EXPLORATIVE DATA STRUCTURE COMPARISON WITH APPLICATION TO ...

Anne H. Petersen, Bo Markussen, Karl Bang Christensen
university of copenhagen

Abstract

abstract...

Key words: keywords...

1. Introduction

Classical statistical methodology is aimed at analyzing data from designed experiments and historically statistical analyses have been done by researchers who knew the design and origin story of the data set well. The origin stories of data sets have changed over time and today a lot of data is accumulated without specific purpose. This is due to vast amounts of data being registered online and to a trend towards more open source research. The latter phenomenon in particular poses new challenges wrt. data quality assessment. When data are collected and made public without a specific end-point in mind, how do we ensure that differences in, say, choice measurement instruments, mode of administration, or sampling frame do not cause the data to be effectively divided into subsets that are simply not comparable?

Surveys often use mixed modes of administration, e.g. mail and telephone, and while this can improve response rates, the mode of administration can affect results Brambilla and McKinlay (1987); McHorney et al. (1994) and differences in response behavior can lead to biased results. Powers, Mishra and Young (2005) report effects of mode of administration on changes in mental health scores that are of a magnitude that is considered to be clinically meaningful.

The rapid growth of web surveys, due to low cost, timeliness, and other factors, generate large data sources that lack a sampling frame of the general population. However, it can be problematic to combine online panels (pre-recruited profiled pools of respondents) with intercept samples (a pool of respondents obtained through banners, ads, or promotions) Liu (2016).

Sophisticated methods for addressing this question are available when we are willing to assume a statistical model, but when these models are taken away, a remarkable void of methods is left behind. What is needed is a procedure that compares differences in overall data structures in two (or more) subsets of a dataset without assuming neither directional nor hierarchical relationships between the variables. We propose a new method for this task, namely Principal Component Analysis-based Data Structure Comparison (PCADSC). This method employs the principal component decomposition of the data matrix performed on two subsets of a dataset in order to create intuitive visualizations of data structure differences. Mention R package.

This manuscript is structured as follows: First, in Section 2, we present the data structure comparison problem in more detail and discuss what statistical methods are already available for solving similar challenges. Next, in Section 3, we move on to a description of the PCADSC procedure, including a brief introduction to principal component analysis (PCA) in general. In Section 4, we present a worked data example using the open source, online available PISA data (ref), which is an example of a dataset where multiple data collection methods Eller mske lande? have been employed.

2. Something about state of the art

2.1. More detailed description of the type of problem we wish to address

- Two subsets of a dataset, i.e. to datasets with the same variables, but different observations
- Wish to compare structures without specifying a model or even any variables of interest

- The most central example is the question of whether the two subsets can readily be combined in a (unknown) data analysis, or if the subset-inducing variable actually implies heterogeneity across the subset division

  - Examples: Large scale open source datasets such as the PISA data and ESS (European Social Survey) data and ...(?). In these datasets, the data producers are very far away from the majority of the data analysts. Therefore, problem-specific recommendations about potential instrument-induced challenges in the datasets are not available for the data analysts. How can data producers ensure that this will not be an issue, at least not related to known data gathering differences?
  - Other examples?
  - Perhaps a description of what happens if we are to combine the two subsets of the datasets without taking a e.g. an instrument-effect into account. When will it cause problems (maybe: causal graph style?)?
  - Mention somewhere: We want a solution that is largely independent of the sizes of the two subsets of data. Thereby, a lot of methods that compare each subset to the full dataset in some sense are excluded.

2.2. Describe existing methods used to solve similar questions or parts of the question we are addressing

- The simplest case: variable-by-variable tests in distributional differences

  - Simple, but scales poorly
  - Only relates to marginal differences and not to the interplay between variables

- Karl's papers?
- Anne's papers: IRT-based methods for surveys

  - Moves beyond the marginal approach, but needs a model pre-specified
  - Thus, it is not 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. This is fundamentally a different (though related) question.

3. PCADSC - description of the method

Description from Anne's master's thesis. Rewrite.

As mentioned above, the purpose of PCADSC is comparing overall data structures in two or more subsets of a dataset. But before we can get further into describing this procedure, we must first define what exactly is meant by "overall structure". One such definition is the structure of the covariance matrix of the dataset. If we assume all variables in the dataset to be jointly normal with zero means, the covariance matrix is a sufficient statistic for describing the simultaneous distribution of all the variables. This gives it a very nice interpretation as a measure of the overall structure. If we do not accept the normality assumption, pairwise correlations and variable variances are still interesting quantities that say something about the interrelations between the variables, at least in terms of linear relationships. All in all, the empirical covariance matrix is a reasonable place to start looking for differences in "overall data structures".

Though the idea sounds appealing, it is quite difficult to assess similarity of matrices, and moreover, this becomes increasingly difficult for large numbers of variables and thus high dimensional covariance matrices. There is simply too much information to consider at once.

However, by clever use of linear algebra, we can construct a decomposition of the covariance matrix that makes it easier to gain an overview of the data. We propose a new method based on principal component analysis that seems to be able to identify differences in datasets based on intuitive, visual inspections. We refer to this method as principal component analysis-based data structure comparison (PCADSC) and we present the procedure below. But first, we give a minimal introduction to principal component analysis in general with reference to Koch (2014).

3.1. Principal component analysis

Consider n observations $x_1, \dots, x_n \in \mathbb{R}^d$ of d variables, let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_n$ denote their average and let $S = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top \in \mathbb{R}^{d\times d}$ denote the empirical covariance. Suppose that we want to describe the observations by q numbers instead of the original d numbers. The associated *rank-q-reconstruction error* is defined as the minimal squared error that is achievable by linear subspaces $K_q \subset \mathbb{R}^d$ of dimension $q < d$, that is

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

Principal component analysis (PCA) ensures the existence of a subspace $\hat{K}_q \subset \mathbb{R}^d$ that attains this minimum, and it provides an explicit description of $\hat{K}_q$ and the rank-q-reconstruction error. Thus, let $S = U\Lambda U^\top$ be the eigenvalue decomposition of S. Here $\Lambda \in \mathbb{R}^d$ is the diagonal matrix with the eigenvectors $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ in the diagonal, and $U \in \mathbb{R}^d$ is the orthogonal matrix with the associated eigenvectors $\eta_1, \dots, \eta_d \in \mathbb{R}^d$ in the columns. The eigenvalues are uniquely defined, and the eigenvectors are uniquely defined up to a change of sign whenever the eigenvalues are different. If some of the eigenvalues are identical, e.g. $\lambda_i = \lambda_{i+1} = \cdots = \lambda_j$, then the associated eigenvectors $\eta_i, \eta_{i+1}, \dots, \eta_j$ are uniquely defined up to a common rotation. In practice this only happens if $n < d$, in which case the last $d - n$ eigenvalues will be zero. It is a result from

linear algebra that the rank-q-reconstruction error for $q < d$ is achieved for

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

and equals $\sum_{j=q+1}^{d} \lambda_j$. The eigenvectors $\eta_j \in \mathbb{R}^d$ are called *loadings*, and the eigenvalues $\lambda_j \geq 0$ may be understood as *variation components*. The projections $\eta_j^\top (x_i - \bar{x})$ of the observations onto the loadings are called *scores*. The jth loading can also found as the unit vector $u \in \mathbb{R}^d$ orthogonal to $\hat{K}_{j-1}$, where the initial subspace is defined as $\hat{K}_0 = \{0\}$, that maximizes the variation of the associated scores

$$\eta_j = \text{argmax}_{u \in \mathbb{R}^d :\, u \perp \hat{K}_{j-1}} \sum_{i=1}^{n} \|u^\top (x_i - \bar{x})\|^2, \qquad \lambda_j = \frac{1}{n-1} \sum_{i=1}^{n} \|\eta_j^\top (x_i - \bar{x})\|^2.$$

It is worth emphasizing that the greedy approach of successively adding the next direction η_j explaining most of the remaining variation, also gives the sequence $\hat{K}_q = \hat{K}_{q-1} \oplus \text{span}\{\eta_q\}$ of subspaces minimizing the rank-q-reconstruction error. This strong interpretation of PCA, which is often overlooked in the literature, means that the sequence of loadings η_j and their associated variation components λ_j yield a simultaneous description of the structure of the data set for all approximating dimensions q. This implies that the loadings and variation components can be used to investigate the structure of the data set without the need to decide on an approximating dimension q.

3.2. PCA-based data structure comparison

Above, we promised a method for intuitive, visual inspection of data structure similarities, but as of now, all intuition might have been lost in technicalities. The main point we want to emphasize from PCA is that whereas the scores describe the observations, the variation components and the accompanying loadings describe the usage of the variables. If two different datasets with the same variables, but different samples of observations, have similar loading patterns, then the variables appear to be measuring the same underlying quantities in both data situations. This can be the case while the two sets of scores could be arbitrarily different, which e.g. could happen if the two datasets were taken from two different populations of subjects. On the other hand, if the loading patterns are different in the two datasets, then this indicates that the variables are used differently in the two data situations, and hence it would be criticizable to use these variables for comparisons across the two datasets.

In this paper we propose three diagnostic plots, referred to as the *CumEigenPlot*, *HairPlot* and *PancakePlot*, for comparing the loading patterns in two datasets. In order to describe these plots we consider two different datasets in the same d variables and with n_1 and n_2 observations, respectively. For dataset $i = 1, 2$ let $S_i \in \mathbb{R}^{d \times d}$ be the empirical correlation matrix (find a way to introduce the empirial correlation matrix instead), and let $\lambda_{i1} \geq \cdots \geq \lambda_{id} \geq 0$ and $\eta_{i1}, \ldots, \eta_{id} \in \mathbb{R}^d$ be the corresponding variation components and loadings. The correlation matrices correspond to the covariance matrices for the variables after these have been standardized to unit standard deviation separately within the two datasets. Similarly, let $S \in \mathbb{R}^{d \times d}$ be the empirical correlation matrix for the combined dataset with $n_1 + n_2$ observations,

and let $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ and $\eta_1, \ldots, \eta_d \in \mathbb{R}^d$ be the corresponding variation components and loadings.

To do: justification for using correlation matrices should be further elaborated. And also: Something about what to do if not all variables are numerical. Using correlation matrices is not the same as assuming standadized variables; there is a difference in the permutations.

Description of the *CumEigenPlot*: This plot compares the variation components. If these are the same in the two sample populations, then the best estimate for the variation components are $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ found in the combined dataset. And we would expect $\lambda_{i1} \geq \cdots \geq \lambda_{id} \geq 0$ for $i = 1, 2$ to be alike excepts sample variation. 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

$$(0,0), \quad (\lambda_1, \lambda_{11} - \lambda_{12}), \quad (\lambda_1 + \lambda_2, \lambda_{11} + \lambda_{12} - \lambda_{12} - \lambda_{22}), \quad \ldots, \quad \Big(\sum_{j=1}^{d} \lambda_j, \sum_{j=1}^{d} \lambda_{1j} - \sum_{j=1}^{d} \lambda_{2j} \Big).$$

This may be seen as a cumulated Bland-Altman plot for the variation components (reference to cumulated residuals and to Bland-Altman). Note that since we are using correlation matrices the last point will always be $(d, 0)$. Thus, this curve will begin and end at the x-axis. And the larger excursions it makes away from the x-axis the less alike the cumulated variation components for the two datasets are.

Whether the excursions in the observed curve are large or within the range of sample variation can be quantified by a permutation test. The idea is that we randomly reallocate the $n_1 + n_2$ observations in the combined dataset to two datasets with n_1 and n_2 observations, respectively, and then repeat the procedure described above. The random reallocation by construction ensures that the two resampled datasets are alike except their sample size and sampling variation. In the *CumEigenPlot* we make 1000 independent random reallocations, and plot the observed curve together with 20 of the resampled curves as well as a shaded region visualizing pointwise 95pct coverage intervals. This provides a graphical validation tool. P-values for the null hypothesis that the variation components are the same are easily provided for any appropriate test statistic by the resampling procedure as well. We have implemented the *Kolmogorov-Smirnov* and the *Cramér-von Mises* test statistics, which are given by

$$\mathrm{KS} = \max_{k=1,\ldots,d} \Big| \sum_{j=1}^{k} \lambda_{1j} - \sum_{j=1}^{k} \lambda_{2j} \Big|, \qquad \mathrm{CvM} = \sum_{k=1}^{d-1} \frac{\lambda_k + \lambda_{k+1}}{2} \Big(\sum_{j=1}^{k} \lambda_{1j} - \sum_{j=1}^{k} \lambda_{2j} \Big)^2.$$

Description of the *HairPlot:* Let $\lambda_{\max} = \max\{\lambda_{11}, \lambda_{21}\}$ be the largest variation component for the two datasets. The empirical correlation matrix S_1 for the first dataset has the following orthogonal decomposition in the coordinate system of the second dataset

$$S_1 = \sum_{k=1}^{d} \lambda_{1j} \eta_{1k} \eta_{1k}^\top = \lambda_{\max} \sum_{k=1}^{d} \Big(\sum_{j=1}^{d} \sqrt{\frac{\lambda_{1k}}{\lambda_{\max}}} (\eta_{1k} \eta_{2j}^\top) \eta_{2j} \Big) \Big(\sum_{j=1}^{d} \sqrt{\frac{\lambda_{1k}}{\lambda_{\max}}} (\eta_{1k} \eta_{2j}^\top) \eta_{2j} \Big)^\top,$$

and we have a similar decomposition of S_2 in the coordinate system of the first dataset. We propose to visualize these two decompositions in a $d \times d$ grid display. In the jth row and kth column of this display we plot two arrows based at the lower left corner of the grid cell. The first arrow has length μ_{jk} and angle $\theta_{jk}/2$ anticlockwise from the diagonal, and the second arrow has length ν_{jk} and angle $\theta_{jk}/2$ clockwise from the diagonal. To facilitate the following description we will refer to the arrows drawn anticlockwise as the blue arrows, and the arrows drawn clockwise as the red arrows. The lengths μ_{jk} and ν_{jk} and the angle θ_{kj} are given by

$$\mu_{jk} = \sqrt{\frac{\lambda_{1k}}{\lambda_{\max}}}|\eta_{1k}\eta_{2j}^\top|, \qquad \nu_{jk} = \sqrt{\frac{\lambda_{2j}}{\lambda_{\max}}}|\eta_{2j}\eta_{1k}^\top|, \qquad \theta_{jk} = \arccos(|\eta_{1k}\eta_{2j}^\top|).$$

The absolute value of the projection $\eta_{1k}\eta_{2j}^\top$ is inserted due to the indeterminacy of the direction of loading vectors. This indeterminacy implies that the angle between loadings from the two datasets always can be chosen to be in the interval $[0, \pi/2]$, and hence the decomposition of S_1 and S_2 can be visualized in a joint plot by dividing the angles by two and using anticlockwise and clockwise shifts from the diagonal. Furthermore, the scaling of the lengths by $\lambda_{\max}$ is made so that the longest arrow has at most unit length.

In the *HairPlot* the blue arrows in the kth column of the grid display visualize the decomposition of the kth principal components for the first dataset in the coordinate system of the second dataset. Similarly, the red arrows in the jth row visualize the decomposition of the jth principal components for the second dataset in the coordinate system of the first dataset. If the structures of the two datasets are identical, then we will have coinciding blue and red arrows along the diagonal in the grid display, and nothing else as arrows in the off-diagonal cells would have zero length. Differences in the variation components 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.

Remarks: Presently permutation tests based on the Kolmogorov-Smirnov and the Cramér-von Mises test statistics has been implemented for CumVarAgreement. This should also be possible for the Hairplot based on the off-diagonal components. For the Pancakeplot a Wallyplot has been implemented (however, in Claus' Wally plot the position of the "observed dataset" is also random).

Description of the *PancakePlot*: Our proposal of a PCADSC method consist of three steps. These steps should be performed separately for each of the two (or more) datasets that we wish to compare. Note that the datasets must have the same variables, but different sample sizes are allowed. The three steps are:

1. Standardize. Full data or subsets? And also: Something about what to do if not all variables are numerical.
2. Compute the PCA loadings and the variance contributions of each principal component.

3. For each principal component, standardize the loadings, i.e. scale them such that they sum to one.
4. Produce a plot consisting of a bar for each principal component, decorated with the cumulative variance contribution corresponding to this component. The bar should be of length one and colored according to the variables loading the component.

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. Refer to example/show plot.

4. Data example stuff

We will now turn to a concrete data example in order to illustrate the possibilities of each of the methods presented above more than just PCADSC?. We use data from the 2012 version of the European Social Survey (ESS) project, a very large dataset that is freely available online at `www.europeansocialsurvey.org`. As the name suggests, the data comes from a survey that was conducted primarily in Europe aiming to collect information about the social conditions of the citizens reference?. As with all international (or, simply, multi-center) studies, one might be concerned about whether or not the data from different countries can readily be combined. This is the question we will address using PCADSC in the current section.

All computations and figures presented in this section where created using our `R` package `PCADSC`, which is available online at `www.github.com/AnnePetersen1/PCADSC` maybe do a CRAN submission instead?.

4.1. Data

The ESS 2012 data contains a total of 626 variables collected from 54673 citizens in 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 (REF: ESS6 Topline Results Series 5), as illustrated in Figure 1. We represent each of these scales by a single variable, which is calculated as the average score within the items related to that variable (ref to someone that says that is sensible?). 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.

In the following, we only compare two countries, namely Denmark and Bulgaria, which have $n_{BG} = 1798$ and $n_{DK} = 1498$ complete cases in the variables of interest, respectively. For these two countries, the ESS authors different term? particularly highlight differences in the relationship between the psychological well-being scales, at least on an nation-aggregated level (REF: ESS6 Topline Results Series 5). This might be the result of large, cultural and socio-economic differences between the two countries. Simply put, we do not expect happiness and psychological well-being to be the same phenomena in Bulgaria and Denmark more here about why, about the cultures?. Therefore, a successful method for data comparisons should be able to detect these differences by looking at the differences in the interplay between the 6 scales of psychological well-being.

Table 1: Items from the wellbeing module grouped by the dimension of wellbeing they relate to

WELLBEING DIMENSION	ESS SURVEY ITEM
Evaluative wellbeing	How satisfied with life as a whole
	How happy are you
Emotional wellbeing	Felt sad, how often past week
	Felt depressed, how often past week
	Enjoyed life, how often past week
	Were happy, how often past week
	You felt anxious, how often past week
	You felt calm and peaceful, how often 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
Functioning	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 an effort, how often past week
	Sleep was restless, how often past week
	Could not get going, how often past week
	Had lot of energy, how often past week
Community wellbeing	Most people can be trusted / can't be too careful
	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 past week

Source: European Social Survey Round 6, 2012

FIGURE 1.
Make a new table here

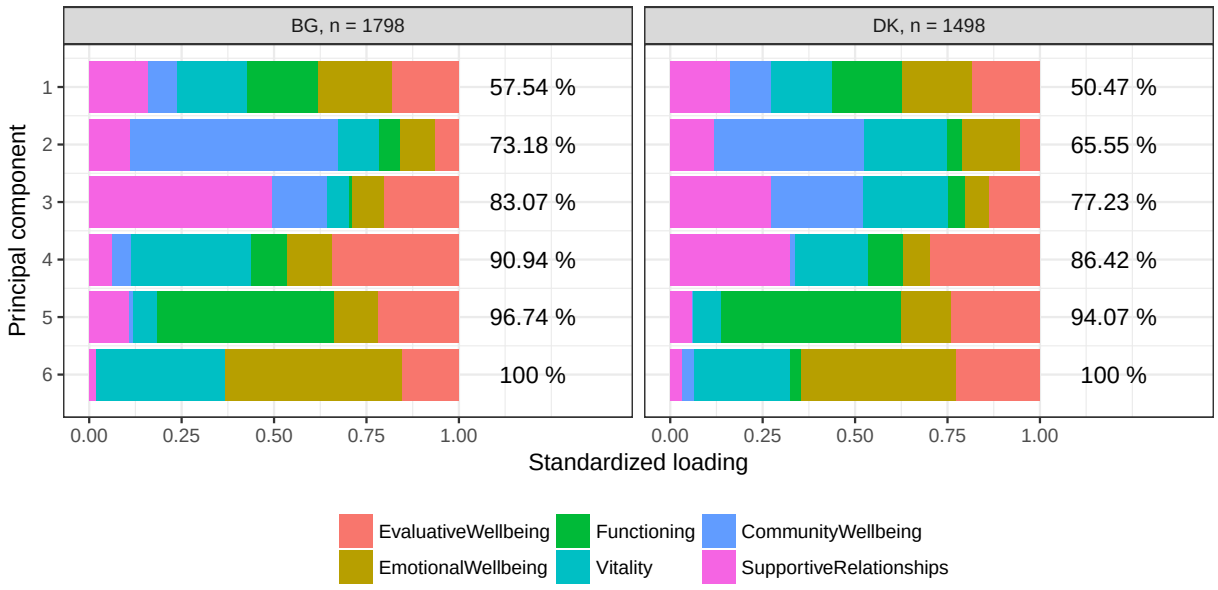

FIGURE 2.
something

4.2. PCADSC

Figure 2 illustrates the results of conducting PCADSC on the psychological well-being data from the ESS. While the first principal component, which is responsible for explaining 50-60 % of the variance in the data, is very similar for the two countries, we see quite large differences in the remaining components. In the second component, we see that the two countries disagree in the relative importance of the scales *Community wellbeing* and *Vitality*. In the third and fourth components, general disagreement is found. All in all, components 2-4, representing almost half of the variability in the data, are not very similar across the two countries. Moreover, the two subsets of the data also differ with respect to how much variance is explained by each component, and the difference is particularly big for the first component. This component has approximately 15 % more explanatory power in the Bulgarian subsample than it does in the Danish.

Figure 4 shows the hairplot. The blue arrows visualize the decomposition of the principal components for the first dataset in the coordinate system of the second 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 second dataset in the coordinate system of the first dataset. We see that PC2 also loads on PC4, that PC3 also loads on PC2 and PC4, and that PC4 also loads on PC3. For PC1, PC5 and PC6 the main difference is in the size of the variation component.

In summary, we find that the datasets are fundamentally different and that 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.

FIGURE 3.
something.

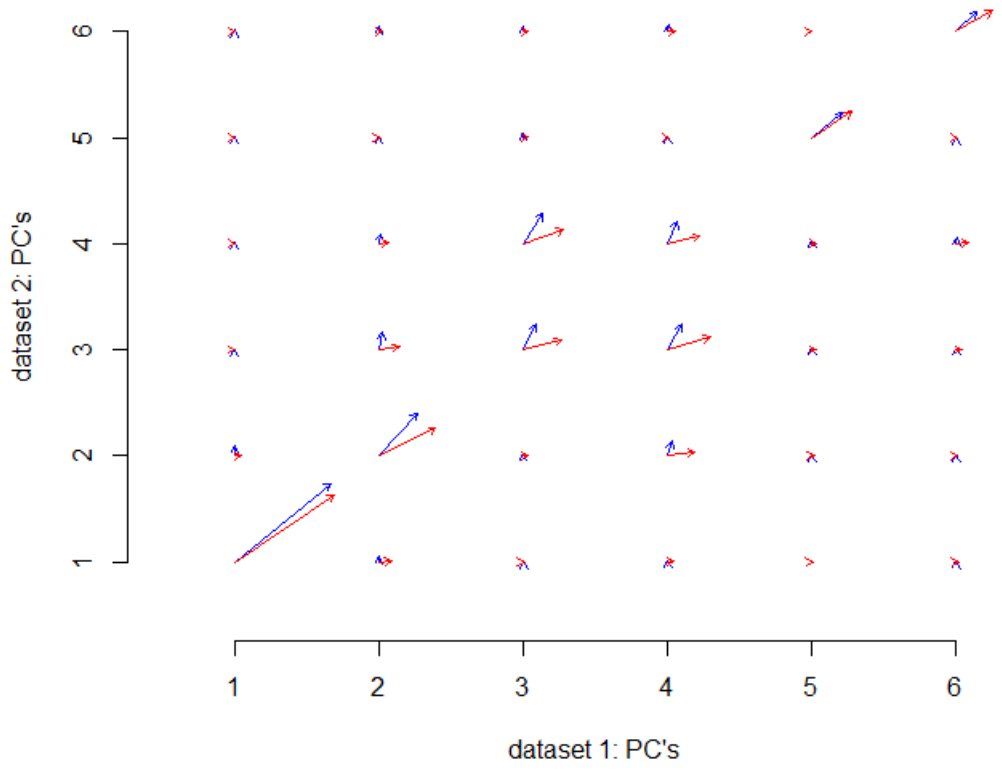

FIGURE 4.
The blue arrows show the principal components of the first dataset decomposed in the coordinate system of the principal components of the second dataset. And the red arrows the reverse.

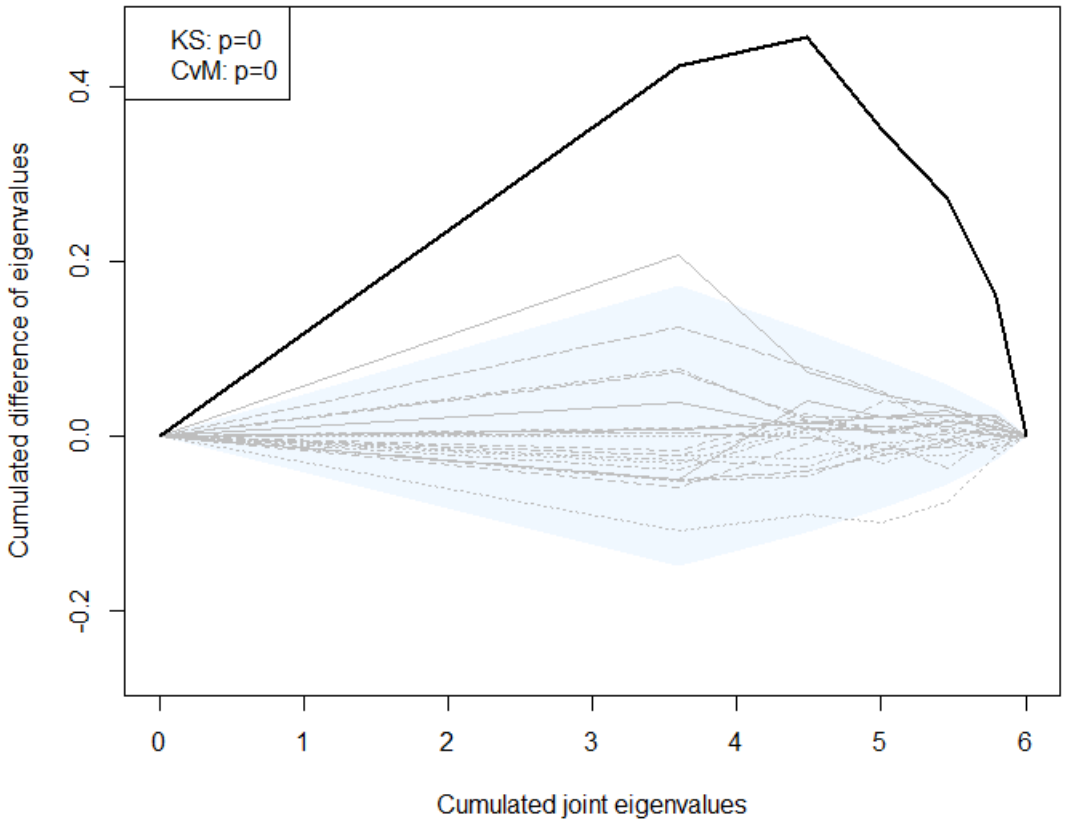

FIGURE 5.
something.

But did we really illustrate a data structure difference due to country differences or did we just illustrate the variability of the results of the PCADSC method? In order to investigate this further, we look at a so-called *Wally plot* (ref: Claus Ekstrm). In this plot, we compare the results of PCADSC conducted with grouping by country with several random, but similar grouping variables. Specifically, we produce 7 PCADSC plots where the country variable was replaced by a randomly generated variable that divides the observations into two groups of the same sizes as the country samples. The results are illustrated in Figure 3. Here, we see that the differences in the second component from the original PCADSC results are not matched in any of the randomly grouped PCADSC runs. In fact, the 7 runs are remarkably similar, thereby illustrating that PCADSC seems to be very robust with respect to random groupings: The signal in the data is not blurred by the random subdivisions. When it comes to the differences in the third component for the two groups, we find much larger variability in the 7 random runs. more comments here... Wait until we are sure exactly what we think about the results and what other PCA-based methods, we will do before/after. Particularly, how do we deal with eigen value differences?

5. Discussion

- Generalizing the results to non-numeric variables?
- Generalizing the results to covariance matrices that are not of full rank?
- ?

6. Concluding Remarks

References

Brambilla, D. J. and McKinlay, S. M. (1987). A comparison of responses to mailed questionnaires and telephone interviews in a mixed mode health survey. *American journal of epidemiology*, 126(5):962–71.

Liu, M. (2016). Comparing data quality between online panel and intercept samples. *Methodological Innovations*, 9:2059799116672877.

McHorney, C. A., Kosinski, M., and Ware, J. E. (1994). 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(6):551–67.

Powers, J. R., Mishra, G., and Young, A. F. (2005). 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(2):149–54.