\mathbf{Y}}_j = \mathbf{y}_j \mathbf{a}'_j$ and $\mathbf{y}_j \in \mathbf{C}_j$ (equality, rank – one, and monotonicity restrictions).

The monotonicity restriction $\mathbf{y}_j \in \mathbf{C}_j$ means that $\mathbf{y}_j$ must be located in the convex cone of all $k_j$ -vectors with nondecreasing elements.

**Numerical** $\underline{\mathbf{Y}}_j = \mathbf{y}_j \mathbf{a}'_j$ and $\mathbf{y}_j \in \mathbf{L}_j$ (equality, rank – one, and linearity restrictions).

The linearity restriction $\mathbf{y}_j \in L_j$ means that $\mathbf{y}_j$ must be located in the subspace of all $k_j$ -vectors that are a linear transformation of the vector consisting of $k_j$ successive integers.

For each variable, these levels can be chosen independently. The general requirement for all options is that equal category indicators receive equal quantifications. The general requirement for the non-multiple options is $\underline{\mathbf{Y}}_j = \mathbf{y}_j \mathbf{a}'_j$; that is, $\underline{\mathbf{Y}}_j$ is of rank one; for identification purposes, $\mathbf{y}_j$ is always normalized so that $\mathbf{y}'_j \mathbf{D}_j \mathbf{y}_j = n_w$.

## Optimization

Optimization is achieved by executing the following iteration scheme:

1. Initialization I or II
2. Update category quantifications
3. Update object scores
4. Orthonormalization
5. Convergence test: repeat (2)—(4) or continue
6. Rotation and reflection

The first time (for the initial configuration) initialization I is used and variables that do not have optimal scaling level Multiple Nominal or Numerical are temporarily treated as numerical, the second time (for the final configuration) initialization II is used . Steps (1) through (5) are explained below.

**(1) Initialization**

I. If a fixed configuration is not specified, the object scores $\mathbf{X}$ are initialized with random numbers. Then $\mathbf{X}$ is normalized so that $\mathbf{u}'\mathbf{M}_*\mathbf{W}\mathbf{X} = \mathbf{0}$ and $\mathbf{X}'\mathbf{M}_*\mathbf{W}\mathbf{X} = n_w m_w \mathbf{I}$, yielding $\tilde{\mathbf{X}}$. The initial component loadings are computed as the cross products of $\tilde{\mathbf{X}}$ and the centered original variables $\left(\mathbf{I} - \mathbf{M}_j\mathbf{u}\mathbf{u}'\mathbf{W}/(\mathbf{u}'\mathbf{M}_j\mathbf{W}\mathbf{u})\right)\mathbf{h}_j$, rescaled to unit length.

II.

All relevant quantities are copied from the results of the first cycle.

**(2) Update category quantifications; loop across variables $j = 1, \ldots, m$ ( variables $1, \ldots, m$ are analysis variables):**

With fixed current values $\mathbf{X}_w^+$ the unconstrained update of $\mathbf{Y}_j$ is

$$\tilde{\mathbf{Y}}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{X}_w^+$$

<u>Multiple nominal</u>: $\mathbf{Y}_j^+ = \tilde{\mathbf{Y}}_j$ .

For non-multiple scaling levels first an unconstrained update is computed in the same way:

$$\tilde{\mathbf{Y}}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{X}_w^+$$

next one cycle of an ALS algorithm (De Leeuw et al., 1976) is executed for computing a rank-one decomposition of $\tilde{\mathbf{Y}}_j$, with restrictions on the left-hand vector, resulting in

$$\tilde{\mathbf{y}}_j = \tilde{\mathbf{Y}}_j \mathbf{a}_j$$

<u>Nominal</u>: $\mathbf{y}_j^* = \tilde{\mathbf{y}}_j$.

For the next four optimal scaling levels, if variable *j* was imputed with an extra category, $\mathbf{y}_j^*$ is inclusive category $k_j$ in the initial phase, and is exclusive category $k_j$ in the final phase.

<u>Spline nominal and spline ordinal</u>: $\mathbf{y}_j^* = d_j + \mathbf{S}_j \mathbf{b}_j$.

The spline transformation is computed as a weighted regression (with weights the diagonal elements of $\mathbf{D}_j$) of $\tilde{\mathbf{y}}_j$ on the I-spline basis $\mathbf{S}_j$. For the spline ordinal scaling level the elements of $\mathbf{b}_j$ are restricted to be nonnegative, which makes $\mathbf{y}_j^*$ monotonically increasing

<u>Ordinal</u>: $\mathbf{y}_j^* \leftarrow WMON(\tilde{\mathbf{y}}_j)$.

The notation $WMON(\ )$ is used to denote the weighted monotonic regression process, which makes $\mathbf{y}_j^*$ monotonically increasing. The weights used are the diagonal elements of $\mathbf{D}_j$ and the subalgorithm used is the up-and-down-blocks minimum violators algorithm (Kruskal, 1964; Barlow et al., 1972).

<u>Numerical</u>: $\mathbf{y}_j^* \leftarrow WLIN(\tilde{\mathbf{y}}_j)$.

The notation $WLIN(\ )$ is used to denote the weighted linear regression process. The weights used are the diagonal elements of $\mathbf{D}_j$.

Next $\mathbf{y}_j^*$ is normalized (if variable *j* was imputed with an extra category, $\mathbf{y}_j^*$ is inclusive category $k_j$ from here on):

$$\mathbf{y}_j^+ = n_w^{1/2} \mathbf{y}_j^* (\mathbf{y}_j^{*\prime} \mathbf{D}_j \mathbf{y}_j^*)^{-1/2}$$

Then we update the component loadings:

$$\mathbf{a}_j^+ = n_w^{-1}\mathbf{Y}'_j\mathbf{D}_j\mathbf{y}_j^+$$

Finally, we set $\underline{\mathbf{Y}}_j^+ = \mathbf{y}_j^+\mathbf{a}_j'^+$.

**(3) Update object scores**

Firs the auxiliary score matrix $\mathbf{Z}$ is computed as

$$\mathbf{Z} \leftarrow \sum_j \mathbf{M}_j\mathbf{G}_j\underline{\mathbf{Y}}_j^+$$

and centered with respect to $\mathbf{W}$ and $\mathbf{M}_*$:

$$\mathbf{X}^* = \left(\mathbf{I} - \mathbf{M}_*\mathbf{u}\mathbf{u}'\mathbf{W}/(\mathbf{u}'\mathbf{M}_*\mathbf{W}\mathbf{u})\right)\mathbf{Z}$$

These two steps yield locally the best updates when there would be no orthogonality constraints.

**(4) Orthonormalization**

To find an $\mathbf{M}_*$-orthonormal $\mathbf{X}^+$ that is closest to $\mathbf{X}^*$ in the least squares sense, we use the singular value decomposition $m_w^{1/2}\mathbf{M}_*^{-1/2}\mathbf{W}^{1/2}\mathbf{X}^* = \mathbf{K}\mathbf{\Lambda}^{1/2}\mathbf{L}'$, then $n_w^{1/2}m_w^{1/2}\mathbf{M}_*^{-1/2}\mathbf{W}^{1/2}\mathbf{K}\mathbf{L}'$ yields $\mathbf{M}_*$-orthonormal weighted object scores:

$$\mathbf{X}_w^+ \leftarrow n_w^{1/2}m_w\mathbf{M}_*^{-1}\mathbf{W}\mathbf{X}^*\mathbf{L}\mathbf{\Lambda}^{-1/2}\mathbf{L}', \text{ and } \mathbf{X}^+ = \mathbf{W}^{-1}\mathbf{X}_w^+.$$

The calculation of $\mathbf{L}$ and $\mathbf{\Lambda}$ is based on tridiagonalization with Householder transformations followed by the implicit QL algorithm (Wilkinson, 1965).

**(5) Convergence test**

The difference between consecutive values of the quantity

$$\textbf{TFIT} = (pn_w)^{-1}\sum_{j\in J} v_j \operatorname{tr}\left(\mathbf{Y}_j'\mathbf{D}_j\mathbf{Y}_j\right) + \sum_{j\notin J} v_j\mathbf{a}_j'\mathbf{a}_j,$$

is compared with the user-specified convergence criterion $\varepsilon$ - a small positive number. It can be shown that TFIT = $m_{w1} + pm_{w2} - \sigma\left(\mathbf{X};\underline{\mathbf{Y}}\right)$. Steps (2) through (4) are repeated as long as the loss difference exceeds $\varepsilon$.

After convergence TFIT is also equal to $\mathrm{tr}(\mathbf{\Lambda}^{1/2})$, with $\mathbf{\Lambda}$ as computed in step (4) during the last iteration. (See also Model Summary, and Correlations Transformed Variables for interpretation of $\mathbf{\Lambda}^{1/2}$).

**(6) Rotation and reflection**

To achieve principal axes orientation, $\mathbf{X}^{+}$ is rotated with the matrix $\mathbf{L}'$. In addition the $s^{\text{th}}$ column of $\mathbf{X}^{+}$ is reflected if for dimension $s$ the mean of squared loadings with a negative sign is higher than the mean of squared loadings with a positive sign. Then step (2) is executed, yielding the rotated and possibly reflected quantifications and loadings.

# Supplementary Objects

To compute the object scores for supplementary objects, after convergence steps (2) and (3) are repeated, with the zero's in $\mathbf{W}$ temporarily set to ones in computing $\mathbf{Z}$ and $\mathbf{X}^{+}$. If a supplementary object has missing values, passive treatment is applied.

# Supplementary Variables

The quantifications for supplementary variables are computed after convergence. For supplementary variables with multiple nominal scaling level step (2) is executed once. For non-multiple supplementary variables, an initial $a_j$ is computed as in step (1). Then the rank-one and restriction substeps of step (2) are repeated as long as the difference between consecutive values of $\mathbf{a}_j'\mathbf{a}_j$ exceeds .00001, with a maximum of 100 iterations.

# Diagnostics

## Maximum Rank (may be issued as a warning when exceeded)

The maximum rank $p_{\max}$ indicates the maximum number of dimensions that can be computed for any data set. In general

$$p_{\max} = \min\left( n-1, \left( \sum_{j \in J} k_j \right) - m_1 + m_2 \right)$$

if there are variables with optimal scaling level multiple nominal without missing values to be treated as passive. If variables with optimal scaling level multiple nominal do have missing values to be treated as passive, the maximum rank is

$$p_{\max} = \min\left( n-1, \left( \sum_{j \in J} k_j \right) + m_2 \right)$$

Here $k_j$ is exclusive supplementary objects (that is, a category only used by supplementary objects is not counted in computing the maximum rank). Although the number of nontrivial dimensions may be less than $p_{\max}$ when $m = 2$, CATPCA does allow dimensionalities all the way up to $p_{\max}$. When, due to empty categories in the actual data, the rank detoriates below the specified dimensionality, the program stops.

## Descriptives

The descriptives tables gives the weighted univariate marginals and the weighted number of missing values (system missing, user defined missing, and values $\leq 0$) for each variable.

## Fit and Loss Measures

When the HISTORY option is in effect, the following fit and loss measures are reported:

(a) Total fit. This is the quantity TFIT as defined in step (5).

(b) Total loss. This is $\sigma\left(\mathbf{X}; \underline{\mathbf{Y}}\right)$, computed as the sum of multiple loss and single loss defined below.

(c) Multiple loss. This measure is computed as

$$\text{TMLOSS} = (m_{w1} + p m_{w2}) - \left( (n_w p)^{-1} \sum_{j \in J} v_j \ \text{tr}(\mathbf{Y}_j' \mathbf{D}_j \mathbf{Y}_j) + n_w^{-1} \sum_{j \notin J} v_j \ \text{tr}(\mathbf{Y}_j' \mathbf{D}_j \mathbf{Y}_j) \right)$$

(d) Single loss. This measure is computed only when some of the variables are single:

$$\text{SLOSS} = n_w^{-1} \sum_{j \notin J} v_j \ \text{tr}(\mathbf{Y}_j' \mathbf{D}_j \mathbf{Y}_j) - \sum_{j \notin J} v_j \mathbf{a}_j' \mathbf{a}_j$$

## Model Summary

### Cronbach's Alpha

Cronbach's Alpha per dimension ( $s = 1, \ldots, p$ ):

$$\alpha_s = m_w (\lambda_s^{1/2} - 1) / (\lambda_s^{1/2} (m_w - 1)),$$

Total Cronbach's Alpha is

$$\alpha = m_w \left( \sum_s \lambda_s^{1/2} - 1 \right) \Big/ \sum_s \lambda_s^{1/2} (m_w - 1)$$

with $\lambda_s$ the $s^{th}$ diagonal element of $\mathbf{\Lambda}$ as computed in step (4) during the last iteration.

### Variance Accounted For

Variance Accounted For per dimension ( $s = 1, \ldots, p$ ):

#### Multiple Nominal variables

$$\text{VAF1}_s = n_w^{-1} \sum_{j \in J} v_j \ \text{tr}(\mathbf{Y}_{js}' \mathbf{D}_j \mathbf{Y}_{js}) \text{, (\% of variance is } \text{VAF1}_s \times 100 / m_{w1} \text{),}$$

#### Non-Multiple variables

$$\mathrm{VAF2}_s = \sum_{j \notin J} v_j a_{js}^2 \text{ , (\% of variance is } \mathrm{VAF2}_s \times 100 / m_{w2} \text{ ).}$$

**Eigenvalue**

Eigenvalue per dimension:

$$\lambda_s^{1/2} = \mathrm{VAF1}_s + \mathrm{VAF2}_s \text{ ,}$$

with $\lambda_s$ the $s^{th}$ diagonal element of $\mathbf{\Lambda}$ as computed in step (4) during the last iteration. (See also Optimization step (5), and Correlations Transformed Variables for interpretation of $\mathbf{\Lambda}^{1/2}$).

The Total Variance Accounted For for multiple nominal variables is the mean over dimensions, and for non-multiple variables the sum over dimensions. So, the total eigenvalue is

$$\mathrm{tr}(\mathbf{\Lambda}^{1/2}) = p^{-1} \sum_s \mathrm{VAF1}_s + \sum_s \mathrm{VAF2}_s \text{ .}$$

If there are no passive missing values, the eigenvalues are those of the correlation matrix (see the Correlations and Eigenvalues section) weighted with variable weights:

$$\mathbf{R}_{jj}^{\mathrm{w}} = v_j \mathbf{R}_{jj} \text{ , and } \mathbf{R}_{jl}^{\mathrm{w}} = \mathbf{R}_{lj}^{\mathrm{w}} = v_j^{1/2} \mathbf{R}_{jl}$$

If there are passive missing values, then the eigenvalues are those of the matrix $m_w \mathbf{Q}_{\mathrm{c}}' \mathbf{M}_*^{-1} \mathbf{Q}_{\mathrm{c}}$, with

$$\mathbf{Q}_c = n_w^{-1/2} \left( \mathbf{I} - \mathbf{M}_* \mathbf{u}\mathbf{u}'\mathbf{W} / (\mathbf{u}'\mathbf{M}_* \mathbf{W}\mathbf{u}) \right) \mathbf{Q} \text{ ,}$$

(for $\mathbf{Q}$ see the Correlations and Eigenvalues section) which is not necessarily a correlation matrix, although it is positive semi-definite. This matrix is weighted with variable weights in the same way as $\mathbf{R}$ .

## Variance Accounted For

The Variance Accounted For table gives the VAF per dimension and per variable for centroid coordinates, and for non-multiple variables also for vector coordinates (see quantification section):

### Centroid Coordinates

$$\mathrm{VAF}_{js} = v_j \ \mathrm{tr}(\mathbf{Y}'_{js}\mathbf{D}_j\mathbf{Y}_{js}) \,.$$

### Vector Coordinates

$$\mathrm{VAF}_{js} = v_j a_{js}^2 \text{, for } \ j \notin J \,.$$

## Correlations and Eigenvalues

### Before transformation

$\mathbf{R} = n_w^{-1}\mathbf{H}'_{\mathrm{c}}\mathbf{W}\mathbf{H}_{\mathrm{c}}$, with $\mathbf{H}_{\mathrm{c}}$ weighted centered and normalized $\mathbf{H}$.

For the SVD of $\mathbf{R}$ to compute the eigenvalues, first column and row $j$ are removed from $\mathbf{R}$ if $j$ is a supplementary variable, and then $\mathbf{R}_{ij}$ is multiplied by $\left(v_i v_j\right)^{1/2}$.

If passive missing treatment is applicable for a variable, missing values are imputed with the variable mode, regardless of the passive imputation specification.

### After transformation

When all analysis variables are non-multiple, and there are no missing values, specified to be treated as passive, the correlation matrix is:

$$\mathbf{R} = n_w^{-1}\mathbf{Q}'\mathbf{W}\mathbf{Q} \text{, with } \mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j \,.$$

The first $p$ eigenvalues of $\mathbf{R}$ equal $\mathbf{\Lambda}^{1/2}$. (See also Optimization step (5) and Model Summary for interpretation of $\mathbf{\Lambda}^{1/2}$).

When there are multiple nominal variables in the analysis, $p$ correlation matrices are computed ( $s = 1, \ldots, p$ ):

$$\mathbf{R}_s = n_w^{-1}\mathbf{Q}_s'\mathbf{W}\mathbf{Q}_s ,$$

with $\mathbf{q}_{js} = \mathbf{G}_j\mathbf{y}_j$ for non-multiple variables

and $\mathbf{q}_{js} = \mathbf{G}_j\mathbf{Y}_{js}\left(\mathbf{Y}_{js}'\mathbf{D}_j\mathbf{Y}_{js}\right)^{-1/2}$ for multiple nominal variables.

The $s^{th}$ eigenvalue of $\mathbf{R}_s$ is equal to $\lambda_s^{1/2}$ (see Model Summary Table).

If there are missing values, specified to be treated as passive, the mode of the quantified variable or the quantification of an extra category (as specified in syntax; if not specified, default (mode) is used) is imputed before computing correlations. Then the eigenvalues of the correlation matrix do not equal $\mathbf{\Lambda}^{1/2}$ (see Model Summary section). The quantification of an extra category for multiple nominal variables is computed as

$$\mathbf{Y}(k_j+1)js = \left(\sum_{i\in I} w_i\right)^{-1}\sum_{i\in I} w_i x_{is} ,$$

with $I$ an index set recording which objects have missing values.

For the quantification of an extra category for non-multiple variables first $\mathbf{Y}(k_j+1)js$ is computed as above, and then

$$\mathbf{y}(k_j+1)j = n_w^{1/2}\left(\sum_s a_{js}^2\right)^{-1}\sum_s a_{js}\mathbf{Y}(k_j+1)js .$$

For the SVD of $\mathbf{R}$ to compute the eigenvalues, first column and row $j$ are removed from $\mathbf{R}$ if $j$ is a supplementary variable, and then $\mathbf{R}_{ij}$ is multiplied by $\left(v_i v_j\right)^{1/2}$.

## Object Scores and Loadings

### Normalization

Normalization partitions the first $p$ singular values of $n_w^{1/2}\mathbf{Q}$ over the objects scores $\mathbf{X}$ and the loadings $\mathbf{A}$ (for $\mathbf{Q}$ see the Correlations and Eigenvalues section). The singular value decomposition of $n_w^{1/2}\mathbf{Q}$ is

$$\mathrm{SVD}(n_w^{1/2}\mathbf{Q}) = \mathbf{K} n_w^{1/2}\mathbf{\Phi}^{1/2}\mathbf{L}' = n^{b/2} n^{a/2}\mathbf{K}\mathbf{\Phi}^{a/2}\mathbf{\Phi}^{b/2}\mathbf{L}'.$$

The first $p$ singular values $\mathbf{\Phi}_p^{1/2}$ equal $\mathbf{\Lambda}^{1/4}$, with $\mathbf{\Lambda}$ as computed in step (4) during the last iteration. (See also Optimization step (5) and Model Summary for interpretations of $\mathbf{\Lambda}^{1/2}$).

With $\mathbf{X} = n^{b/2}\mathbf{K}_p\mathbf{\Lambda}^{a/4}$ and $\mathbf{A} = n^{a/2}\mathbf{L}_p\mathbf{\Lambda}^{b/4}$, $\mathbf{X}\mathbf{A}'$ gives the best $p$-dimensional approximation of $\mathbf{Q}$.

During the optimization phase, variable principal normalization is used. Then, after convergence $\mathbf{X} = n_w^{1/2}\mathbf{K}_p$ and $\mathbf{A} = \mathbf{L}_p\mathbf{\Lambda}^{1/4}$.

If variable principal normalization is requested, $\mathbf{X}_\mathrm{n} = \mathbf{X}$ and $\mathbf{A}_\mathrm{n} = \mathbf{A}$, else

$$\mathbf{X}_\mathrm{n} = n_w^{1/2(b-1)}\mathbf{X}\mathbf{\Lambda}^{a/4}$$

$$\mathbf{A}_\mathrm{n} = n_w^{a/2}\mathbf{A}\mathbf{\Lambda}^{1/4(b-1)},$$

with $a = (1+q)/2$, $b = (1-q)/2$, and $q$ any real value in the closed interval [-1,1], except for independent normalization: then there is no $q$ value and $a = b = 1$.

$q = -1$ is equal to variable principal normalization, $q = 1$ is equal to object principal normalization, $q = 0$ is equal to symmetrical normalization.

For $\mathbf{Q}$ and singular values when not all variables have non-multiple scaling level, see the Correlations Transformed Variables section.

## Quantifications

The centroid coordinates are $\mathbf{Y}_j$ if variable principal normalization is requested, else the centroid coordinates are computed as $\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{W}\mathbf{X}_{\mathrm{n}}$ .

For multiple nominal variables the quantifications are the centroid coordinates. For non-multiple variables the quantifications $\mathbf{y}_j$ are displayed, and the vector coordinates $\mathbf{y}_j\mathbf{a}_{\mathrm{n}_j}'$ .

If a category is only used by supplementary objects (i.e. treated as a passive missing), only centroid coordinates are displayed for this category, computed as

$$\mathbf{y}_{(j)k} = n_w^{1/2} n_{jk}^{-1} \sum_{i \in I} \mathbf{x}_{\mathrm{n}_i}$$

where $\mathbf{y}_{(j)k}$ is the $k^{\text{th}}$ row of $\mathbf{Y}_j$ , $n_{jk}$ is the number of objects that have category $k$ ,and $I$ as n index set recording which objects are in category $k$ .

## Residuals

For non-multiple variables, Residuals gives a plot of the quantified variable $j$ ($\mathbf{G}_j\mathbf{y}_j$) against the approximation, $\mathbf{X}\mathbf{a}_j$ . For multiple variables plots per dimension are produced of $\mathbf{G}_j\mathbf{Y}_{js}$ against the approximation $\mathbf{X}_s$ .

## Projected Centroids

The projected centroids of variable $l$ on variable $j$ , $j \notin J$ ,are

$$\mathbf{Y}_l\mathbf{a}_j(\mathbf{a}_j'\mathbf{a}_j)^{-1/2}$$

# References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. 1972. *Statistical inference under order restrictions.* New York: John Wiley & Sons, Inc.

Cliff, N. 1966. Orthogonal rotation to congruence. *Psychometrika*, 31: 33–42.

De Leeuw, J., Young, F. W., and Takane, Y. 1976. Additive structure in qualitative data: an alternating least squares method with optimal scaling features. *Psychometrika*, 31: 33–42.

Gifi, A. 1990. *Nonlinear multivariate analysis*. Leiden: Department of Data Theory.

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29: 115–129.

Max, J. (1960), Quantizing for minimum distortion. IRE Transactions on Information Theory, 6, 7-12.

Pratt, J.W. (1987). *Dividing the indivisible: using simle symmetry to partition variance explained.* In T. Pukkila and S. Puntanen (Eds.), Proceedings of the Second International Conference in Statistics (245-260). Tampere, Finland: University of Tampere.

Ramsay, J.O. (1988) Monotone regression Splines in action, Statistical Science, 4, 425-441.

Wilkinson, J. H. 1965. *The algebraic eigenvalue problem*. Oxford: Clarendon Press.