# Exploratory data structure comparisons: Three new visual tools based on Principal Component Analysis

| | |
|---|---|
| Journal: | *Journal of Applied Statistics* |
| Manuscript ID | CJAS-2018-0825 |
| Manuscript Type: | Original Research Paper |
| Date Submitted by the Author: | 11-Sep-2018 |
| Complete List of Authors: | Petersen, Anne; University of Copenhagen, Department of Public Health<br>Markussen, Bo; University of Copenhagen<br>Christensen, Karl; University of Copenhagen |
| Keywords: | Principal component analysis, exploratory data analysis, covariance matrix, data structure comparison, psychological well-being, European Social Survey, International ranking |
| <a href="http://www.ams.org/mathscinet/msc/msc2010.html" target="_blank">2010 Mathematics Subject Classification</a>: | 62P15   Applications to psychology,<br>62H25   Factor analysis and principal components; correspondence analysis |
| | |

ORIGINAL RESEARCH ARTICLE

# Exploratory data structure comparisons: Three new visual tools based on Principal Component Analysis

**ABSTRACT**
Datasets are sometimes divided into distinct subsets, e.g. due to multi-center sampling, or to variations in instruments, questionnaire item ordering or mode of administration, and the data analyst then needs to assess whether a joint analysis is meaningful. The Principal Component Analysis-based Data Structure Comparisons (PCADSC) tools are three new non-parametric, visual diagnostic tools for investigating differences in structure for two subsets of a dataset through covariance matrix comparisons by use of principal component analysis. The PCADCS tools are demonstrated in a data example using European Social Survey data on psychological well-being in three countries, Denmark, Sweden, and Bulgaria. The data structures are found to be different in Denmark and Bulgaria, and thus a comparison of for example mean psychological well-being scores is not meaningful. However, when comparing Denmark and Sweden, very similar data structures, and thus comparable concepts of well-being, are found. Therefore, inter-country comparisons are warranted for these countries.

## 1. Introduction

Data comparability is a recurring topic in a wide array of statistical applications. Often, data will be collected in such a way that it is essentially divided into several subsets whose comparability needs to be assessed empirically. This happens for instance when data are collected across several centers (or countries) or when different versions of an instrument, the mode of administration of a survey, or the order of survey questionnaire items are applied. While a mixture of modes of administration, for example mail and telephone, can improve response rates in surveys, it can also be quite problematic by inducing differences in response behavior that may lead to biased results [3, 14]. Powers, Mishra, and Young report effects of mode of administration on changes in mental health scores that are of a magnitude that is considered to be clinically meaningful [16], so the comparability issue is not just a statistical puzzle. Similarly, combining data obtained from different sampling schemes can also be problematic, as illustrated by Liu, who warns against combining online panel data with intercept samples (a pool of respondents obtained through banners, ads, or promotions)[12]. But mixing several data collection methods is essential for acheiving further advances in surveying- and data collection methodology, as it allows us to simultaneously build on existing methods and answer new empirical questions. However, if we do not address the question of comparability among the data subsets arriving from e.g. different countries, we risk

conducting analyses whose most fundamental assumptions are not satisfied.

The comparability issue in these examples can be summarized as follows: Assume that we have two datasets with the same variables, but different observations, often represented as a single dataset with a subset-inducing variable, and that we wish to compare them without specifying a model, or even a variable of interest. The central question is then whether the two datasets can readily be combined for the purpose of later data analysis, or if the subset-inducing variable implies heterogeneity that must be dealt with in later statistical modelling.

Sophisticated methods for addressing this question are available when we are willing to assume a statistical model. But this places the effort of assessing data comparability very late in the data analysis process and it makes the comparability assessment ad-hoc, as it essentially relies on the appropriateness of the modeling choices, which again depends on the structure of the data. The use of (parametric) models does thus not constitute a general data structure comparison method, but rather a fitted-model comparison method. It addresses the interplay between the model and the data, not the data alone. However, useful tools for initial, exploratory investigations of data comparability are mostly absent. What is needed are procedures that compare differences in data structures in two subsets of a dataset without assuming neither directional nor hierarchical relationships between the variables. Simple methods like variable-by-variable tests of distributional differences suffer from the drawback that they only address marginal differences and ignore the interplay between variables, but entry-by-entry comparisons of two empirical correlation matrices quickly become unmanageable as the number of variables increase.

We propose three visual tools for data structure comparisons, which we will refer to collectively as Principal Component Analysis-based Data Structure Comparisons (PCADSC). These methods use principal component decomposition of the empirical covariance matrix in the two subsets to create intuitive visualizations of data structure differences. The use of PCA implies that we can only compare datasets with numerical variables and that we focus on the linear aspect of the data structures. The proposed tools are implemented in the R package PCADSC [15].

## 1.1. Comparing psychological well-being

An example of where the data structure comparability issue must be considered is when conducting international rankings of countries according to e.g. educational quality (e.g. the PISA project) and citizen happiness (e.g. the UN *World Happiness Report* project). From a methodological point of view, such international rankings are problematic, as they rely on the fundamental assumption that the measured concepts are inherently the same across countries.

We use data from the 2012 version of the European Social Survey (ESS) project to investigate inter-country differences in psychological well-being and happiness. In the rankings of happiness, not much work has yet been devoted to evaluating the assumption of international comparability, though Veenhoven presents a theoretically thorough, but empirically simplistic, summary of possible reasons for lack of comparability [18], while Lolle & Andersen show highly potent translation issues for the term *happiness* [13]. If for instance two countries differ in terms of how social networks are typically built and structured, with one emphasizing family relations and the other mostly focusing on other social relations, then having a weak family connection does not have the same implications in the first country as it does in the second one. More specifically,

2

whereas in the first country, lack of familial network might be related to loneliness, lack of general social capital, and isolation, in the second country, the quality of the family network might not be informative at all about other aspects of a person's social or psychological well-being. The two countries thus differ in how different aspects or measures of psychological well-being are interrelated, which is essentially a difference in data structures. And therefore, comparing the two countries in these measures is not a meaningful endeavour.

Our starting point is Denmark, a small, Northern European country that has repeatedly been awarded with the title of "happiest country in the world" by the *World Happiness Report*, most recently in 2016 [7], and we wish to investigate if this title is really meaningful at all. In order to do this, we compare the Danish ESS psychological well-being data with that of Bulgaria. Though both countries are European and thus neither very far apart geographically nor culturally, these two countries have previously been highlighted to be very different in terms of what defines happiness [8]. Moreover, intra-European, regional differences in the relationship between social capital and happiness have also been demonstrated [17]. When comparing Northern European countries to other European countries, a much less pronounced relationship between the two concepts is found. In particular, interpersonal relations play a less important role in Denmark, compared to Bulgaria. Therefore, a successful method for data comparisons should be able to detect these differences by looking at data on psychological well-being from these two countries. We also compare the Danish data with Swedish data in order to investigate if the PCADSC tools actually do hold discriminatory power. Denmark and Sweden are both Northern European countries and are often deemed very similar in terms of culture and history. Therefore, we expect fundamental concepts, such as psychological well-being, to be similar across these two countries.

## 2. Methodology

PCADSC compares the covariance matrices of two subsets of a dataset. If all variables in the subsets are jointly normal with known means, the covariance matrices are sufficient statistics describing the joint distribution of all the variables, but even without the normality assumption, pairwise correlations and marginal variable variances are still interesting quantities describing linear interrelations between variables. This makes the empirical covariance matrix a reasonable place to start looking for differences in data structures. Computing empirical covariance matrices for the two subsets and comparing them entry-by-entry becomes increasingly difficult as the number of variables increases and for this reason we decompose and recompose the covariance matrices using *principal component analysis* (PCA) to get an overview of differences between the two subsets. We refer to Appendix A for a brief introduction to PCA.

We assume that the variables have been standardized before PCA is conducted, which is equivalent to working with the empirical correlation matrices rather than the empirical covariance matrices. We also assume that all considered variables have a numerical interpretation, for example by being continuous or ordinally categorical. Moreover, we use the following notation: $\mathbf{X} \in \mathbb{R}^{n_x \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n_y \times d}$ are datasets containing the same number of variables, $d$, but possibly different numbers of observations, $n_x$ and $n_y$. We use $X_j$ to refer to the $j$th variables of $\mathbf{X}$ and $x_{ij}$ to refer to the $i$th observation within that variable, while $x_{i\cdot}$ is the full $i$th observation row. Note that $X_j \in \mathbb{R}^{n_x}$, $x_{ij} \in \mathbb{R}$, and $x_{i\cdot} \in \mathbb{R}^d$. We let $\bar{X} = (\bar{X}_1,...,\bar{X}_d)^T = \frac{1}{n_x}\sum_{i=1} x_{i\cdot}$ denote the variable averages and use $S_x = \frac{1}{n_x-1}\sum_{i=i}^n (x_{i\cdot} - \bar{X})(x_{i\cdot} - \bar{X})^T \in \mathbb{R}^{d \times d}$ to denote the empirical

3

covariance matrix of $\mathbf{X}$.

### 2.1. Using PCA for data structure comparisons

The tools presented here are all based on comparing the PCA results across two different datasets that contain the same variables, $\mathbf{X}$ and $\mathbf{Y}$. Let

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \in \mathbb{R}^{(n_x + n_y) \times d}$$

be the combined dataset. For each of these three datasets, we complete the following steps (here described for $\mathbf{X}$ only):

(1) Standardize each of the variables to have mean zero and unit standard deviation. Let $\tilde{\mathbf{X}}_j \in \mathbb{R}^{n_x \times d}$ be the standardized dataset.
(2) Form the principal component analyses

$$S_x = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} \tilde{x}_{i\cdot} \tilde{x}_{i\cdot}^\top = \sum_{j=1}^{d} \lambda_j^x \eta_j^x (\eta_j^x)^\top$$

thereby obtaining loadings $\eta_j^x$ and eigenvalues $\lambda_j^x$ for $j = 1, \dots, d$.

The hereby obtained PCA decompositions of the correlation matrices can then be compared. The standardization implies that the diagonal elements of $S_x$, $S_y$, and $S_z$ all equal 1, and thus also that

$$\sum_{j=1}^{d} \lambda_j^x = \sum_{j=1}^{d} \lambda_j^y = \sum_{j=1}^{d} \lambda_j^z = d.$$

This identity will simplify some expressions below. We note that the sequence of loadings $\eta_j$ and their associated eigenvalues $\lambda_j$ yield a simultaneous description of the structure of the dataset for all approximating dimensions $q$. This implies that the loadings and the eigenvalues can be used to investigate the structure of the dataset without the need to decide on an approximating dimension, $q$, a priori.

We present three diagnostic plots that are designed to shine a light on different types and levels of data structure differences, namely the *cumulative eigenvalue (CE) plot*, the *angle plot*, and the *chroma plot*. These plots are dsecribed in turn below. For the deepest understanding of the data structure differences in two datasets, we suggest using all of the three plots in the same order as they are presented. While we describe the three plot types from a purely theoretical point of view below, we also provide example illustrations as a supplement to the somewhat technical definitions. These plots are available in Figures 2, 3, and 4 for the CE-, angle-, and chroma plots, respectively, and they are based on two simulated datasets:

**Dataset A:** This dataset contains 1000 independent simulations from the same underlying model, namely *model 1* from Figure 1. The observations are randomly divided into two groups.
**Dataset B:** This dataset contains 500 independent simulations from *model 1* from Figure 1 and 500 independent simulations from *model 2* in the same figure. The

4

observations are naturally divided into groups corresponding to which model they were simulated from.

In *model 1* the observed variables $V_3$ and $V_4$ are caused by a common, unobserved variable $L_1$. Further details on the models and the simulated data are available in the supplementary material, including covariance matrices for the two models, in Appendix A. In dataset **A**, all observations stem from the same, underlying data structure and thus, the PCADSC plots should point towards no notable differences in data structures. In dataset **B**, on the other hand, there are two fundamentally different, underlying data structures and therefore, the PCADSC plots should illustrate this lack of homogeneity in data structures.

### 2.2. *The cumulative eigenvalue plot*

The cumulative eigenvalue (CE) plot compares the eigenvalues of the correlation matrices. These eigenvalues represent how much information is withheld in each component in terms of explained variance. Thus, by comparing cumulative sums of eigenvalues, it is possible to obtain a detailed picture of how the two datasets differ in terms of what components are the most informative. In order to investigate whether the same proportion of the total variation can be described by the same number of principal components in the two datasets, we plot a piecewise linear curve connecting the points

$$(0,0), \ (\lambda_1^z, \lambda_1^x - \lambda_1^y), \ (\lambda_1^z + \lambda_2^z, \lambda_1^x + \lambda_2^x - \lambda_1^y - \lambda_2^y), \ \dots,$$
$$\left( \sum_{j=1}^{d} \lambda_j^z, \sum_{j=1}^{d} \lambda_j^x - \sum_{j=1}^{d} \lambda_j^y \right)$$

Due to the standardization, the last point will always be equal to $(d, 0)$. Thus, this curve will begin and end at the x-axis. And the larger excursions it makes away from the first axis, the less alike the cumulative eigenvalue sums for the two datasets are. Moreover, a positive cumulative difference implies that dataset **X** holds more information in the first components than dataset **Y** does. By using the cumulative eigenvalues of **Z** as the first coordinate, we obtain a visualization where a component take up an amount of horizontal space that corresponds to the variance explained by the component (that is, its eigenvalue), thereby letting the more influential components drive the visual impression.

In order to test whether these cumulative differences are statistical artefacts or if they represent something real, we have implemented both the *Kolmogorov-Smirnov* and the *Cramér-von Mises* test statistics, which are given by

$$\text{KS} = \max_{k=1,\dots,d} \left| \sum_{j=1}^{k} \lambda_j^x - \sum_{j=1}^{k} \lambda_j^y \right|, \quad \text{CvM} = \sum_{k=1}^{d-1} \frac{\lambda_k^z + \lambda_{k+1}^z}{2} \left( \sum_{j=1}^{k} \lambda_j^x - \sum_{j=1}^{k} \lambda_j^y \right)^2.$$

We conduct the tests as *permutation tests*, that is, by randomly reallocating the combined and separately standardized datasets into two new datasets of $n_x$ and $n_y$ observations, respectively, and then redoing the CE plot steps and recalculating the test statistics. This should be done a large (e.g. 10000) number of times. Then, a $p$-value is obtained by computing the proportion of reallocated datasets that lead to

5

test statistics at least as large as the one we found for the original datasets.

The permutation test results are also used to visualize the uncertainty of the CE curve in the plots. In the CE plots shown in the following section, we plot the observed curve together with 20 of the resampled curves, as well as a shaded region visualizing pointwise 95 % coverage intervals. If the observed curve is very different from the resampled curves or if it is substantially outside the shaded region, then this also indicates differences between the two datasets.

Two examples of cumulative eigenvalue plots are available in Figure 2. In the top panel, we see a cumulative eigenvalue curve that falls well within the confidence region, thereby suggesting no difference in eigenvalues (as expected). This also agrees with the large $p$-values and the fact that only a single underlying model was used for generating this data. In the bottom panel, the conclusion is reversed, which is also in accordance with the nature of the simulated data, which contains two distinct groups.

## 2.3. The angle plot

The angle plot compares both loadings and eigenvalues at once and it can be used to understand the information loss if the data structure of one dataset is superimposed on the other, thereby revealing which principal components (i.e. loading and eigenvalue pairs) are most similar and most different across the two datasets. Let $\lambda_{\max} = \max\{\lambda_1^x, \lambda_1^y\}$ be the largest eigenvalue for the two datasets. Then the empirical correlation matrix for the $\mathbf{X}$ dataset, $S_x$, has the following orthogonal decomposition in the coordinate system of the $\mathbf{Y}$ dataset

$$
S_x = \sum_{j=1}^d \lambda_j^x \eta_j^x (\eta_j^x)^\top = \lambda_{\max} \sum_{k=1}^d \left( \sum_{j=1}^d \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} \eta_j^y (\eta_j^{y\top} \eta_k^x) \right) \left( \sum_{j=1}^d \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} \eta_j^y (\eta_j^{y\top} \eta_k^x) \right)^\top,
$$

and we have a similar decomposition of $S_y$ in the coordinate system of the $\mathbf{X}$ dataset. We propose to visualize these two decompositions in a $d \times d$ grid display. In the $j$th row and $k$th column of this display we plot two arrows based at the lower left corner of the grid cell. The first arrow has length $\mu_{jk}$ and angle $\theta_{jk}/2$ counterclockwise from the diagonal, and the second arrow has length $\nu_{jk}$ and angle $\theta_{jk}/2$ clockwise from the diagonal. To facilitate the following description we will refer to the arrows drawn counterclockwise as the blue arrows, and the arrows drawn clockwise as the red arrows. The lengths $\mu_{jk}$ and $\nu_{jk}$, and the angle $\theta_{jk}$, are given by

$$
\mu_{jk} = \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} |\eta_k^{x\top} \eta_j^y|, \qquad \nu_{jk} = \sqrt{\frac{\lambda_j^y}{\lambda_{\max}}} |\eta_j^{y\top} \eta_k^x|, \qquad \theta_{jk} = \arccos(|\eta_k^{x\top} \eta_j^y|).
$$

For two $d$-dimensional, unit length vectors $a$ and $b$, it holds that $a^\top b = \langle a, b \rangle = \tilde{c}(a, b)$, where $\tilde{c}$ denotes the empirical correlation function. Thus, in the angle plot, we are essentially looking at the absolute values of correlations between loadings that have been scaled according to their variance contributions. The absolute value of the projection $\eta_k^{x\top} \eta_j^y$ is inserted due to the indeterminacy of the direction of loading vectors. This indeterminacy implies that the angle between loadings from the two datasets can always be chosen to be in the interval $[0, \pi/2]$, and hence the decomposition of $S_x$ and $S_y$ can be visualized in a joint plot by dividing the angles by two and using counterclockwise and clockwise shifts from the diagonal. Furthermore, the scaling of the lengths by $\lambda_{\max}$

6

is made so that the longest arrow has at most unit length.

In the *angle plot*, the blue arrows in the $k$th column of the grid display visualize the decomposition of the $k$th principal component of the **A** dataset in the coordinate system of the **B** dataset. Similarly, the red arrows in the $j$th row visualize the decomposition of the $j$th principal components of the **B** dataset in the coordinate system of the **A** dataset. If we have coinciding blue and red arrows along the diagonal in the grid display, and arrows of vanishing lengths in the off-diagonal cells, then the structures of the two datasets are identical. In contrast, differences in the eigenvalues are visualized as differences in the lengths of the blue and the red arrows, also in the diagonal. And loadings in other directions than the corresponding loading from the other dataset are visualized as angle separation of the blue and the red arrows in the diagonal cells, as well as arrows of non vanishing length in the off-diagonal cells.

However, there is a generic situation where the angles between the arrows lack discriminative power, namely when two or more neighboring eigenvalues are identical or close. We refer to this situation as *eigenvalue indeterminancy*. For instance, multiple eigenvalues close to 1 arise if more than one of the variables are uncorrelated, or close to uncorrelated, of all the other variables. In this situation only the corresponding eigenspace is well estimated, but the rotation of the associated eigenvectors will be subject to sampling variability. To investigate if we are in such a situation of identical data structures with neighbouring eigenvalues being close, where comparably long arrows may occur in the corresponding blocks of off-diagonal cells, we have implemented a permutation test similar as done for the CE-plot. For each random reallocation we recompute the angles $\theta_{jk}$. The resampled angles may both be used for quantitative tests, where large angles are critical for the diagonal cells and small angles are critical for the off-diagonal cells, as well as for the visualisation of the sampled angle regions. In the visualization we propose to present the entire sampled region of angles for each cell, but gray shaded according to percentiles such that more critical angles are lighter gray.

Figure 3 presents two examples of angle plots based on simulated data. For dataset **A**, we find very short off-diagonal arrows except for the off-diagonal cells corresponding to the 3rd and 4th eigenvalues. But since the arrows in cells (3,4) and (4,3) lie well within the shaded angle regions (supported by p-values 0.9119 and 0.9103) this is in agreement with identical data structures, where the 3rd and 4th eigenvalues suffer from indeterminancy (in this situation due to the two independent variables in *model 1*). For dataset **B**, on the other hand, we find very little agreement in all components, which should be expected, as the two groups in the dataset correspond to different data structures.

## 2.4. The chroma plot

The chroma plot is primarily an illustration of differences in loading patterns and it targets the question of how the roles of the original variables are different between the two datasets, thus leading the data structure comparison question back to its original, empirical context. The chroma plot consists of two panels, one for each dataset, made up of colored bars. These bars each represent a principal component and their coloring illustrates the relative weights of the $d$ original variables, that is, their absolute, normalized loading contributions. More specifically, when illustrating the $i$th principal component in the two data subsets, we plot vertical bars of length one that has been divided into $d$ segments of different colors. The width of the $j$th

colored segment is given by

$$\omega_{ij}^x = \frac{|\eta_{ij}^x|}{\sum_{k=1}^d |\eta_{ik}^x|}, \qquad\qquad \omega_{ij}^y = \frac{|\eta_{ij}^y|}{\sum_{k=1}^d |\eta_{ik}^y|},$$

where $\eta_{ij}^x$ and $\eta_{ij}^y$ denotes the $j$th entry of $\eta_i^x$ and $\eta_i^y$, respectively. Due to the indeterminacy of the loading signs, all the signs are removed from the coefficients in the loadings. The bars are ordered according to the eigenvalues and they are annotated with the cumulative percentage explained variance of that component, that is, the scaled and summed variance contributions,

$$\tilde\Sigma_i^x = \frac{\sum_{j=1}^i \lambda_j^x}{\sum_{k=1}^d \lambda_k^x} = \frac{\sum_{j=1}^i \lambda_j^x}{d}, \qquad\qquad \tilde\Sigma_i^y = \frac{\sum_{j=1}^i \lambda_j^y}{\sum_{k=1}^d \lambda_k^y} = \frac{\sum_{j=1}^i \lambda_j^y}{d}.$$

Especially when $d$ is large, we recommend plotting only a select set of interesting principal components (for instance identified by use of the angle plot). In this scenario, the annotations should rather be the non-cumulative variance contributions, $\tilde\sigma_i^x = \frac{\lambda_i^x}{d}$ and $\tilde\sigma_i^y = \frac{\lambda_i^y}{d}$. If there is eigenvalue indeterminancy for neighboring components, as discussed above, the loading patterns for these components will not have a meaningful interpretation. Therefore, one should not compare the chroma bars for component that were identified to suffer from eigenvalue indeterminancy by the angle plot.

In Figure 4, two examples of chroma plots are available. For dataset **A**, we see very similar visual patterns across the two panels for components 1, 2, 5 and 6, corresponding well with the fact that both groups are drawn from the same underlying model. Components 3 and 4 do not have similar loading patterns across the two subsets of the data, but this is expected since we identified these components to suffer from eigenvalue indeterminancy in the angle plot. For dataset **B**, we see different color compositions for the bars in the left- and the right panel, illustrating that the data structures are not identical for the two groups. The plots resulting from this procedure should be inspected focusing on two properties: Similarities in loading patterns, which will correspond to similar visual impressions, and similarities in variance contributions. For each component, the loadings describe how influential the different variables are on that component. Therefore, the chroma plot allows us to make qualitative statements about the original datasets, such as *"variable 1 is generally more influential in subset A than it is in subset B"*, thereby helping us to understand where and why potential data structure differences are found.

### 2.5. A data structure comparison workflow

Finally, we turn to a few comments about how the three plots presented thus far can be used together. We suggest that the plots are applied in the order in which they were presented, thereby moving from overall assesments of similarity to explanations about the nature of potential data structure differences. This workflow is summarized in Figure 5 and should follow three steps:

(1) Make a CE plot. If this plots suggests similarity in eigenvalues then it is appropriate to move on to the angle plot to assess similarity in loadings. Otherwise, it can be concluded that the structures of the two datasets are not similar.

(2) Make an angle plot. If this plots suggests similarities in both eigenvalues and

8

loadings, then the data structure comparison exploration is concluded. If the angle plots suggests differences in data structures, identify which components are responsible for the differences.

(3) Make a chroma plot of the components selected to be different by the angle plot, but whose differences are not due to eigenvalue indeterminancy among neighboring components. Use this information to discuss *how* and *why* the data structures might be different.

We will now turn to a data example, showing how the PCADSC tools can be used in practice.

## 3. Application

We now illustrate the three PCADSC tools using data from the 2012 version of the European Social Survey (ESS) project to investigate inter-country differences in psychological well-being and happiness. All computations and figures presented in this section were created using the R package PCADSC [15].

### *3.1. Data*

The ESS 2012 data contain a total of 626 variables collected from 54673 citizens of 29 countries. Here, we will only work with a subset of 35 questionnaire items that are all related to psychological well-being. These 35 items can be divided into 6 distinct scales, namely *Evaluative well-being, Emotional well-being, Functioning, Vitality, Community well-being*, and *Supportive relationships*. More details on these scales can be found in [8] and the relationship between questionnaire items and scales is summarized in Appendix C. We represent each of the scales by a single variable, which is calculated as the average score within the items related to that variable and scaled such that it takes a value between 0 and 10. For simplicity, we use only complete cases for this construction and thus exclude all participants that did not answer all the 35 questionnaire items used below. This corresponds to approximately 9% of the observations in the Danish sample, 20% in the Bulgarian sample and only 6% in the Swedish sample. All in all, we have $n_{DK} = 1498$ complete case observations in the Danish sample, $n_{BG} = 1798$ observations in the Bulgarian sample and $n_{SE} = 1736$ Swedish observations. Figure 1 summarizes the marginal distributions of the six dimensions of psychological well-being, stratified by country.

### *3.2. Comparing Denmark and Bulgaria*

Figure 6 presents the CE plot and the angle plot obtained from comparing the Danish and Bulgarian psychological well-being scales. The CE plot show a remarkable degree of lacking comparability: The cumulative differences in the eigenvalues by far exceed what could come about randomly if there really were no difference in the data structures. This is also confirmed by the Kolmogorov-Smirnov and the Cramér-von Mises tests, which both result in *p*-values that are virtually zero. The CE curve generally lies above zero, suggesting that the Bulgarian dataset has larger eigenvalues for all but the last component, thereby explaining more variance at lower dimensionalities.

Moving on to the angle plot, we find that the differences are primarily to be found in the second, third, and fourth principal components (PCs). The blue arrows visualize the

9

decomposition of the principal components for the Bulgarian dataset in the coordinate system of the Danish dataset. We see that PC2 also loads on PC3, that PC3 also loads on PC4, and that PC4 also loads on PC2 and PC3. The red arrows visualize the decomposition of the principal components for the Danish dataset in the coordinate system of the Bulgarian dataset. Here, we see that PC2 also loads on PC4, that PC3 also loads on PC2 and PC4, and that PC4 also loads on PC3. Thus, if we wish to understand why differences in the data structures occur, an inspection of the loadings of components 2, 3 and 4 might be informative.

The chroma plot in Figure 7 allows us to look closer into these components. Here, we find that the relative importance of the *Community well-being* and *Supportive relationships* scales is much larger in the Bulgarian sample than in the Danish. In the Danish data, on the other hand, we find that *Vitality* and *Emotional well-being* seem to play bigger roles, as they appear with larger loadings in more high-ranking components in this sample, relative to the Bulgarian.

All in all, we find that psychological well-being does *not* seem to be the same concept in Bulgaria and Denmark. The two countries disagree both in how many dimensions are needed to capture the most important parts of the concept (as illustrated by the differences in eigenvalues) and in how these dimensions are then weighted among the 6 scales (as illustrated by the angle- and chroma plots). In Bulgaria, interpersonal features seem to be more informative of psychological well-being, whereas in Denmark, individual characteristics play a relatively larger role, corresponding well with the previous findings mentioned above. Thus, the datasets are fundamentally different and we should therefore be wary about combining them in a joint analysis, which was also the conclusion of the ESS authors, though based on country-level aggregated statistics [8]. Moreover, this also implies that two countries cannot be ranked in terms of which country is "the most happy", at least not by referring to psychological well-being dimensions such as those encountered here.

### 3.3. Comparing Denmark and Sweden

We now turn to the comparison of Denmark and Sweden in terms of psychological well-being. Figure 8 shows the CE- and angle plots for these two countries. In the CE plot, we now find the cumulative eigenvalue curve to be just within the acceptance region of the null-hypothesis. This is also reflected by the two tests, which now produce $p$-values of $p_{KS} = 0.17$ and $p_{CvM} = 0.11$, respectively. Thus, it is not unreasonable to assume equal eigenvalues in the two datasets.

The angle plot in Figure 8 shows that the two datasets agree very strongly about the relative importance of the six scales in the six PCs, as almost all off-diagonal arrows are practically non-existent. This implies that if one already has the information held in the first PC from the Danish data, this information is in itself mostly sufficient to describe the first PC of the Swedish data.

Looking at the chroma plot in Figure 9, the same tale is told once again: Here, we find remarkably similar loading patterns in the first three components (which are responsible for almost 80% of the variance in both datasets), and slight, but increasing, differences in the remaining three components. We therefore conclude that any differences in the data structures of the Danish and the Swedish samples are related to the least important dimensions of the datasets and that these dimensions are only responsible for less than 25% of the variance in both datasets. In particular, this means that we can combine and compare the Danish and Swedish datasets in a meaningful way and

10

for instance conclude using Figure 1 that in general, Danes seem to be somewhat more happy than Swedes, and in particular that the least happy people in Denmark (represented by the 1st quartiles) are generally happier than the least happy people in Sweden. A more thorough, statistical investigation could now be put to work on answering *why* this seems to be the case.

## 4. Conclusion and discussion

Three new tools, referred to collectively as Principal Component Analysis-based Data Structure Comparisons (PCADSC), for the task of deciding if two datasets can be combined for analysis were proposed and discussed. They all employ the principal component decomposition of the empirical covariance matrix performed on two subsets of a dataset in order to create intuitive visualizations of data structure differences.

In an analysis of data from the 2012 version of the European Social Survey (ESS) project, we illustrated that the PCADSC tools help inform analyses of inter-country differences in psychological well-being and happiness. It should be noted that even though concerns about pooling data from different countries are not yet very prevalent when it comes to happiness rankings, the subject has been very well scrutinized in the field of international educational rankings. For instance, the PISA tests have repeatedly been criticized for not being meaningful objects of international comparisons, especially due to problems with differential item functioning [9, 10, 19] and translation problems [2]. This highlights lack of data comparability as an empirically relevant and concerning issue that there is no way around: When conducting rankings, the issue of data structure heterogeneity should always be addressed.

Usually, when using PCA, the main interest lies in the *scores*, i.e. the coordinates $\eta_j^\top (x_i - \bar{x})$ of the projections observations onto the loadings, but the eigenvalues and the accompanying loadings are in fact more suitable objects for assessing data structure similarity. If two dataset subsets have similar loadings, then they can be argued to measure the same underlying quantities. Similar loadings can occur in situations where the two sets of scores are different, for example if the two datasets come from two different populations of subjects. On the other hand, if the loading patterns are different in the two subsets, then this indicates that their respective variable interplays differ, and hence it would be criticizable to use these variables for comparisons across the two subsets. The PCADSC tools are novel in their emphasis on the information held in the loadings rather than the scores.

Though the PCADSC toolbox comprises three diagnostic plots as of now, the PCA decomposition can of course be utilized further to bring about new PCADSC plots that illustrate more aspects of data structure differences. For instance, the scree plot (as introduced by Cattell [4]), which plots the ordered eigenvalues against their component numbers, could easily be recreated as a PCADSC tool by creating scree plots for both data subsets in the same coordinate system, yielding a simple graphical comparison. The resampling approach used in the CE plot could then be adopted, thus annotating the scree plots with resampled curves. This general approach to model-fit evaluation is very helpful in combination with permutation tests, as inspired by Lin and colleagues [11].

The proposed methods also have limitations. First of all, as PCA is performed after standardization of the variables, the PCADSC methods cannot detect differences in neither mean values nor in marginal variances - this information is thrown away as the very first step. However, such differences are of a fundamentally different nature

11

than those we have discussed so far. Differences in mean values are often the main interest of the analysis and should of course not be regarded as a data comparability issue. Differences in marginal variances, on the other hand, can pose some modeling challenges, but none that are not generally solvable by use of random effects models. We therefore do not consider this property of PCADSC to be a major drawback.

A more prominent limitation is the fact that the PCASDC methods are only valid for variables that have a numerical interpretation. In particular, this means that the methods cannot be used for data with nominal categorical variables. This limitation is inherited from the PCA method and thus, moving beyond it will entail replacing the PCA framework with a more general one. Multiple correspondence analysis (MCA) is often suggested as a categorical extension of PCA [1] and including MCA in the PCADSC tools is thus a natural next step.

Moreover, though we have illustrated the discriminatory power of PCASDC in one simulated- and on real-life data example, of course further evaluation of its performance is needed. While the best feature of PCADSC is perhaps the intuitive, visual nature of the tools, it is also the biggest weakness. For statistical methods that result in a number (or a set of numbers), the standard way of evaluating performance is to report on simulation studies, but in the case of statistical methods that report graphics this is not feasible. Therefore, a systematic evaluation of the performance of the PCADSC methods is not very straight-forward to design: It really depends on whether or not users can detect actual differences in data structures by using the diagnostic plots. We suggest that the variability of the methods could be studied further using *Wally plots* [5], which are learning tools developed for teaching students how to interpret residual plots. First in line as a topic for more thorough investigations of performance is the sensitivity towards the sample sizes of the two datasets, $n_1$ and $n_2$.

A different aspect of the data size, namely the number of variables $d$, may also limit the useability of the PCADSC tools. While the CE plot scales well with large $d$, the angle plot and the chroma plot can become difficult to read if there are a lot of variables and thus a large number of principal components. However, because the angle plot lets the arrow length depend on the size of the corresponding eigenvalue, its visual impression should help guide a user towards influential *and* dissimilar components even though $d$ is large. And when these components have been identified, the chroma plot can focus only on them, thereby avoiding plotting the loading patterns of too many components.

All in all, the PCADSC methods represent a first step towards addressing a known issue that has otherwise been met with unsatisfactory ad-hoc methods whose assumptions rest upon the very hypothesis they are testing. The PCADSC tools perform quite well in the data example provided here, yielding conclusions that are in accordance with well-established theories and they are simple to use and test further using the PCADSC R package.

### References

[1] H. Abdi and L.J. Williams, *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics 2 (2010), pp. 433–459. Available at http://dx.doi.org/10.1002/wics.101.

[2] M. Asil and G.T.L. Brown, *Comparing OECD PISA Reading in English to Other Languages: Identifying Potential Sources of NonInvariance*, International Journal of Testing 16 (2016), pp. 71–93.

12

[3] D.J. Brambilla and S.M. McKinlay, *A comparison of responses to mailed questionnaires and telephone interviews in a mixed mode health survey.*, American journal of epidemiology 126 (1987), pp. 962–71. Available at http://www.ncbi.nlm.nih.gov/pubmed/3661543.

[4] R.B. Cattell, *The Scree Test For The Number Of Factors*, Multivariate Behavioral Research 1 (1966), pp. 245–276.

[5] C.T. Ekstrøm, *Teaching 'Instant Experience' with Graphical Model Validation Techniques*, Teaching Statistics 36 (2014), pp. 23–26. Available at http://doi.wiley.com/10.1111/test.12027.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., chap. 14.5.1, Springer (2009).

[7] J. Helliwell, R. Layard, and J. Sachs, *World Happiness Report 2016 Update*, New York: Sustainable Development Solutions Network (2016).

[8] K. Jeffrey, S. Abdallah, and A. Quick, *Europeans' Personal and Social Wellbeing: Topline Results from Round 6 of the European Social Survey*, ESS Topline Results (Series 5) (2015).

[9] M. Kankaras and G. Moors, *Analysis of Cross-Cultural Comparability of PISA 2009 Scores*, Journal of Cross-Cultural Psychology 45 (2014), pp. 381–399.

[10] S. Kreiner and K.B. Christensen, *Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy*, Psychometrika 79 (2014), pp. 210–231.

[11] D.Y. Lin, L.J. Wei, and Z. Ying, *Model-Checking Techniques Based on Cumulative Residuals*, Biometrics 58 (2002), pp. 1–12.

[12] M. Liu, *Comparing data quality between online panel and intercept samples*, Methodological Innovations 9 (2016), p. 2059799116672877. Available at http://dx.doi.org/10.1177/2059799116672877.

[13] H.L. Lolle and J.G. Andersen, *Measuring Happiness and Overall Life Satisfaction: A Danish Survey Experiment on the Impact of Language and Translation Problems*, Journal of Happiness Studies 17 (2016), pp. 1337–1350. Available at http://dx.doi.org/10.1007/s10902-015-9646-4.

[14] C.A. McHorney, M. Kosinski, and J.E. Ware, *Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey.*, Medical care 32 (1994), pp. 551–67. Available at http://www.ncbi.nlm.nih.gov/pubmed/8189774.

[15] A.H. Petersen and B. Markussen, *PCADSC: Tools for Principal Component Analysis-Based Data Structure Comparisons* (2017). Available at https://CRAN.R-project.org/package=PCADSC, R package version 0.8.0.

[16] J.R. Powers, G. Mishra, and A.F. Young, *Differences in mail and telephone responses to self-rated health: use of multiple imputation in correcting for response bias.*, Australian and New Zealand journal of public health 29 (2005), pp. 149–54. Available at http://www.ncbi.nlm.nih.gov/pubmed/15915619.

[17] A. Rodríguez-Pose and V. von Berlepsch, *Social Capital and Individual Happiness in Europe*, Journal of Happiness Studies 15 (2014), pp. 357–386.

[18] R. Veenhoven, *Cross-national differences in happiness: Cultural measurement bias or effect of culture?*, International Journal of Wellbeing 2 (2012), pp. 333–353.

[19] R.J. Zwitser, S.S.F. Glaser, and G. Maris, *Monitoring Countries in a Changing World: A New Look at DIF in International Surveys*, Psychometrika 82 (2017), pp. 210–232.

## A. A brief introduction to principal component analysis

In this appendix we introduce PCA. We suppress the subscripts and superscripts used in the article because we only consider a single dataset here, denoted by $\mathbf{X} \in \mathbb{R}^{d \times n}$.

For a given $q \leq d$, principal component analysis (PCA) is a tool for finding a new representation of the dataset of dimension $q$ such that the least possible amount of

13

information is lost. More specifically, we wish to minimize the loss when looking at a projection of $\mathbf{X}$ onto a $q$-dimensional space, rather than the original $d$-dimensional one. Formally, we define the *rank-q-reconstruction error* as the minimal squared error that is achievable by linear subspaces $K_q \subset \mathbb{R}^d$ of dimension $q < d$, that is

$$\min_{K_q} \sum_{i=1}^n \min_{z \in K_q} \|x_{i\cdot} - \bar{X} - z\|^2 = \min_{K_q} \sum_{i=1}^n \|x_{i\cdot} - \bar{X} - \mathrm{proj}_{K_q}(x_{i\cdot} - \bar{X})\|^2.$$

The theory of *Principal component analysis* (PCA) not only ensures the existence of a subspace $\hat{K}_q \subset \mathbb{R}^d$ that attains this minimum, it also provides an explicit description of $\hat{K}_q$ and the rank-q-reconstruction error [6]. More specifically, the rank-q-reconstruction error is attained when we choose

$$\hat{K}_q = \mathrm{span}\{\eta_1, \ldots, \eta_q\},$$

where $\eta_1, \ldots, \eta_q$ are the first $q$ eigenvectors of the empirical covariance matrix, $S$, as ordered by the size of their associated eigenvalues, $\lambda_1, \ldots, \lambda_d$. In the PCA framework, we refer to the eigenvectors as *loadings*. The eigenvalues may be understood as *variance components*, as the sum of the marginal empirical variances is preserved under eigenvalue decomposition, that is,

$$\mathrm{trace}(S) = \sum_{j=1}^d \hat{V}(X_j) = \sum_{j=1}^d \hat{V}(\eta_j^\top \mathbf{X}^\top) = \sum_{j=1}^d \lambda_j,$$

where $X_j \in \mathbb{R}^n$ denotes the $j$th variable of $\mathbf{X}$ and $\hat{V}$ is the empirical variance function. This again emphasizes that we do not change the covariance structure of a dataset when performing PCA; we merely use linear algebra to make it easier to describe. And as eigenvalues are uniquely defined, and eigenvectors are uniquely defined up to a change of sign whenever the eigenvalues are distinct, the representation hereby obtained is a valid object of inference.

The $j$th loading can also be found iteratively as the unit vector $u \in \mathbb{R}^d$ orthogonal to $\hat{K}_{j-1}$ that maximizes the variation of the associated scores:

$$\eta_j = \operatorname*{argmax}_{u \in \mathbb{R}^d \,:\, u \perp \hat{K}_{j-1}} \sum_{i=1}^n \|u^\top(x_{i\cdot} - \bar{X})\|^2, \qquad \lambda_j = \frac{1}{n-1} \sum_{i=1}^n \|\eta_j^\top(x_{i\cdot} - \bar{X})\|^2,$$

where the initial subspace is defined as $\hat{K}_0 = \{0\}$. It is worth emphasizing that this greedy approach of successively adding the next direction $\eta_j$ explaining most of the remaining variation also gives the sequence $\hat{K}_q = \hat{K}_{q-1} \oplus \mathrm{span}\{\eta_q\}$ of subspaces minimizing the rank-q-reconstruction error.

At this point a general remark about standardization is in place. PCA deconstructs the covariance matrix in components according to the most explained variance. This implies that if a variable has a very large sample variance (possibly because of its scale), this variable will always be deemed highly influential, no matter the structure of the data. Therefore, the variables are typically standardized prior to performing PCA. The covariance matrix of the standardized variables is the same as the correlation

14

matrix of the original variables, so post-standardization PCA simply corresponds to performing data structure comparisons of the correlation matrices rather than the covariance matrices. The standardization makes the variables comparable on the same scale, namely units of standard deviation. We have adopted this standardization policy.

## B. Details about the simulated data

This supplementary appendix describes the origin of the simulated data used for presenting the PCADSC plots. Dataset **A** consisted of 1000 independent simulated realizations from the $N(0, \Sigma_1)$-distribution, while dataset **B** consisted of 500 independent realizations from $N(0, \Sigma_1)$ and 500 independent realizations from $N(0, \Sigma_2)$. Dataset **A** was furthermore appended with a grouping variable, randomly dividing the observations into two groups. Dataset **B** also included a grouping variable, and this variable contained information about which of the two normal distributions each observation was simulated from. The covariance matrices, $\Sigma_A$ and $\Sigma_B$, were defined by

$$
\Sigma_A = \begin{pmatrix}
1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 1.0 & 0.7 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.7 & 1.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.4 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.4 & 1.0
\end{pmatrix}
$$

and

$$
\Sigma_B = \begin{pmatrix}
1.0 & 0.2 & 0.1 & 0.0 & 0.0 & 0.0 \\
0.2 & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 \\
0.1 & 0.1 & 1.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 1.0 & 0.3 & 0.1 \\
0.0 & 0.0 & 0.0 & 0.3 & 1.0 & 0.2 \\
0.0 & 0.0 & 0.0 & 0.1 & 0.2 & 1.0
\end{pmatrix},
$$

respectively.

## C. ESS psychological well-being scale definitions

**Table C1.**  ESS psychological well-being scale definitions. Relationship between questionnaire items and scales, as defined in [8]. Before we construct the scale scores as item means, we transform the individual item scores such that they are all on a scale from 0 to 10 and such that 10 always corresponds being the most happy.

| Scale | Items |
| --- | --- |
| Evaluative well-being | How satisfied with life as a whole |
| | How happy are you |
| Emotional well-being | Felt sad, how often in the past week |
| | Felt depressed, how often in the past week |
| | Enjoyed life, how often in the past week |
| | Were happy, how often in the past week |
| | Felt anxious, how often in the past week |
| | Felt calm and peaceful, how often in the past week |
| Functioning | Free to decide how to live my life |
| | Little chance to show how capable I am |
| | Feel accomplishment from what I do |
| | Interested in what you are doing |
| | Absorbed in what you are doing |
| | Enthusiastic about what you are doing |
| | Feel what I do in life is valuable and worthwhile |
| | Have a sense of direction |
| | Always optimistic about my future |
| | There are lots of things I feel I am good at |
| | In general feel very positive about myself |
| | At times feel as if I am a failure |
| | When things go wrong in my life it takes a long time to get back to normal |
| | Deal with important problems |
| Vitality | Felt everything did as effort, how often in the past week |
| | Sleep was restless, how often in the past week |
| | Could not get going, how often in the past week |
| | Had a lot of energy, how often in the past week |
| Community well-being | Most people can be trusted |
| | People try to take advantage |
| | Most of the time people are helpful |
| | Feel people in local area help one another |
| | Feel close to the people in local area |
| Supportive relationships | How many with whom you can discuss intimate matters |
| | Feel appreciated by those you are close to |
| | Receive help and support |
| | Felt lonely, how often in the past week |

16

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

**Table 1.** Median values of each of the six dimensions of psychological well-being, stratified by country. 25 and 75 percentiles are listed in parentheses. The scales are constructed such that they all run from 0-10.

|  | Denmark | Bulgaria | Sweden |
| --- | --- | --- | --- |
| Evaluative well-being | 8.75 (8.00, 9.50) | 5.00 (3.50, 7.00) | 8.00 (7.00, 9.00) |
| Emotional well-being | 8.33 (7.22, 8.89) | 6.67 (5.00, 7.78) | 7.78 (6.67, 8.89) |
| Functioning | 7.57 (6.93, 8.21) | 6.68 (5.50, 7.68) | 7.04 (6.39, 7.68) |
| Vitality | 7.50 (6.67, 8.33) | 7.50 (5.83, 8.33) | 7.50 (6.67, 9.17) |
| Community well-being | 6.77 (5.83 ,7.57) | 4.67 (3.70, 5.70) | 6.57 (5.66, 7.37) |
| Supportive relationships | 8.25 (7.42, 8.92) | 7.25 (6.17, 8.08) | 8.25 (7.42, 8.75) |

17
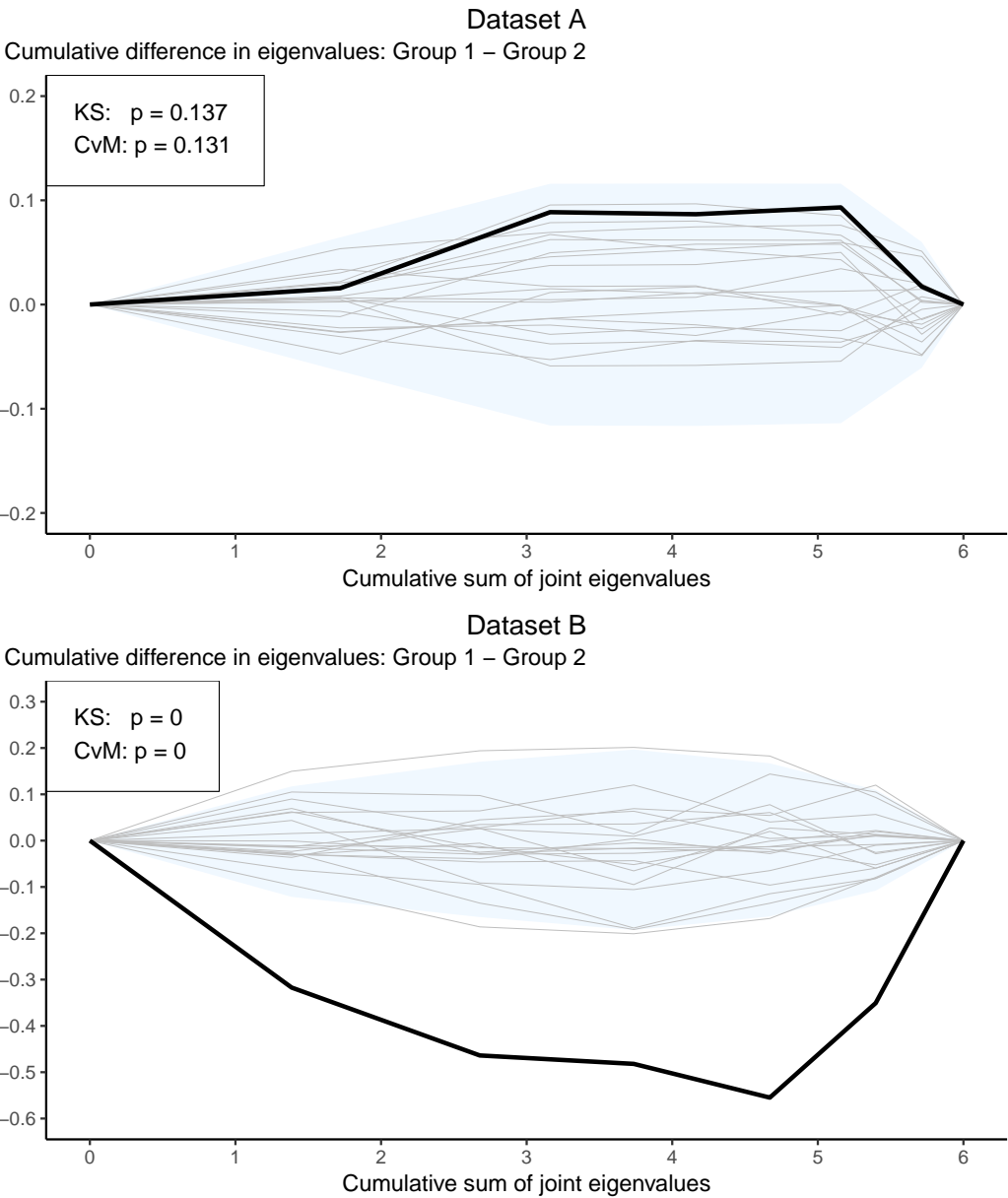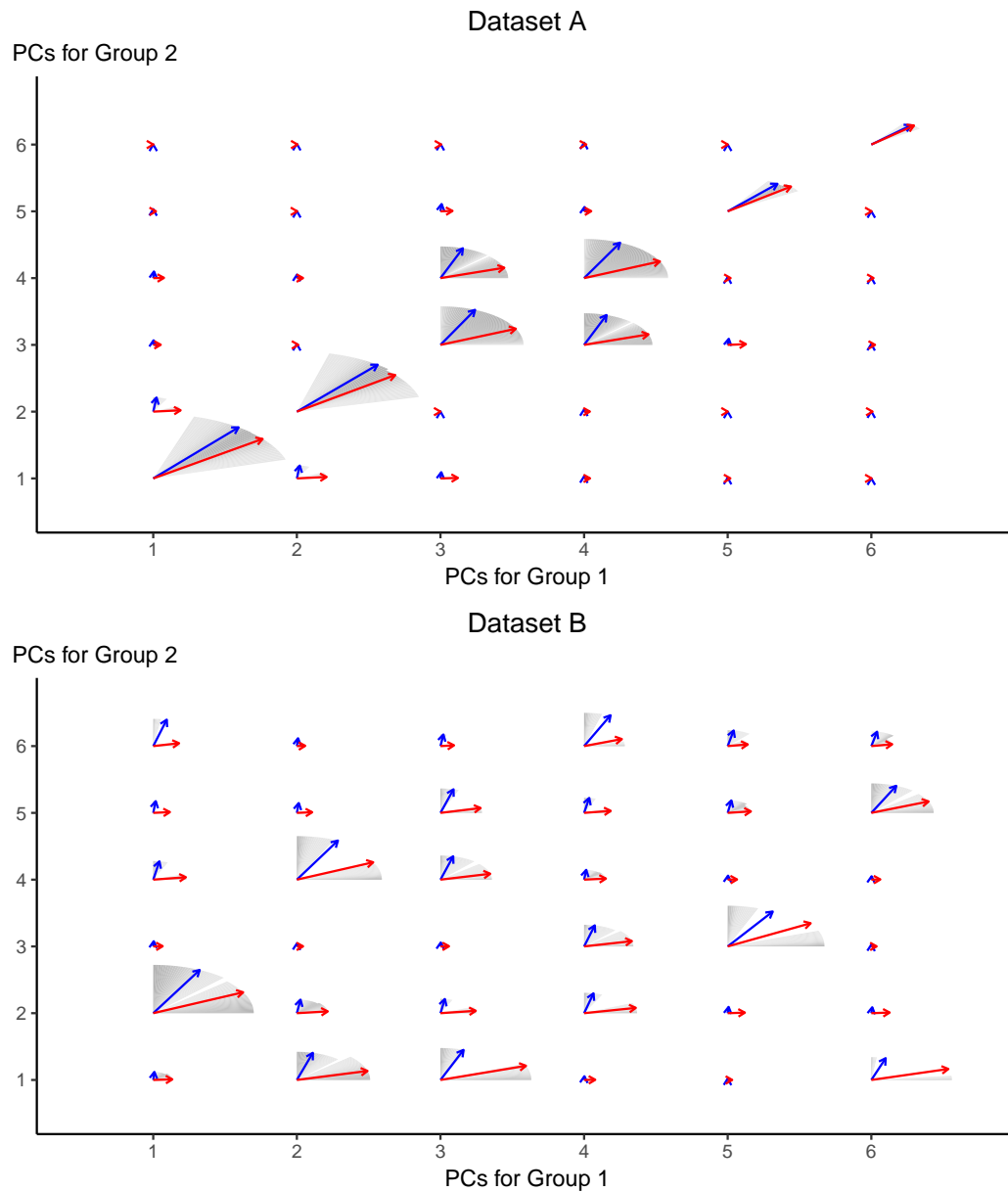
Model 1:



Model 2:



**Figure 1.**

Figure 2.

## Dataset A

PCs for Group 2

PCs for Group 1

## Dataset B

PCs for Group 2

PCs for Group 1

**Figure 3.**

20

**Figure 4.**

**Figure 5.**

22

Cumulative difference in eigenvalues: BG – DK



PCs for DK



**Figure 6.**

**Figure 7.**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
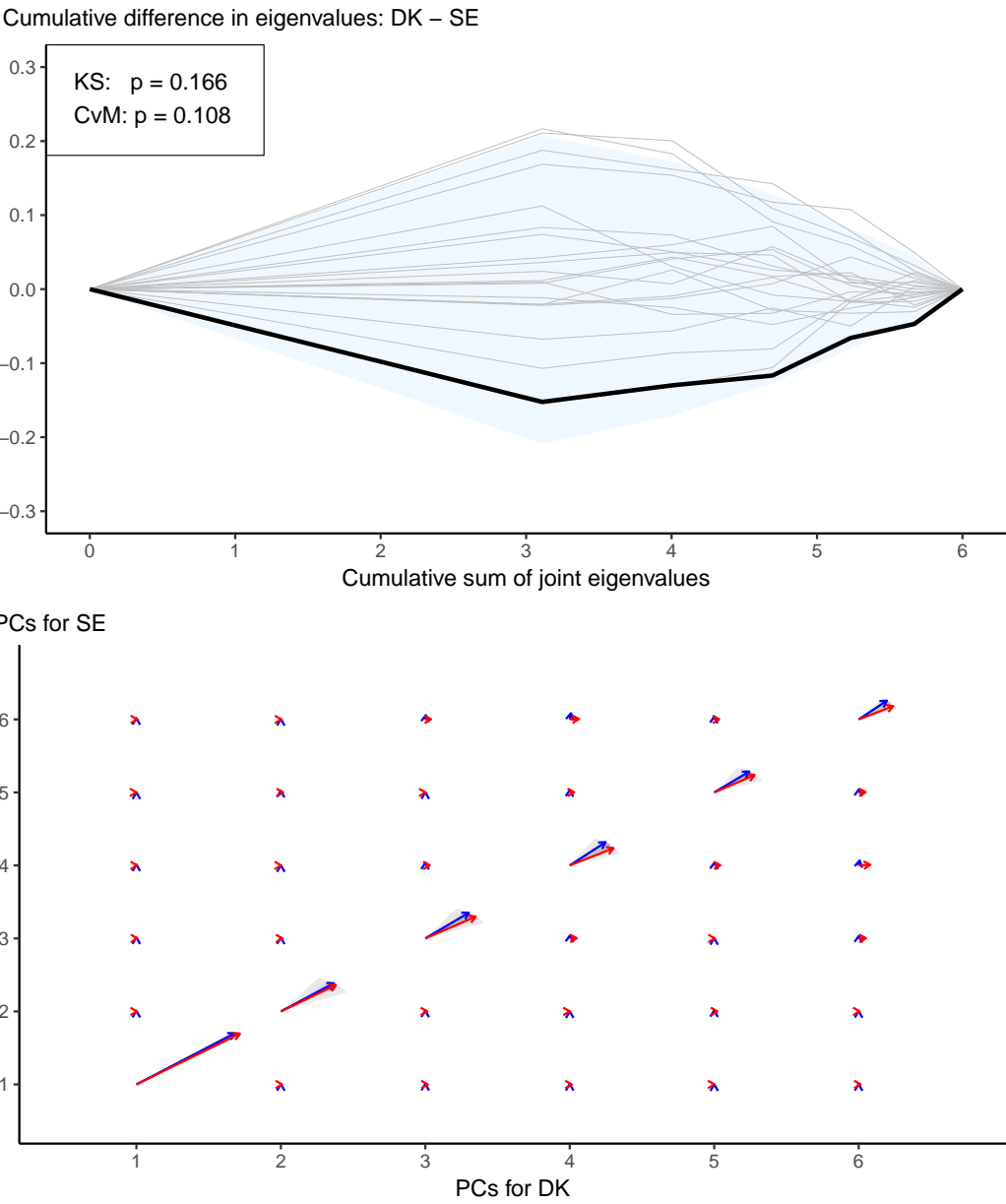50
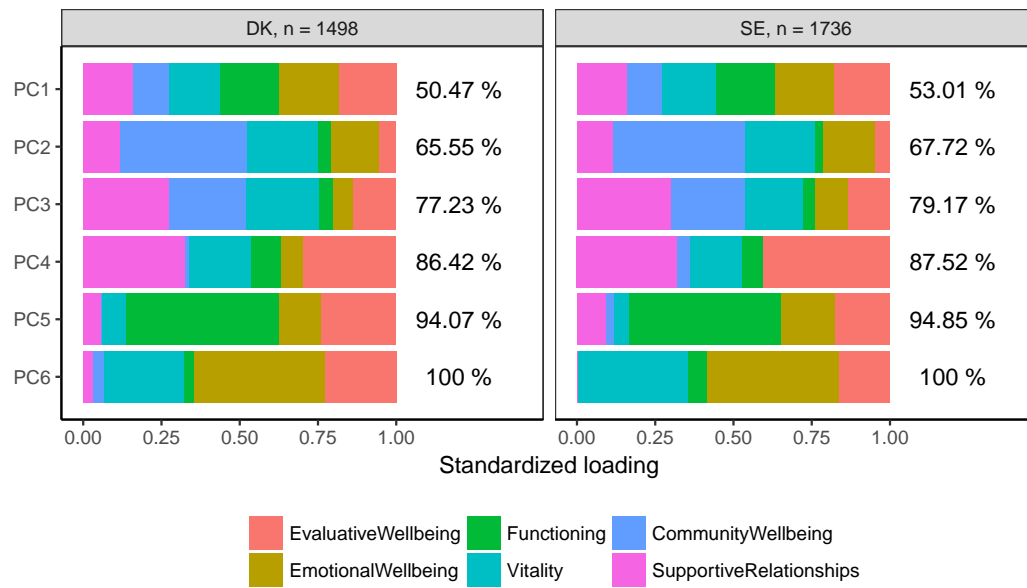51
52
53
54
55
56
57
58
59
60
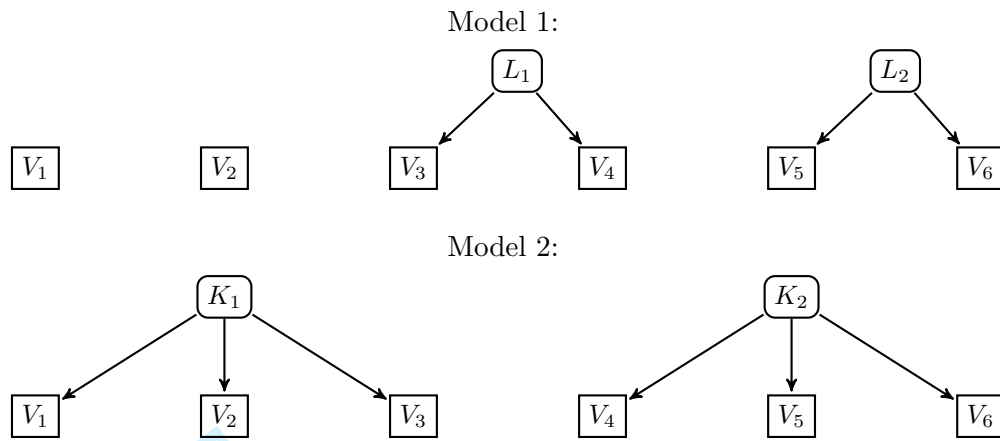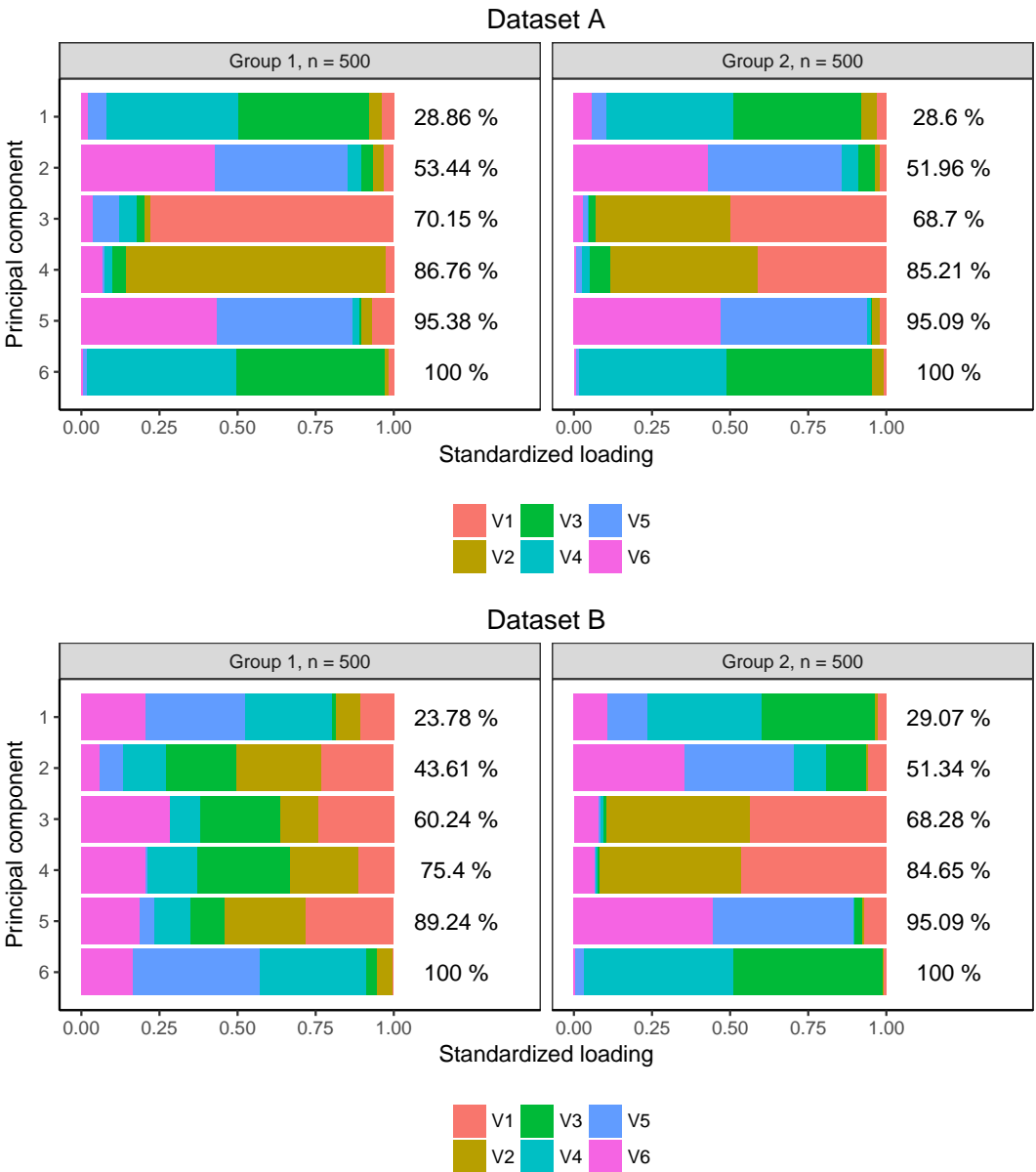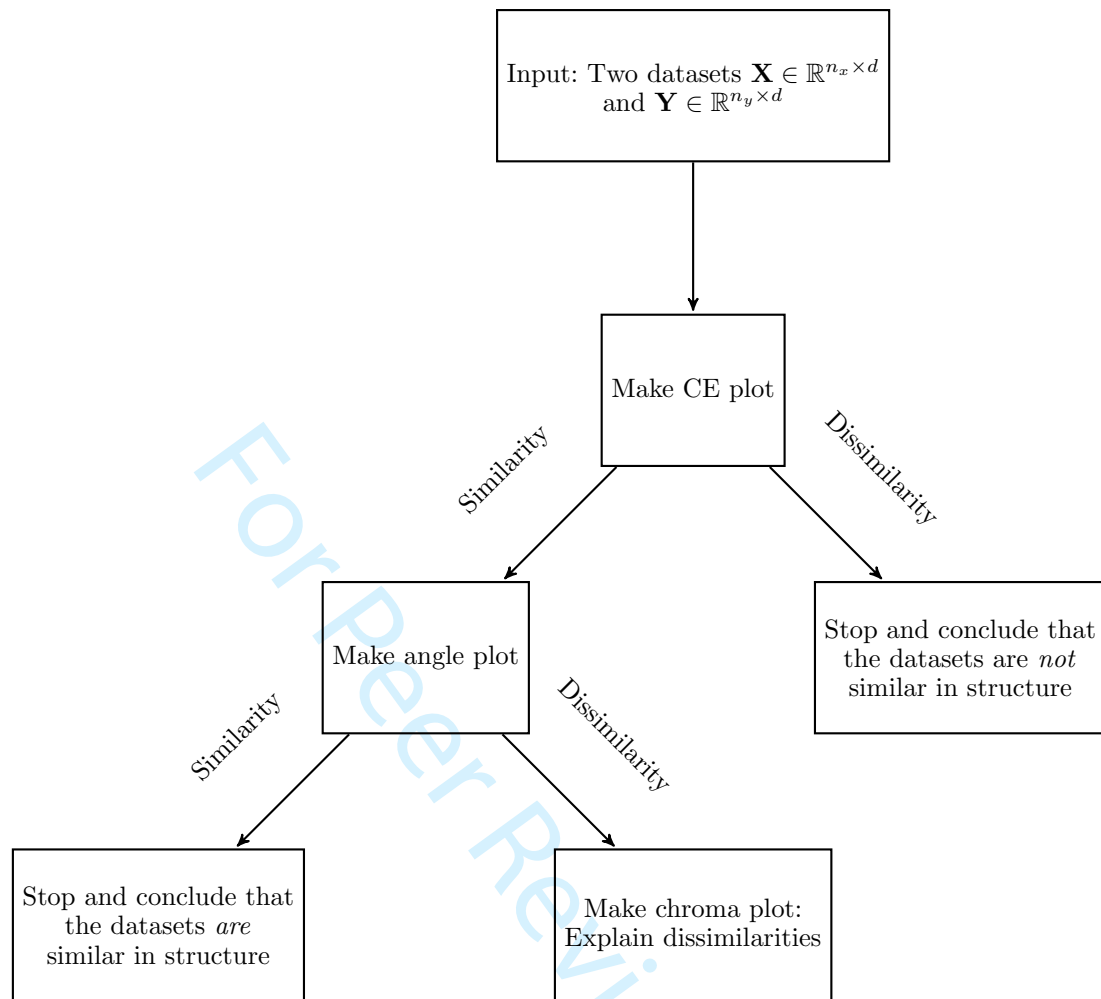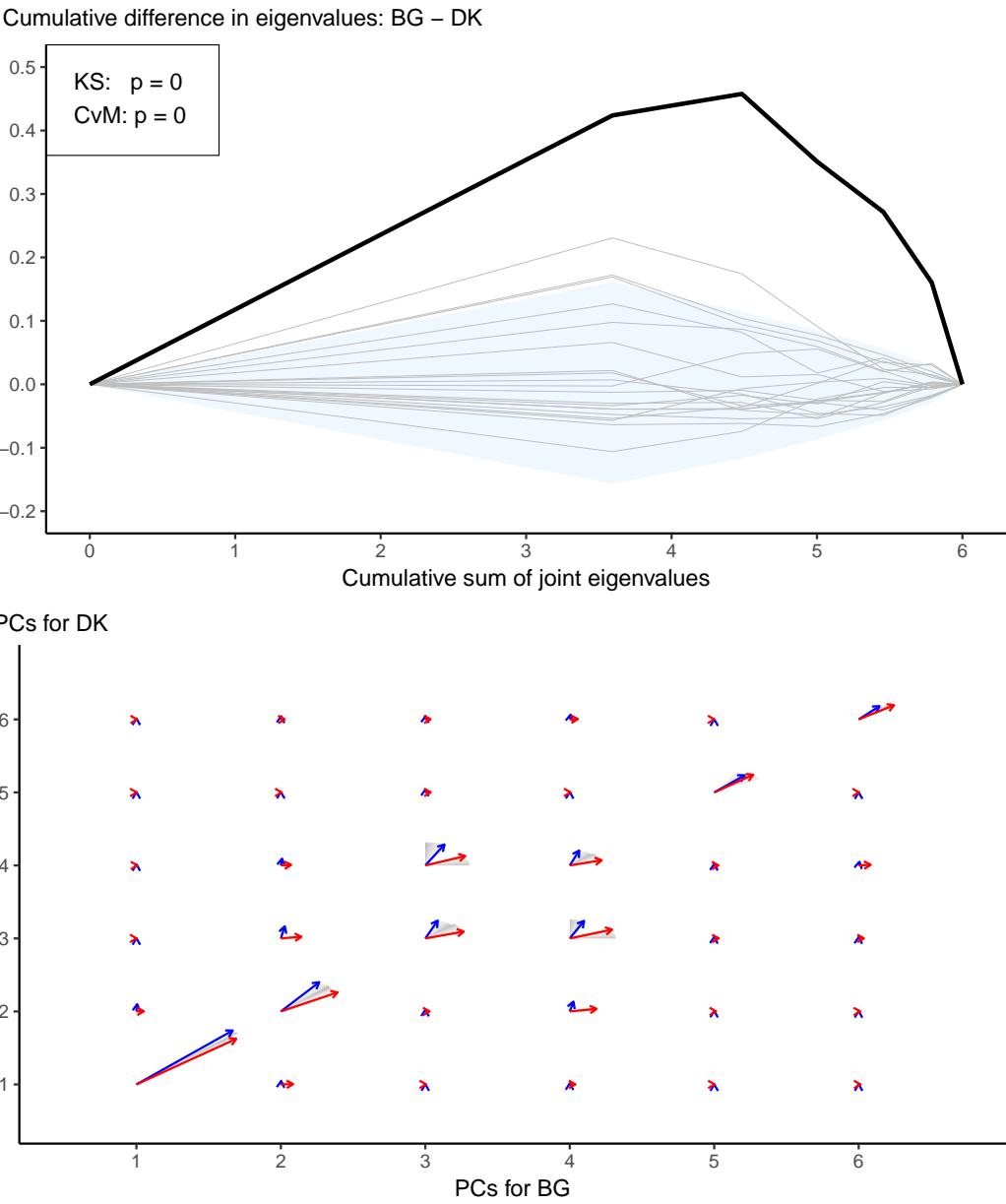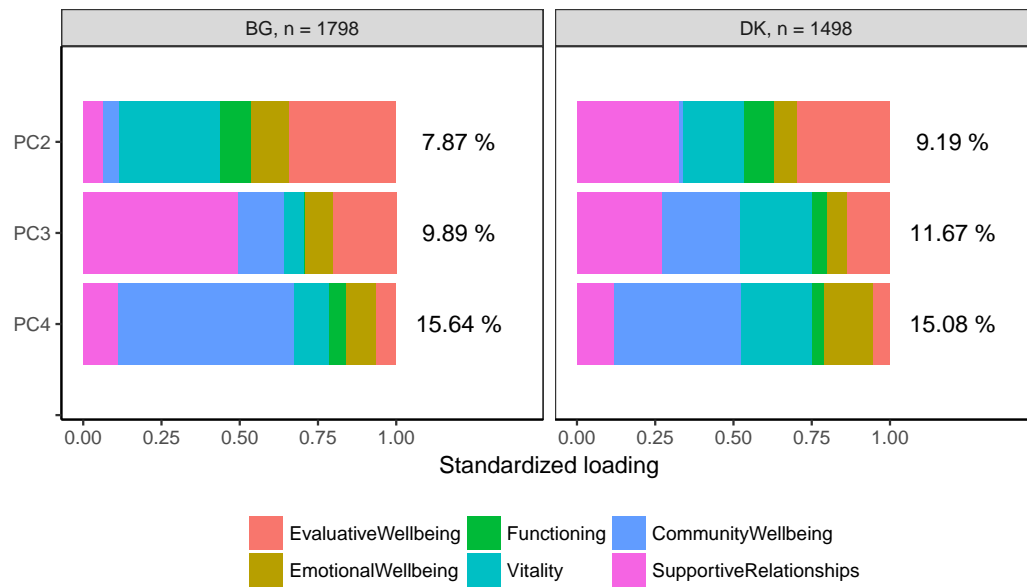


**Figure 8.**

Figure 9.

**Figure 1** Graphical representations of models used for simulating example data. Square nodes represent observed variables, while rounded nodes represent latent variables. Arrows are used to illustrate causal links.

**Figure 2** CE plots for the simulated datasets. Dataset A (with no differences in data structure) in top panel, and dataset B (where observations stem from different models) in bottom panel. Plots are annotated with the $p$-values of the Kolmogorov-Smirnov (KS) and the Cramér-von Mises (CvM) tests of the null hypothesis of no difference in data structures. The bold, black lines illustrate the observed cumulative eigenvalue differences, while the shaded areas are 95 % confidence bands under the null hypothesis of no difference in data structures.

**Figure 3** Angle plots for the simulated datasets. Dataset A in top panel and dataset B in bottom panel. Blue arrows show the principal components of the observations in group 1 decomposed in the coordinate system of the principal components of group 2, while the red arrows illustrate the reverse.

**Figure 4** Chroma plots for the simulated datasets. Dataset **A** in top panel and dataset **B** in bottom panel. Each bar represents a principal component and the bars are annotated with their cumulative percentages of explained variance.

**Figure 5** The PCADSC workflow, starting with two datasets and then successively applying the three visual tools for obtaining a deeper and deeper understanding of similarities and dissimilarities in the data structures.

**Figure 6** CE plots (top panel) and angle plots (bottom panel) comparing Danish and Bulgarian data on psychological well-being. The CE plot is annotated with the $p$-values of the Kolmogorov-Smirnov and the Cramér-von Mises tests of the hypothesis of no difference in data structures. In the angle plot, the blue arrows show the principal components of the Bulgarian dataset decomposed in the coordinate system of the principal components of the Danish dataset, while the red arrows illustrate the reverse.

**Figure 7** Chroma plot for comparing Danish and Bulgarian data on psychological well-being. A chroma plot comparing the 2nd, 3rd, and 4th principal components of the Bulgarian- and Danish psychological well-being data. The component-bars are annotated with their relative variance contributions.

**Figure 8** A CE plot (top panel) and an angle plot (bottom panel) comparing Danish and Swedish data on psychological well-being. The blue arrows show the principal components of the Danish dataset decomposed in the coordinate system of the principal components of the Swedish dataset, and the red arrows illustrate the reverse.

**Figure 9** Chroma plot comparing the loading patterns of the Danish and the Swedish subsamples. For each component, the bar is annotated with its cumulative variance contribution, that is, how much variance can be explained by having information of this and the preceding components.

27