**WorldPCA manual**

**About the software.** The software performs principal component analysis (PCA) based on correlation matrices where analyzed individuals can be divided into training and test sets, or analyzed as a whole. The PC equations are calculated using only the training set, and coefficients from those equations are used for calculating PC values for the individuals included in the test set.

Results of analyses can be visualized in WorldPCA in the form of scatterplots and diagrams. But, more importantly, the software implements a number of analytical functions tools aimed at exploring the results of the PCA. Virtually all of these functions are based on a very simple statistic: mean pair-wise Euclidean distance (MPED) which can be calculated between individuals of the same sample or different samples using various combinations of PCs (see Evteev et al., ... for details).

The functions implemented in WorldPCA are the following:

1) Search for individuals that can be called "morphological outliers", i.e. those lying well out of the range of the sample's morphological variation. This is carried out by comparing MPED of the individual to other individuals of the sample with mean MPED of this sample. The output is presented graphically in the form of bar charts.

2) Assessment of the total range of variation in a sample (mean MPED) in the morphospace of chosen PCs. This range can be compared with ranges of other samples, and the difference can be tested for significance using the Mann-Whitney U-test.

3) Search for clusters of individuals of samples using different methods of clustering: k-means, Birch, agglomerative clustering. Individuals can also be assigned to clusters manually. The software calculates MPED of individuals of a particular cluster to other individuals of the sample (i.e. not from this cluster). When compared with MPED of the whole sample, it can demonstrate how distinct the cluster is.

4) Intergroup MPED can be calculated as well.

The software can be used for exploring any type of continuously varying data. In contrast to some existing methods of analyzing intergroup variation, no particular assumptions about the nature of the data should be made, and any number of samples can be compared simultaneously instead of just one or two. As the approach is based on principal component analysis, it is well suited for assessing not only differences but also the similarity between samples.

**Download and installation.** If you prefer to use the *Python version* of the software, please download all the necessary files and manual ("PythonWorldPCAmanual") at https://github.com/Nikolay-Staroverov/WorldPCA. Otherwise, it is possible to download an executable file of the software at https://drive.google.com/drive/folders/1C6g4ckwCIOlnMGBX4aqMMLEPX8KHcz5V?usp=sharing. Please download the "Icons" folder as well, and store it in the same folder as the exe. file.

**Input file preparation.** Use one of the xlsx files stored at https://github.com/Nikolay-Staroverov/WorldPCA ("SampleFullHowells" or "HowellsFemales") as an example. The data you plan to analyze should be placed on the very first sheet and contain numerical values only. The top line must contain names of the columns in the following order: "Title", names of the variables, "Active", "target". Please keep the names of the first and two last columns exactly as they are spelled here, or better just copy them from the example files. In the column called "Title" names or codes of individuals (units of analysis) are listed. The following columns contain values to be analyzed (column=variable). The "Active" column provides information regarding what individuals should be included in a training set (1) or a test (0) sets (see Evteev, Dvurechensky, 2017; Evteev et al., ... for details). Finally, the "target" column specifies groups (or populations) to which the individuals belong.

**Input file upload and PCA analysis.** Load your file into the software using the "load base" button then press the "fit" button. In a pop-up window which emerges after pressing the "fit" button the variables to be employed in the analysis and groups to be included in the training set can be specified. After this, a PCA analysis *per se* is performed, and the following steps are about visualizing and exploring the results of the analysis.

**Main window.** By pressing the "analyze" button, a pop-up window is activated where the user can specify what principal components (PCs) and what groups to visualize. It is also possible to choose to add or not to add annotations to individual points on the plot. After the "apply" button is activated, the graphical and numerical results are presented in the left and right windows, respectively.

*Scatterplot.* On the top of the scatterplot in the left window, there is a number of icons that can be used to change the appearance of the graph. It is possible to pan or zoom the plot with respect to the axes, pick a rectangular area of the plot to zoom in, change the graph margins, and edit labels and ranges of the axes. Some additional changes can be done by simply changing the form size of the software's window. The user can move between different views (i.e. before or after making some changes), or reset the original view. The graph can be exported by pressing the "floppy-disk" button.

*Table.* In the spreadsheet on the right, individual values of the first four PCs can be found and exported by pressing the button on the bottom left of the table. The "show explained variance" button on the bottom right of the table displays eigenvalues of the PCs and the proportion of explained variance.

The appearance of the window and the visualization tools are basically similar in all the modules described below.

**Clustering.** The user should specify what PCs to use for calculating clusters and what to present graphically. Keep in mind that these are not necessarily the same, e.g. clustering can be based on all the four PCs while only two of them can be presented on the scatterplot. Groups to be employed should be specified as well (buttons "groups" and "data for visualization"). Then one of the clustering methods must be chosen and the desired number of clusters specified. It is possible to add or not to add annotations to individual points on the plot. The scatterplot on the left and the table on the right are very similar to that in the main window; both can be exported in the same way as described in the previous section. The crosses on the graph depict means of respective clusters. An additional column entitled "clusters" is presented in the table which indicates which cluster belongs one or another individual. This cluster affiliation can be changed manually by simply changing the number in the respective cell. After pressing the "approve clusters" button, two new windows appear. A first contains MPEDs for clusters, i.e. MPEDs of individuals of a particular cluster to other individuals of the sample (i.e. not from this cluster) along with the MPED of the whole sample (see Evteev et al., ... for details). This can help to assess how distinct the clusters are. A second window displays U-test values between clusters' and whole samples' MPEDs. As an empirical suggestion, only those clusters should be considered that a) are significant, b) MPEDs of which are substantially larger compared to the whole sample's MPEDs.

**Histograms.** The PCs to be used for creating histograms and groups to be employed (buttons "groups" and "data for visualization") should be specified. The bar chart on the graph displays the distribution of individuals' MPEDs (the y-axis), i.e. the mean MPEDs between a particular individual and the rest of the group (used for creating the histogram). The bottom red line stands for the mean MPED of the group, the upper red line - for the mean + 1SD of the group. Two buttons are visible on the bottom left of the graph. By pressing the "show all dist matrix" button, the user can obtain and export MPEDs between all individual in the sample used for the PC analysis. After pressing the "show group dist matrix" button, only MPEDs between individuals represented on the histogram will appear.

**Distances.** This module calculates *intergroup* Euclidean (MPED) or Makhalanobis distances for a set of groups and a pair of PCs chosen by the user. The output includes two tables. The first table entitled "distance" contains pair-wise between-group MPEDs and, in the very last column, intragroup MPEDs for all the groups compared. The second table contains U-test values showing if intragroup MPEDs of the samples differ significantly from each other.

Please see the previous section regarding calculation of *between-individual* MPEDs.