***Clustering tendency***

SPSS macros by Kirill Orlov

kior@akado.ru, ttnphns@gmail.com

<https://www.spsstools.net/en/KO-spssmacros>

All rights reserved

*Clustering tendency*. Block-diagonalization of a distance matrix between objects allows to judge preliminary before cluster analysis whether there are clusters in the data and how many.

*Read “*[*About SPSS macros*](https://www.spsstools.net/en/KO-aboutmacros)*” what are they and how to run them.*

*The “Protected directory” error.* Some of the macros described in the current document write temporary files to hard disc. If you don't have full Administrator rights of your computer, it may cause error saying, among things: *“SPSS Statistics cannot access a file... specifies a protected directory...”*, meaning that the default directory the macro wants to use is protected on your PC. To solve the problem, in Syntax window issue command: CD 'myfolder'., where 'myfolder' is the path/name of some folder where you are allowed to save files to.

# MACRO !KO\_BLOCKDIAG: BLOCK-DIAGONALIZATION OF DISTANCE MATRIX

Version 1, Mar 2023. Tested on SPSS Statistics 22, 27, 30.

!KO\_blockdiag matrix= *VAR1 to VAR80* /\*Columns constituting the matrix body (may use "to")

/id= /\*Optionally: numeric identifier variable of cases (rows)

/method= VAT /\*Method: VAT (default) or MDS

/poster= YES /\*"Posterization" after permutation: YES or NO (default)

/plot= /\*Heatmap: GREY (default) or RGREY;

/\*word LABEL can be appended then; or NONE

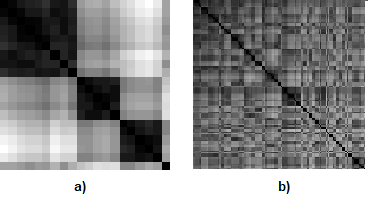
/bounds= /\*Bounds for the heat scale on the map: AUTO (default),

/\*OBS, or min max

/mds= /\*With METHOD=MDS: transformation in MDS: SPLINE (default), ORDINAL, INTERVAL.

Minimal specification MATRIX.

Let there be a matrix of distances (dissimilarities) between objects. Block-diagonalization of the square symmetric distance matrix is the reordering of its rows/columns – i.e., the objects – so that objects with small distances between them become placed next of each other in the matrix (i.e., their indices become close). On the heatmap, such permuted matrix looks consisting of diagonal blocks. It will be the more distinctly, contrastly diagonal-blocked the stronger there exists the tendency towards clusteredness among the objects, that is, the stronger the distances fall themselves apart into “within-cluster” (small) and “between-cluster” (large). Therefore, one can approximately judge by the heatmap about presence or absence of clusters in the data and about the number of clusters, without doing cluster analysis. Each cluster on the heatmap after block-diagonalization of the matrix looks like a block on the diagonal. If there are no clusters in the data, block-diagonalization will not show clear blocks on the heatmap.



**Fig. 1**. Heatmap of a block-diagonalized distance matrix: a) with clear clusters in data, b) without clusters in data.

The macro does block-diagonalization (permutes rows/columns) of the input distance matrix, saves the obtained matrix as a new unnamed dataset and draws heatmap.

**Algorithm**

With METHOD=VAT, the macro does reordering of rows/columns of the distance matrix by VAT algorithm. VAT algorithm (“Visual Assessment of [Cluster] Tendency”) is described detailed in [1], also in [2,3]. It is tightly connected with the Prim’s algorithm of building minimum spanning tree in a weighted graph, and through this is implicitly akin to hierarchical clusterization by the method of single linkage or nearest neighbour [2]. The common feature of these algorithms is the stepwise growing of a spanning tree/cluster/block by adding of closest elements.

iVAT method (“improved VAT”) [3] is a superstructure to VAT and is equivalent in result to application of Floyd–Warshall algorithm in its “identify easiest passes” version to the reordered matrix returned by VAT. iVAT specifically replaces some distances in the matrix by other its distances, thus lessening diversity of distances in the matrix. The iVAT’s idea is simple: if two points as a pair are far from each other but are mediated by a chain of points which all links (distances) are small, one should recognize that the two points are “in fact” close. The effect of iVAT-substitution is that on the heatmap (1) the contrast between between-cluster and within-cluster distances will get stronger, helping to visually uncover clusters-blocks; (2) detectability of clusters of chain-like structure (including strongly elongated, dendric, ring-like) will enhance. iVAT method is performed by the macro when METHOD=VAT /POSTER=YES.

With METHOD=MDS, the macro does reordering of rows/columns of the distance matrix by using multidimensional scaling (SPSS command PROXSCAL) with weighting of distances. It is the macro’s author idea (which doesn’t mean it is a novelty). Each distance *dij*, the matrix element, receives weight (importance) , where *Ri*is the rank of value *dij* in row *i*, *Rj* is the rank of value *dij* in row *j*, *R* is the rank of value *dij* in the triangle of the matrix. PROXSCAL executes the mapping (ordination) in space of dimensionality 1. Rows/columns of the distance matrix are reordered by ascending of coordinates by the dimension. The idea of the method is to force the distances to distribute along the diagonal, giving small distances the priority at that. In VAT method, the size of a distance determines the order of its inclusion in the spanning tree. In MDS method, the size of a distance determines its importance in affecting the ordination. In both methods the final result is that clumps of small distances form blocks threaded on the diagonal.

METHOD=MDS /POSTER=YES applies Floyd–Warshall algorithm in its “identify easiest passes” version to the reordered matrix returned by the MDS method. It has the same effect as iVAT after VAT.

Sources

1. Bezdek, J.C., Hathaway, R.J. VAT: a tool for visual assessment of (cluster) tendency // Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 – 2002 – Volume 3 – p. 2225–2230. [DOI:10.1109/IJCNN.2002.1007487]
2. Havens, T.C., Bezdek, J.C., Keller, J.M, Popescu, M., Huband, J.M. Is VAT really single linkage in disguise? // Annals of Mathematics and Artificial Intelligence. – 2009 – Vol. 55, article 237 [DOI 10.1007/s10472-009-9157-2]
3. Havens, T.C., Bezdek, J.C. An efficient formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm // IEEE Transactions on Knowledge and Data Engineering – 2012 – vol. 24, no. 5, p. 813-822 [DOI: 10.1109/TKDE.2011.33]

**Limitations**

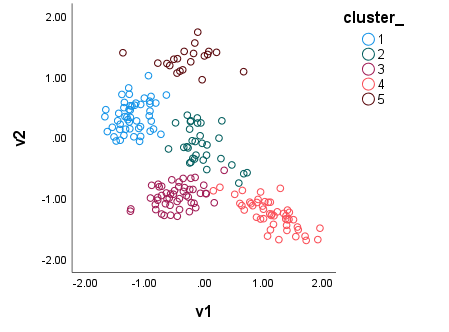
With METHOD=MDS, the matrix compiled for the analysis is up to 700 rows/columns. With METHOD=VAT, the macro does not put limitations, however the recommended matrix size is not above 1000, otherwise heatmap will take long to draw. You can always do the analysis on a random subsample of cases if your data are big.

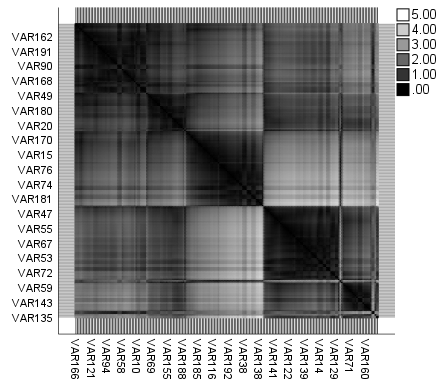
EXAMPLE 1.

proximities v1 v2 /view= case /measure= seuclid /matrix= out(\*) /print= none.

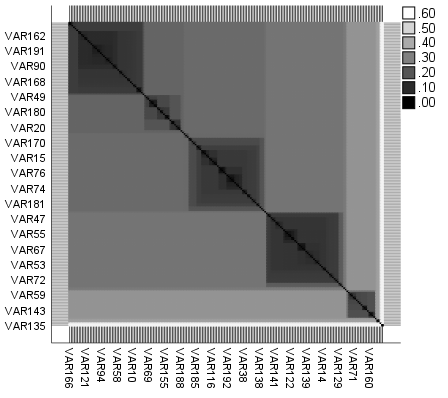
dataset name dist.

!KO\_blockdiag matrix= VAR1 to VAR199.





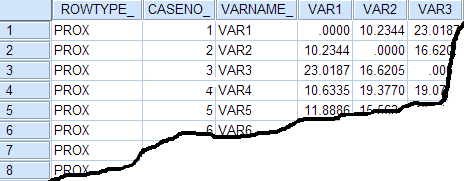
!KO\_blockdiag matrix= VAR1 to VAR199 /poster= YES.



* PROXIMITIES computes squared euclidean distances between cases of the data and saves the matrix as a new dataset which is called *DIST*.
* The macro performs block-diagonalization by VAT method and builds the heatmap. Five clusters can be discerned on it.
* In the second run, the option of “posterization” is added. The five clusters manifest more clearly.

***Matrix structure***

The dataset must be a matrix of pairwise distances (dissimilarities, not similarities). Variable names – matrix columns – up to 8 bytes. Required is the presence of variable VARNAME\_ naming rows in correspondence to columns. Names, which are the values of the variable, should be written in the same case of letters as the identical to them names of the columns. The macro does not require that the rows and the columns go in the same order or their number and list be fully identical: the macro will by itself select from the input matrix rows and columns with the same names and will co-order them so that the compiled matrix be square and diagonalized structure. Variable ROWTYPE\_ and other auxiliary are not necessary in the input matrix.



***Subcommands***

**MATRIX**

Specify variables of the working dataset which are proper the columns of the matrix of distances (dissimilarities). You may list all or just needed columns and in arbitrary sequence. May use “to” to specify by range. If you have similarities, transform them first into dissimilarities the way you find suitable.

*Specification of open range with the help of “?”*. In some instances there may arise need to specify a range between a pair of variables which themselves are not included in the range. Use “?” at the sides, for that. For example, *?VARNAME\_ to ENDVAR?* means all variables found in the dataset between variables *VARNAME\_* and *ENDVAR*, excluding these two. To specify a range which is open from one side, use “?” only from that side. For example: *?VARNAME\_ to VAR100* or *VAR1 to ENDVAR?*.

Since your data are dissimilarities, “diagonal” values – i.e., data in cells on the intersection of rows and columns of the same name – must be zeros, and all other (“offdiagonal”) values must be nonnegative; greater value corresponds to greater dissimilarity.

EXAMPLE 2.

temporary.

sample 0.2.

!KO\_blockdiag matrix= VAR1 to VAR100 /method= MDS /poster= YES .

* SAMPLE command temporarily (under TEMPORARY) selects randomly 20% of rows of the distance matrix.
* The macro takes columns *VAR1* to *VAR100*. The matrix compiled by the macro for the analysis consists of rows and columns that are the intersection of the two lists – selected rows and selected columns.
* The macro performs block-diagonalization by MDS method and “posterization” before building the heatmap.

**ID**

Optional numeric identifier variable of cases (objects). Variable name up to 8 bytes long. There should be no missing values in the variable.

**METHOD**

Indicate the method of block-diagonalization:

VAT - (default/unspecifying) VAT method.

MDS - multidimensional scaling with weighting of distances is performed by SPSS procedure PROXSCAL. It is available in SPSS Statistics Professional Edition or in Categories option.

Both methods often yield very similar results.

**POSTER**

By default/unspecifying and with POSTER=NO, the macro performs only block-diagonalization, i.e., permutation of rows/columns. With POSTER=YES, it does “posterization” after it. Heatmap at “posterization” looks more contrast, and usually clusters (blocks) are easier to see on it. Besides, there somewhat enhances detectability of clusters of chain-like structure (including strongly elongated, dendric, ring-like). Sometimes, though, “posterization” conceals clusters that are close to each other.

“Posterization” with VAT method is known as iVAT (“improved VAT”). “Posterization” affects only the heatmap, it does not affect the matrix being saved. “Posterization” is not the posterization of the obtained image of the heatmap, it is the contrasting of distances in the matrix on which the heatmap is immediately built.

**PLOT**

Heatmap is drawn in greyscale. With PLOT= GREY (default/unspecifying) the higher the value of the element the brighter it is, and with PLOT=RGREY it is opposite – the darker it is. After the keyword, you may add the second keyword LABEL, to label the cells with elements’ values. PLOT=NONE does not produce the heatmap.

**BOUNDS**

This subcommand is not in effect with PLOT=NONE. It sets bounds for the brightness scale.

AUTO - (also default/unspecification) let SPSS automatically define suitable bounds.

OBS - the bounds exactly match with the observed minimal and maximal values in the matrix.

*min* *max* - specify the bounds manually as two numbers, minimum and maximum. Indicate values by and large comparable with the values of the matrix.

Manual specification of the bounds means that you are fixing the brightness mapping on the picture with respect to the size of the elements in the matrix. It becomes possible to compare different matrices with each other by tone immediately.

**MDS**

This subcommand acts with METHOD=MDS. You may choose to use metric or nonmetric multidimensional scaling with PROXSCAL command. Specify INTERVAL (metric), ORDINAL (nonmetric) or SPLINE (in-between metric and nonmetric). By default/unspecifying, MDS=SPLINE.

***Special regimes***

The macro does not obey weighting (however, it doesn’t take in the procedure cases with missing and nonpositive weights). It is not suited for the split state of the dataset (SPLIT FILE). The macro obeys commands selecting cases (SELECT IF, FILTER, USE), including those standing under TEMPORARY command.