EXPLORATORY DATA STRUCTURE COMPARISONS: THREE NEW VISUAL TOOLS BASED ON PRINCIPAL COMPONENT ANALYSIS

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

Abstract

abstract...

Key words: multi-center studies, data heterogeneity, data structure comparisons, principal component analysis (PCA), goodness of fit, European Social Survey.

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. However, the origin stories of data sets have changed over time and today a lot of data is accumulated without a specific purpose in mind. 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? Three examples of this are

- 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 behaviour 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), see Liu (2016).
- Large scale open source data sets such as the PISA data and ESS (European Social Survey) data. In these data sets, data analysts are far away from the data producers and problem-specific recommendations about potential instrument-induced challenges in the data sets are not available for the data analysts.

Maybe broaden the raison d'etre of PCADSC a bit: It is also a useful tool for doing initial, exploratory data analysis and it helps us to identify potential problems or challenges early in the study design development, before concrete model choices have been made, thereby aiding a more economic use of resources and a less ad-hoc approach to the types of problems that PCADSC can find.

Assume that we have two data sets with the same variables, but different observations represented as a data set with a subset-inducing variable. Assume next that we wish to compare structures without specifying a model, or even a variable of interest. The central question is whether the two data sets can readily be combined for the purpose of later data analysis, or if the subset-inducing variable implies heterogeneity that must be taken into account.

Sophisticated methods for addressing this question are available when we are willing to assume a statistical model Skal der referencer her? Nogen forslag?, 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 data set without

assuming neither directional nor hierarchical relationships between the variables. The use of parametric models does not constitute 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. Simple methods like variable-by-variable tests in distributional differences suffer from the drawback that they only address marginal differences and not to the interplay between variables. Entry-by-entry comparison of the two empirical correlation matrices quickly becomes unmanageable as the number of variables increase. Parametric models using, e.g., latent variable models moves beyond the marginal approach, but need a pre-specified model.

We propose a suite of three new tools for this task, which we will refer to collectively as Principal Component Analysis-based Data Structure Comparisons (PCADSC). These methods 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. This yields a solution that is largely independent of the sizes of the two subsets of data. The proposed tools are implemented in the statistical software `R` (R Core Team) statistical software in our package, `PCADSC`, which is available at ? How to refer to a package on Github?.

This manuscript describes the procedures, including a brief introduction to principal component analysis (PCA) in general and presents a worked data example using open source, online available data on psychological well-being in three European countries. More specifically, we compare data from Denmark with data from Bulgaria and Sweden, respectively, to investigate whether or not data on psychological well-being can be combined across countries.

2. PCA-based tools for data structure comparisons

As mentioned above, the purpose of PCADSC is to compare overall data structures in two subsets of a data set. But before we can get further into describing the PCADSC tools, we must first define the exact meaning of *overall structures* in this context. 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 marginal variable variances are still interesting quantities that say something about the interrelations between the variables in the data. All in all, the empirical covariance matrix is a reasonable place to start looking for differences in *overall data structures*.

A naive approach for data structure comparisons might therefore be to compute the empirical covariance matrices on each of the two data subsets and simply compare these matrices. Though the idea perhaps sounds appealing at first, it is quite difficult to assess similarity of matrices, and moreover, it 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 use of linear algebra, we can decompose and recompose the covariance matrix such that the distinct dimensions of information withheld in it are clearly separated and ordered according to their relative importance. In particular, we find a new representation of the covariance matrix that makes it possible to gain an overview of the most interesting aspects of the data. On such decomposition strategy is using *principal component analysis* (PCA), which uses

eigenvalue decomposition in order to obtain a new representation of the covariance matrix.

2.1. A brief introduction to Principal component analysis

Consider n observations $x_1, \dots, x_n \in \mathbb{R}^d$ of d variables, let $\bar{x} = (\bar{x}_1, ..., \bar{x}_d)^\top = \frac{1}{n}\sum_{i=1}^n x_n$ denote their averages, 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 matrix of the full data matrix, X. We assume in the following that all variables of X have a numerical interpretation, e.g. by being continuous or ordinal and categorical. 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 information is lost. More specifically, we wish to minimize the loss when looking at a projection of X onto a q-dimensional space, rather than the orignial 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

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

and 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 (Hastie et al., 2009).

More specifically, the rank-q-reconstruction error is attained when we choose

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

where $\eta_1, ..., \eta_q$ are the q first eigenvectors of the empirical covariance matrix, S, as ordered by the size of their associated eigenvalues, $\lambda_1, ..., \lambda_d$. 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,

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

where $X_j \in \mathbb{R}^n$ denotes the jth variable of 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 different, the representation hereby obtained is a valid object of inference. If $n < d$ or if the covariance matrix does not have full rank, it is possible to obtain non-unique eigenvalues, e.g. $\lambda_i = \lambda_{i+1} = \dots = \lambda_0$, but even in this case, the associated eigenvectors $\eta_i, \eta_{i+1}, \dots, \eta_j$ are uniquely defined up to a common rotation.

The jth 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:

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

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 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 the variation components can be used to investigate the structure of the data set without the need to decide on an approximating dimension, q, a priori.

2.2. Using PCA for data structure comparisons

All in all, PCA qualifies as an appealing first step in structural comparisons of two datasets containing the same variables, and especially the loadings and variance components are meaningful quantities to compare across such different datasets. Usually, when performing PCA with other purposes in mind, the main interest lies in the *scores*, i.e. the projections $\eta_j^\top(x_i - \bar{x})$ of the observations onto the loadings. But where 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 datasets. 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.

The tools presented in this paper are all based on comparing the PCA results across two different datasets that contain the same variables. We denote these two datasets by X_1 and X_2, respectively, with X_1 consisting of n_1 observations of d variables $x_{11}, \ldots, x_{1n_1} \in \mathbb{R}^d$, and similarly, X_2 consisting of n_2 observations of the d variables, $x_{21}, \ldots, x_{2n_2} \in \mathbb{R}^d$. For each of these two datasets, we complete the following steps:

1. Standardized each of the variables to have mean zero and unit standard deviation. Let $\tilde{x}_{ij} \in \mathbb{R}^d$ for $i = 1, 2$ and $j = 1, \ldots, n_i$ be the standardized datasets.
2. Form the principal component analysis

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \tilde{x}_{ij} \tilde{x}_{ij}^\top = \sum_{k=1}^{d} \lambda_{ik} \eta_{ik} \eta_{ik}^\top \quad \text{for } i = 1, 2.$$

   thereby obtaining loadings η_{ik} and variance components λ_{ik} for $i \in \{1, 2\}$ and $k \in 1, ..., d$.

The hereby obtains PCA decompositions of the covariance matrices can then be compared. We present three diagnostics plots that are designed to shine a light on different types and nuances of data structure differences. These three plots are:

The CE plot: The CE (cumulative eigenvalue) plot can be used to illustrate differences in variance components, that is, in the relative importance of the directions identified by the

PCA. The CE plot is accompanied by two permutation-tests, testing the hypothesis of no difference in variance components.

The angle plot: The angle plot compares both loadings and variance components 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 variance component pairs) that are most similar and most different across the two datasets

The chroma plot: The chroma plot is primarily an illustration of the 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.

For the deepest understanding of the data structure differences in two datasets, we suggest using all three steps in the above order.

But before diving deeper into the details of the *CE plot*, the *angle plot* and the *chroma plot*, a general remark about standardization in PCA is in place. We would like to emphasize that PCA is sensitive to scaling, as the procedure 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 be always be deemed highly influential, no matter the structure of the data. Therefore, the variables should always be scaled prior to performing PCA. Note that the covariance matrix for the standardized variables is the same as the correlation matrix for the original variables, so this simply corresponds to performing data structural comparisons of the correlation matrices rather than the covariance matrices. The standardization makes the variables comparable on the same scale, i.e. units of standard deviation, and it implies that the diagonal elements of S, S_1, and S_2 all equals 1, and thus also that

$$\sum_{k=1}^{d} \lambda_k = \sum_{k=1}^{d} \lambda_{1k} = \sum_{k=1}^{d} \lambda_{2k} = d$$

which will simplify some expressions below.

2.2.1. The cumulative eigenvalue plot

The cumulative eigenvalue (CE) plot compares the variation components, i.e. the eigenvalues of the covariance matrix. 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

$$(0,0), \quad (\lambda_1, \lambda_{11} - \lambda_{12}), \quad (\lambda_1 + \lambda_2, \lambda_{11} + \lambda_{12} - \lambda_{21} - \lambda_{22}), \quad \ldots, \quad \left(\sum_{j=1}^{d} \lambda_j, \sum_{j=1}^{d} \lambda_{1j} - \sum_{j=1}^{d} \lambda_{2j} \right)$$

This may be seen as a cumulative Bland-Altman plot for the variation components (reference to cumulative residuals and to Bland-Altman). Note that 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 x-axis, the less alike the cumulative variation components for the two datasets are. Moreover, a positive cumulative differences implies that dataset 1 holds more information in the first components than dataset 2 does.

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

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

We conduct the tests as *permutation tests*, that is, by randomly reallocating the combined and standardized datasets into two new datasets of n_1 and n_2 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 even larger test statistics than 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.

2.2.2. The angle plot

This plot simultaneously compares the variation components and the loadings. 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

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

and we have a similar decomposition of S_2 in the coordinate system of the first dataset Anne: it was not completely clear to me what this meant the first xx times I read it - expand?. 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 $\theta_{jk}/2$ counterclockwise 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 counterclockwise 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

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

Note that for two d-dimensional, unit length vectors a and b, $a^\top b = \langle a, b\rangle = \tilde{c}(a, b)$, where $\tilde{c}$ denotes the sample correlation. 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 component contributions. 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 counterclockwise 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 *angle plot*, the blue arrows in the kth column of the grid display visualize the decomposition of the kth principal component 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. Can we expand on this in a non-technical way? 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.

2.2.3. The chroma plot

This plot compares the loadings of the two datasets. 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 ith principal component, we plot a vertical bar of length one that has been divided into d segments of different colors and the wideness of the jth such segment is given by

$$\omega_{ij} = \frac{|(\eta_i)_j|}{\sum_{k=1}^{d} |(\eta_i)_k|}$$

where $(\eta_i)_j$ denotes the jth entry of η_i. Due to the indeterminacy of the sign, all the signs are removed from the coefficients in the loadings. The bars are ordered according to the variation components and they are annotated with the cumulative percentage explained variance of that component, that is, the scaled and summed variance component contributions,

$$\tilde{\sigma}_i^c = \frac{\sum_{j=1}^{i} \lambda_i}{\sum_{k=1}^{d} \lambda_k} = \frac{\sum_{j=1}^{i} \lambda_i}{d}$$

Especially when d is large, we recommend plotting only a select set of interesting principal components, e.g. as identified by use of the angle plot. In this scenario, the annotations should rather be the non-cumulative variance contributions, $\tilde{\sigma}_i = \frac{\lambda_i}{d}$.

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 that *"variable x is generally more influential in subset 1 than it is in subset 2"*, thereby helping us to understand where and why the data structure differences are found.

3. European differences in psychological well-being: A data example

We will now turn to a concrete data example in order to illustrate the capabilities of the methods presented above. 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. This investigation is motivated by an increasingly popular new tendency to publish miscellaneous rankings of countries in fields as different as educational quality (e.g. the PISA tests) and citizen happiness (e.g. the UN *World Happiness Report* project). From a methodological point of view, such international rankings are very concerning, as they rely on the fundamental assumption that the measured concepts are inherently the same across countries. The PCADSC tools qualify as a suite of methods for exploring the validity of this assumption empirically.

For international comparisons of educational systems, the PISA tests have repeatedly been criticized for not being meaningful objects for international comparisons, especially due to problems with differential item functions (Kankara and Moors (2014); Kreiner and Christensen (2014)) and translation problems (Asil and Brown (2016)). Til Karl: Mske kender du nogle flere studier her, vi br nvne?

In the rankings of happiness, not much work has yet been devoted to evaluating the assumption of international comparabiltity, though Veenhoven (2012) presents a theoretically thorough, but empirically simplistic, summary of possible reasons for lack of comparability and Lolle and Andersen (2016) shows highly potent translation issues for the term *happiness*. The main question is whether or not there exist such a thing as a universal, internationally valid concept of happiness. Or do different aspects of psychological well-being or happiness simply not have the same relative meaning in different cultural and socioeconomic settings? This is in fact a question concerning comparability of data structures. If two countries differ e.g. in terms of how social networks are typically build and structured, with one emphasizing family relations and the other mostly focusing on other social relations, having a weak family connection does not have the same implications in the first country as it does in the second one. More specifically, 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 endeavor. In this paper, we will focus on a single aspect of overall happiness, namely psychological well-being,

In this section, we use the PCADSC tools to unveil international differences in one aspect of

	Denmark			Bulgaria			Sweden		
	Q_1	M	Q_3	Q_1	M	Q_3	Q_1	M	Q_3
Evaluative wellbeing	8.00	8.75	9.50	3.50	5.00	7.00	7.00	8.00	9.00
Emotional wellbeing	7.22	8.33	8.89	5.00	6.67	7.78	6.67	7.78	8.89
Functioning	6.93	7.57	8.21	5.50	6.68	7.68	6.39	7.04	7.68
Vitality	6.67	7.50	8.33	5.83	7.50	8.33	6.67	7.50	9.17
Community wellbeing	5.83	6.77	7.57	3.70	4.67	5.70	5.66	6.57	7.37
Supportive relationships	7.42	8.25	8.92	6.17	7.25	8.08	7.42	8.25	8.75

TABLE 1.
The 1st quartile, the median and the third quartile of the distrubutions of each of the six dimensions of psychological well-being, stratified by country. Note that the scales are constructed such that they all run from 0-10.

happiness, namely psychological well-being. 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 (Helliwell et al. (2016)), 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 not geographically nor culturally as far apart as some other countries might be, these two countries have previously been highlighted to be very different in terms of what defines happiness (Jeffrey et al. (2015) Is this the correct way to refer to a technical report?). Moreover,intra-European, regional differences in the relationship between social capital and happiness have also been demonstrated (Rodríguez-Pose and von Berlepsch (2014)), with a much less strong relationship between the two in Northern- compared to other European countries. In particular, interpersonal relations should 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 illustrate that the PCADSC tools actually do have some discriminatory power. Denmark and Sweden are both Scandinavian countries and are often deemed very similar in terms of culture and history. Therefore, we expect fundamental concepts such as psychological well-being to similar across these two countries.

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`.

3.1. Data

The ESS 2012 data contains 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 wellbeing*, *Emotional wellbeing*, *Functioning*, *Vitality*, *Community wellbeing* and *Supportive relationships*. More details on these scales can be found in (Jeffrey et al. (2015)) and the relationship between questionnaire items and scales is summarized in Table 2 in the appendices. 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 gives us $n_{DK} = 1498$ observations in the Danish sample, $n_{BG} = 1798$ observations in the Bulgarian sample and $n_{SE} = 1736$ Swedish observations. Table 1 summarizes the marginal distributions of the six dimensions of psychological well-being, stratified by country.

3.2. Comparing Denmark and Bulgaria

Figure 1 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.

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 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 blot in Figure 2 allows us to look closer into these components. Here, we find that the relative importance of the *Community wellbeing* 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 characteristic play a relatively larger role, which corresponds with previous findings. Thus, 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 (Jeffrey et al. (2015)). Moreover, the 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.

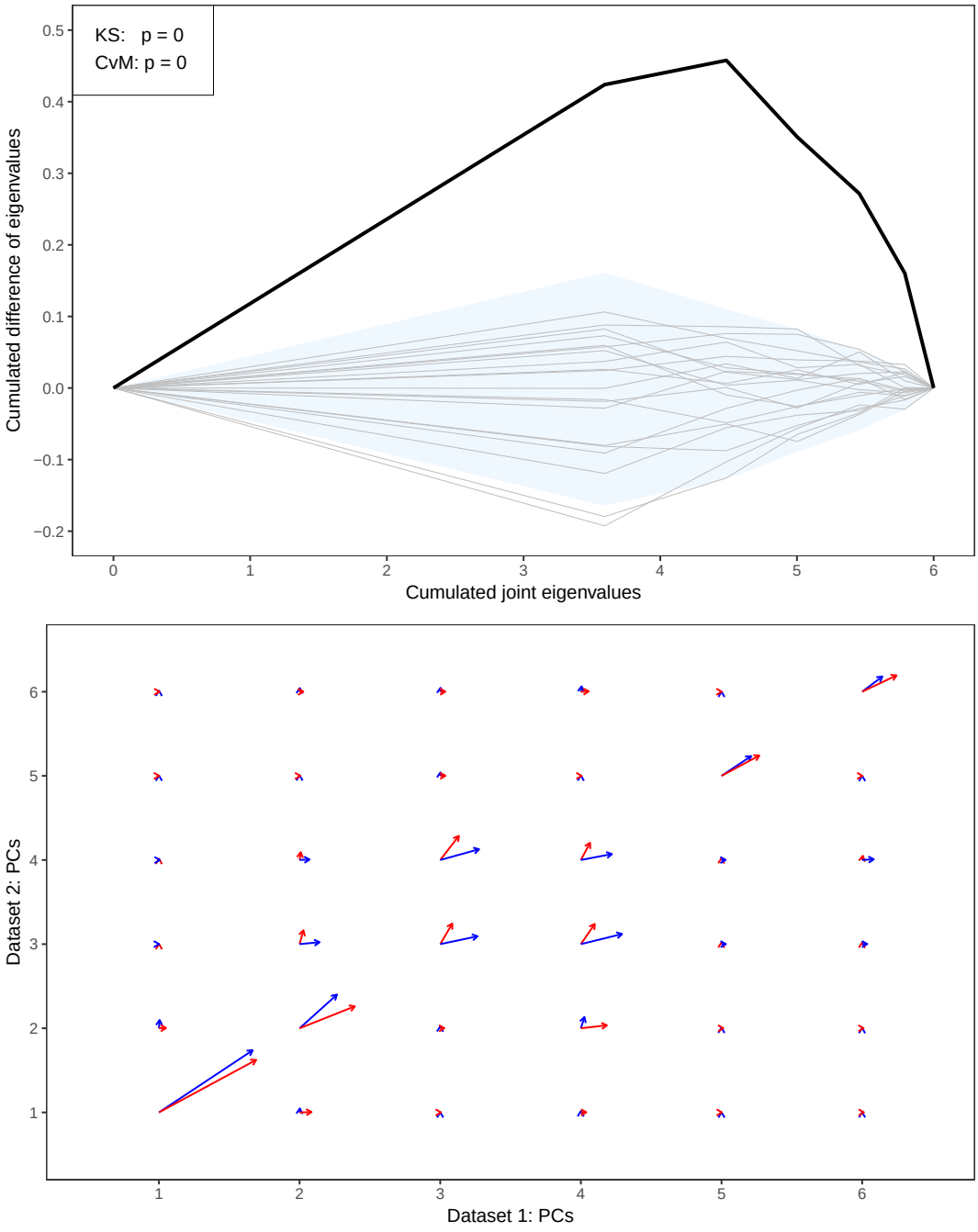

FIGURE 1.
The CE plot (top) and the angle plot (bottom) resulting from comparing Bulgarian and Danish data on psychosocial well-being. Dataset 1 refers to the Bulgarian subsample, while Dataset 2 is the Danish data. The CE plot is annotated with the p-values of the Kolmogorov-Smirnov and the Cramér-von Mises tests of the assumption 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.

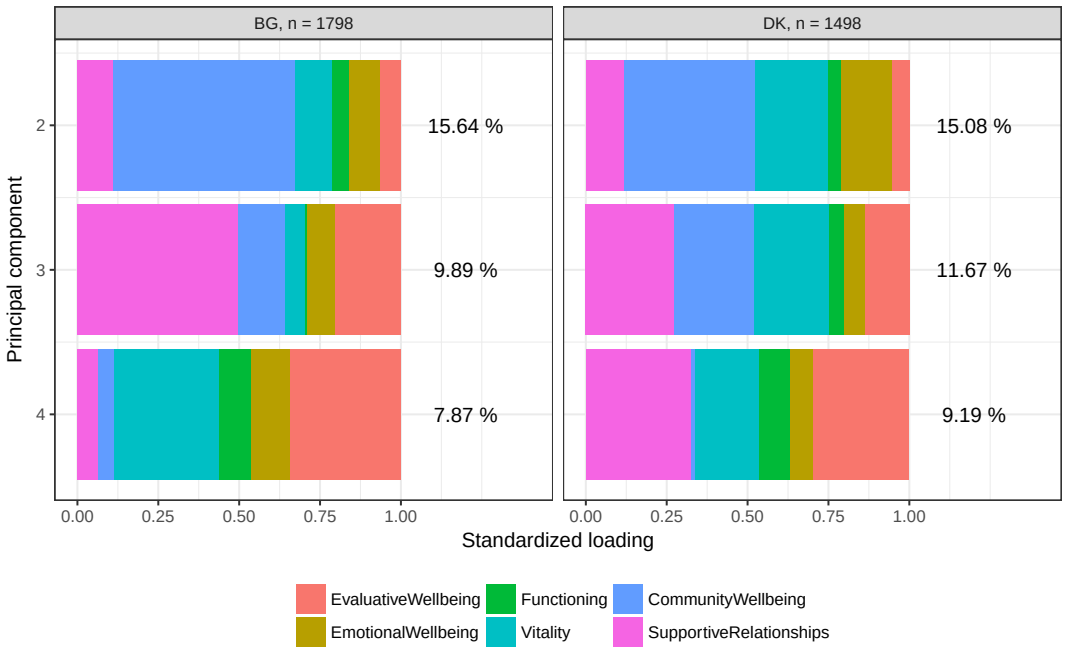

FIGURE 2.
A chroma plot comparing the 2nd, 3rd and 4th principal components of the Bulgiarian- and Danish psychological well-being data. The component-bars are annotated with their relative variance contributions (denoted $\tilde{\sigma}_i$ in the above).

3.3. Comparing Denmark and Sweden

We now turn to the comparison of Denmark and Sweden in terms of psychological well-being. Figure 3 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 $p_{\mathrm{KS}} = 0.14$ and $p_{\mathrm{CvM}} = 0.09$, respectively, thus accepting the null-hypothesis at the typical 5% level, but not with overwhelming evidence.

The angle plot from Figure 3 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 e.g. 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 4, 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 e.g. conclude using Table 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 a lot 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. Discussion

When deciding whether to combine two data sets for analysis, the issue of heterogeneity across data sets must be addressed. Simple methods suffer drawbacks and will likely scale poorly. Parametric models using, e.g., latent variable models moves beyond the marginal approach, but need a pre-specified model.

New tools, referred to collectively as Principal Component Analysis-based Data Structure Comparisons (PCADSC), for the task of deciding if the two two data sets can be combine for analysis were proposed and discussed in the paper. They 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 yielding a solution that is largely independent of the sizes the two data sets.

The methodology is quite general and in principle any plot that can be made can be included ('di-scree plot') ... [rephrase!!]

Further topics need to be addressed. These include generalizing the methods to be able to address (i) binary, ordinal or even nominal categorical variables, (ii) covariance matrices that are not of full rank.

More evaluation of the performance is also needed. Investigating sensitivity towards the sample sizes n_1 and n_2

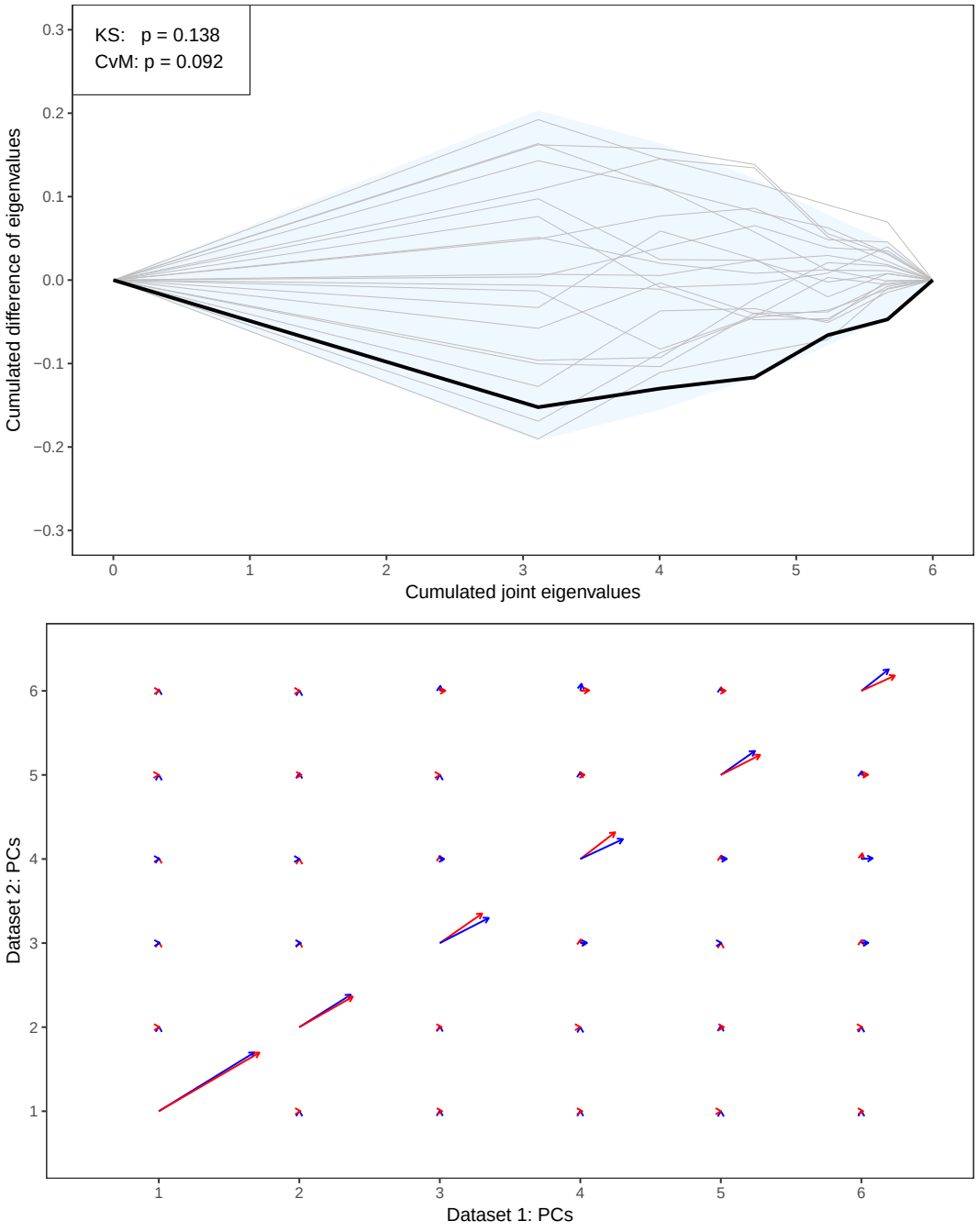

FIGURE 3.
A CE (top) and a hair (bottom) plot for comparing the Danish (Dataset 1) and the Swedish (Dataset 2) psychological well-being data. The blue arrows show the principal components of the Danish dataset decomposed in the coordinate system of the principal components of the Swedish dataset, and the red arrows illustrate the reverse.

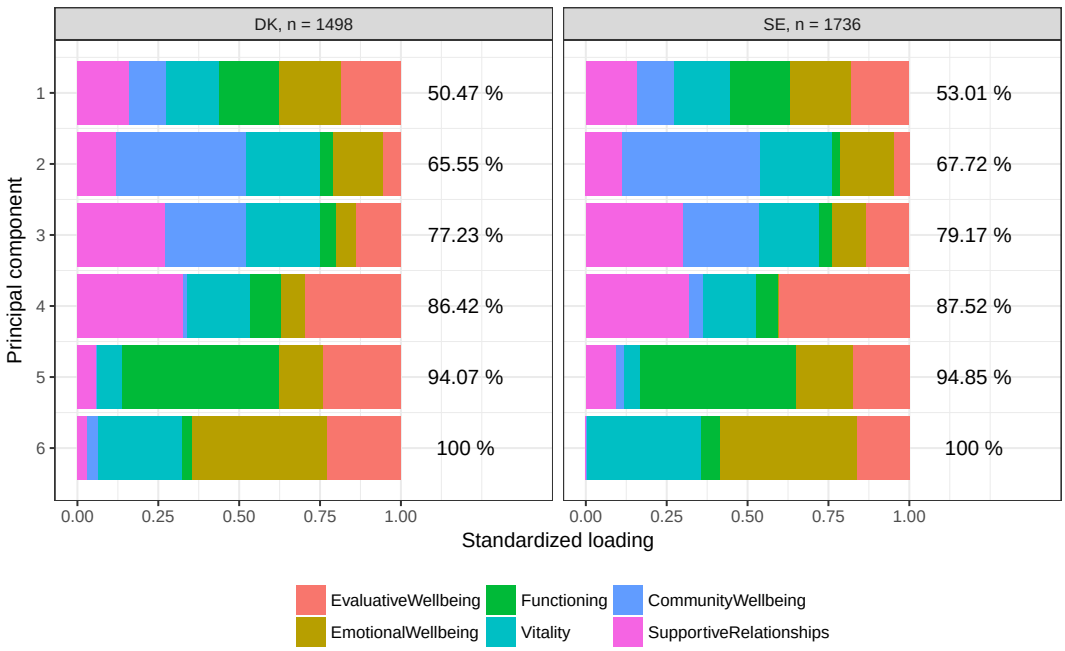

FIGURE 4.
A chroma plot for comparing the loading patterns of the Danish and the Swedish subsamples. Note that the bars for each component is annotated with its cumulative variance score (denoted $\tilde{\sigma}_i^c$ in the above), that is, how much variance can be explained by having information of this and the preceding components.

The limitations of the PCADSC procedures should also be sudied in more detail. It seems unlikely that the procedure would be able to disclose differences in scaling, since all variables are standardized in the procedure. This type of heterogeneity should thus be adjusted for in later analyses of the combined data set.

References

Asil, M. and Brown, G. T. L. (2016). Comparing OECD PISA Reading in English to Other Languages: Identifying Potential Sources of NonInvariance. *International Journal of Testing*, 16(1):71–93.

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.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, chapter 14.5.1. Springer, second edition.

Helliwell, J., Layard, R., and Sachs, J. (2016). World happiness report 2016 update. *New York: Sustainable Development Solutions Network*.

Jeffrey, K., Abdallah, S., and Quick, A. (2015). Europeans' personal and social wellbeing: Topline results from round 6 of the european social survey. *ESS Topline Results (Series 5)*.

Kankara, M. and Moors, G. (2014). Analysis of Cross-Cultural Comparability of PISA 2009 Scores. *Journal of Croos-Cultural Psychology*, 45(3):381–399.

Kreiner, S. and Christensen, K. B. (2014). Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, 79(2):210–231.

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

Lolle, H. L. and Andersen, J. G. (2016). Measuring happiness and overall life satisfaction: A danish survey experiment on the impact of language and translation problems. *Journal of Happiness Studies*, 17(4):1337–1350.

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.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez-Pose, A. and von Berlepsch, V. (2014). Social capital and individual happiness in europe. *Journal of Happiness Studies*, 15(2):357–386.

Veenhoven, R. (2012). Cross-national differences in happiness: Cultural measurement bias or effect of culture? *International Journal of Wellbeing*, 2(4):333–353.

A. Supplementary tables

Scale	Items
Evaluative wellbeing	How satisfied with life as a whole
	How happy are you
Emotional wellbeing	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 an 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 wellbeing	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

TABLE 2.

Relationship between questionnaire items and scales, as defined in Jeffrey et al. (2015). Note that 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.