Exploratory data structure comparisons

A few tools based on PCA

Anne H. Petersen, Bo Markussen & Karl Bang Christensen

March 15, 2017

Problem

- Let X_k be a $n_k \times d$ matrix of data for $k \in \{1, 2\}$ with (the same) ordinal or numeric variables for both k.
- We ask: Are the structures of X_1 and X_2 similar? Can we e.g. combine the two into one dataset $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with $n = n_1 + n_2$ observations of the d variables without problems?
- Example: X_1 stems from a survey administered by telephone, X_2 stems from the same survey administered as a web questionnaire. Can we analyze the "full" dataset jointly?
- Example: X_1 represents data from a study conducted in Denmark, X_2 represents the same measurements from a study in Bulgaria. Can we combine the two?

Requirements for a candidate method

- New origin stories of datasets add a new challenge to this question: Increasingly often, data is *not* collected with a specific endpoint in mind (e.g. PISA, the European Social Survey (ESS), traffic data from websites, ...)
  - $\Rightarrow$ We wish to answer the question *before* specifying what we are going to use the data for
- In particular, we do not want to specify a model (yet)
  - $\Rightarrow$ Cannot use e.g. IRT methods, regression methods
- We do not want just a test, but instead tools for intuitive, exploratory investigations that might lead to insights as to *why* the differences occur

Solution strategy

- Focus on the covariance structures:
  - Let $S_1 = \tilde{\mathrm{V}}(X_1)$ and $S_2 = \tilde{\mathrm{V}}(X_2)$ denote the empirical covariance matrices of the two datasets
  - S_k describes the interplay between variables. And for Gaussian (zero-mean) variables: A sufficient statistic for the joint distribution.
  - In principle, we could compare S_1 and S_2 entry by entry, but this strategy does not scale well
- Deconstruct each covariance matrix such that we can deal with the most informative components first and ignore noise

  $\rightarrow$ Use Principal Component Analysis (PCA)

Principal component analysis: A greedy interpretation

- Let $U \subset \mathbb{R}^d$ be the set of unit vectors of dimension d. For $j \in 1, ..., d$, define:

  The jth *loading vector*:

  $$\eta_j := \text{argmax}_{u \in U : u \perp \hat{K}_{j-1}} \tilde{\mathrm{V}}(u^\top X^\top)$$

  The jth *variance component*:

  $$\lambda_j := \tilde{\mathrm{V}}(\eta_j^\top X^\top)$$

  where $K_0 = \emptyset$ and $K_j = \text{span}\{\eta_1, ..., \eta_j\}$.
- Now, η_1 is the linear transformation of dimension $d \times 1$ of the data that explains the largest possible amount of the variance, and this amount is $\frac{\lambda_1}{trS}$.
- It holds that $S = \sum_{j=1}^d \lambda_j \eta_j \eta_j^\top$

Principal component analysis: A greedy interpretation

- Let $U \subset \mathbb{R}^d$ be the set of unit vectors of dimension d. For $j \in 1, ..., d$, define:

  The jth *loading vector*: (eigenvectors)

  $$\eta_j := \text{argmax}_{u \in U : u \perp \hat{K}_{j-1}} \tilde{\mathrm{V}}(u^\top X^\top)$$

  The jth *variance component*: (eigenvalues)

  $$\lambda_j := \tilde{\mathrm{V}}(\eta_j^\top X^\top)$$

  where $K_0 = \emptyset$ and $K_j = \text{span}\{\eta_1, ..., \eta_j\}$.
- Now, η_1 is the linear transformation of dimension $d \times 1$ of the data that explains the largest possible amount of the variance, and this amount is $\frac{\lambda_1}{trS}$.
- It holds that $S = \sum_{j=1}^{d} \lambda_j \eta_j \eta_j^\top$

Principle Component Analysis-based Data Structure Comparisons (PCADSC)

- PCADSC: Conduct PCA decomposition on each of the (standardized) datasets X_1 and X_2 and compare the results
- We provide three tools for visualizing the results of PCA in order to compare data structures:
  - The cumulated eigenvalue (CE) plot
  - The hair plot
  - The pancake plot
- Available in `R`-package at `www.github.com/AnnePetersen1/PCADSC`

Principle Component Analysis-based Data Structure Comparisons (PCADSC)

- PCADSC: Conduct PCA decomposition on each of the (standardized) datasets X_1 and X_2 and compare the results
- We provide three tools for visualizing the results of PCA in order to compare data structures:
  - The cumulated eigenvalue (CE) plot - compare eigenvalues
  - The hair plot
  - The pancake plot
- Available in R-package at `www.github.com/AnnePetersen1/PCADSC`

Principle Component Analysis-based Data Structure Comparisons (PCADSC)

- PCADSC: Conduct PCA decomposition on each of the (standardized) datasets X_1 and X_2 and compare the results
- We provide three tools for visualizing the results of PCA in order to compare data structures:
  - The cumulated eigenvalue (CE) plot - compare eigenvalues
  - The hair plot - explain S_1 from S_2 (and S_2 from S_1)
  - The pancake plot
- Available in `R`-package at `www.github.com/AnnePetersen1/PCADSC`

Principle Component Analysis-based Data Structure Comparisons (PCADSC)

- PCADSC: Conduct PCA decomposition on each of the (standardized) datasets X_1 and X_2 and compare the results
- We provide three tools for visualizing the results of PCA in order to compare data structures:
  - The cumulated eigenvalue (CE) plot - compare eigenvalues
  - The hair plot - explain S_1 from S_2 (and S_2 from S_1)
  - The pancake plot - compare loadings
- Available in `R`-package at `www.github.com/AnnePetersen1/PCADSC`

Principle Component Analysis-based Data Structure Comparisons (PCADSC)

- PCADSC: Conduct PCA decomposition on each of the (standardized) datasets X_1 and X_2 and compare the results
- We provide three tools for visualizing the results of PCA in order to compare data structures:
  - The cumulated eigenvalue (CE) plot - compare eigenvalues
  - The hair plot - explain S_1 from S_2 (and S_2 from S_1)
  - The pancake plot - compare loadings
- Available in `R`-package at `www.github.com/AnnePetersen1/PCADSC`
- First step of all procedures:
  1. For $k \in \{1, 2\}$, standardize X_k and perform PCA, thereby obtaining $\eta_1^k, ..., \eta_d^k$ and $\lambda_1^k, ..., \lambda_d^k$

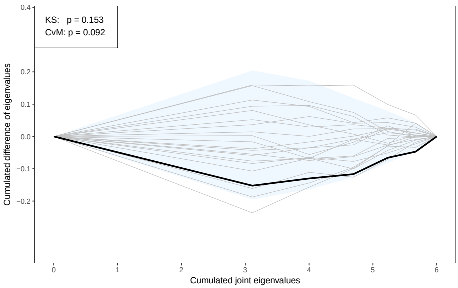
The CE plot

- Let $\lambda_1, ..., \lambda_d$ be the eigenvalues of the covariance matrix of the combined dataset, X.
- Draw a piecewise linear curve through the points

$$(0,0), (\lambda_1, \lambda_1^1 - \lambda_1^2), (\lambda_1 + \lambda_2, \lambda_1^1 + \lambda_2^1 - \lambda_1^2 - \lambda_2^2),$$
$$\dots, \left(\sum_{j=1}^{d} \lambda_j, \sum_{j=1}^{d} \lambda_j^1 - \lambda_j^2 \right).$$

- Get an idea of the variability of the results using repeated random splits:
  - E.g. 10000 times, divide X randomly into two subsets of size n_1 and n_2, respectively and perform the steps from above
  - Draw a shaded region representing a "bootstrapped" pointwise confidence interval obtained in this way
  - Draw a subset of the random curves (e.g. 20) to illustrate how they vary

The CE plot



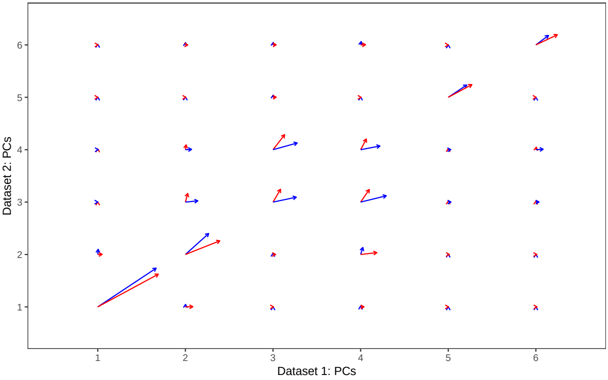
Annotations: p-values from two test statistics:

- Kolmogorov-Smirnov: $\max_{k=1,\dots,d} \left| \sum_{j=1}^{k} \lambda_j^1 - \lambda_j^2 \right|$
- Cramér-von Mises: $\sum_{k=1}^{d-1} \frac{\lambda_k + \lambda_{k+1}}{2} \left(\sum_{j=1}^{k} \lambda_j^1 - \lambda_j^2 \right)^2$

The hair plot

- Let $\lambda_{\max} = \max\{\lambda_1^1, \lambda_1^2\}$.
- Define $\mu_{jk} = \sqrt{\frac{\lambda_k^1}{\lambda_{\max}}} \left|(\eta_k^1)^\top \eta_j^2\right|$, $\nu_{jk} = \sqrt{\frac{\lambda_j^2}{\lambda_{\max}}} \left|(\eta_j^2)^\top \eta_k^1\right|$,
  $\theta_{jk} = \arccos\left(\left|(\eta_k^1)^\top \eta_j^2\right|\right)$
- In the jkth position in a $d \times d$ grid, draw two arrows (hairs):
  - A blue arrow with length μ_{jk} drawn at an angle of θ_{jk} counter-clockwise from the diagonal
  - A red arrow with length μ_{jk} drawn at an angle of θ_{jk} clockwise from the diagonal
- Note: If $\eta_k^1 \perp \eta_j^2$, then $\mu_{jk} = \nu_{jk} = 0$. If $\eta_k^1 = \eta_j^2$, then $\theta_{jk} = \arccos(1) = 0$ (due to unit length).
- Thus: Identical structures imply zero-length arrows for $j \neq k$ and identical red and blue arrows at the diagonal for $j = k$.

The hair plot



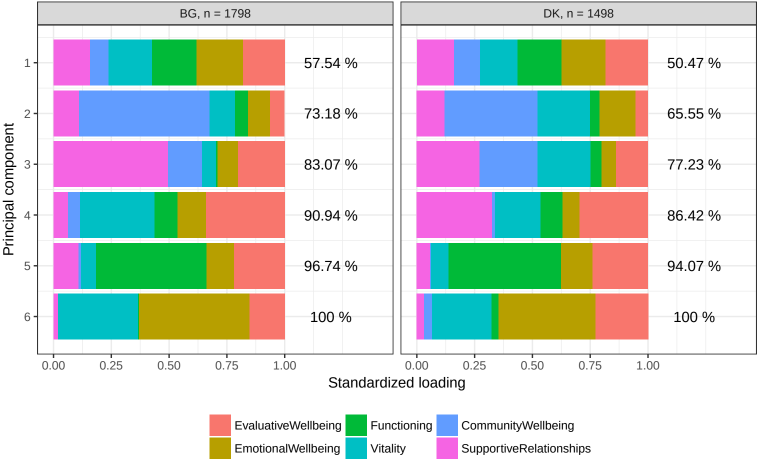
The pancake plot

For $k \in \{1, 2\}$, do:

2. For $i \in 1, ..., d$, normalize η_i^k and add a bar to the plot consisting of differently colored "pancakes" whose widths correspond to their standardized loading weights.
3. Annotate each bar with the amount of (cumulated) explained variance when using the information from this and the previous components, i.e.

$$\frac{\sum_{j=1}^{i} \eta_j^k}{\sum_{j=1}^{d} \eta_j^k}$$

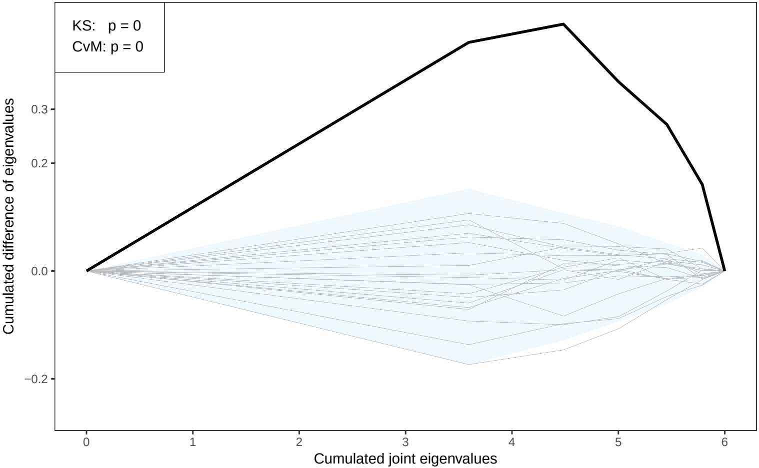
The pancake plot



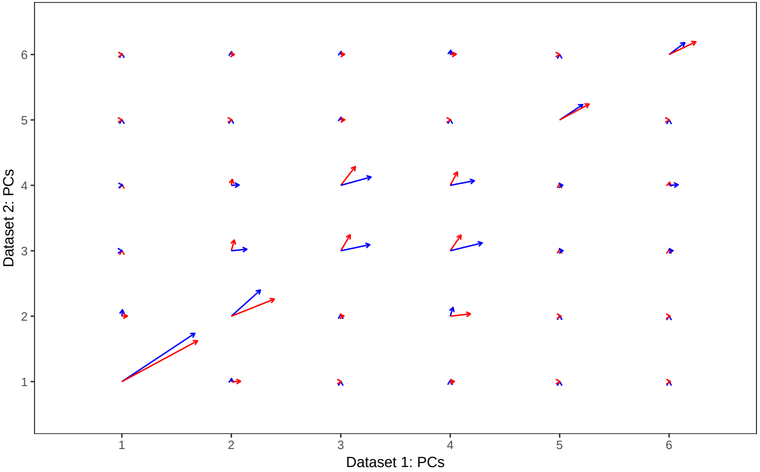
Data example - set-up

- Data on psychological well-being from the 2012 version of the European Social Survey (ESS)
- We use 35 items for producing 6 distinct psychological well-being scales
- ESS report: Bulgaria (BG) and Denmark (DK) are particularly different in the interplay between different aspects of psychological well-being (evaluated at aggregated country-level)
- Postulate: Denmark and Sweden (SE) are quite similar in what defines personal happiness and psychological well-being
- Strategy: Run all three methods on the DK vs. BG and DK vs. SE comparisons and expect to find *different* data structures for DK vs. BG and *similar* structures for DK vs. SE
- Only use complete cases. This results in $n_{\text{DK}} = 1498$, $n_{\text{BG}} = 1798$ and $n_{\text{SE}} = 1736$ observations, respectively.

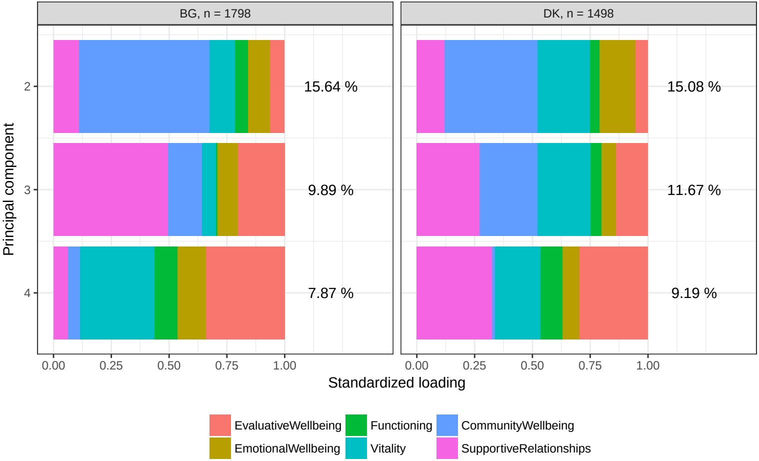
DK vs. BG: CE plot



DK vs. BG: hair plot

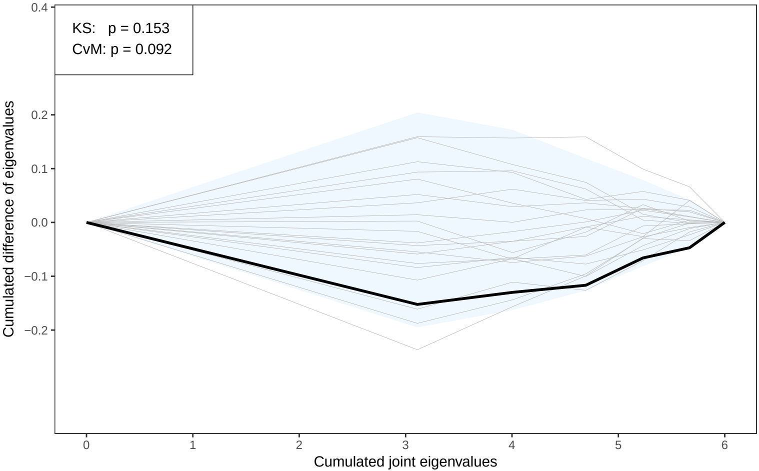


DK vs. BG: Pancake plot

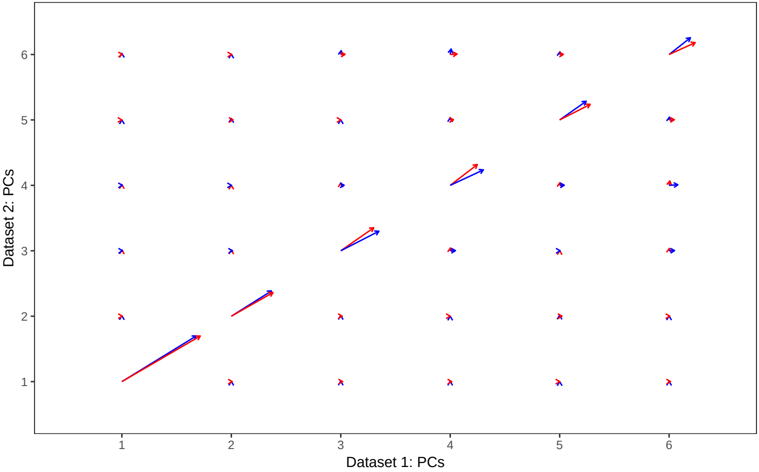


DK vs. BG: Pancake Wally plot

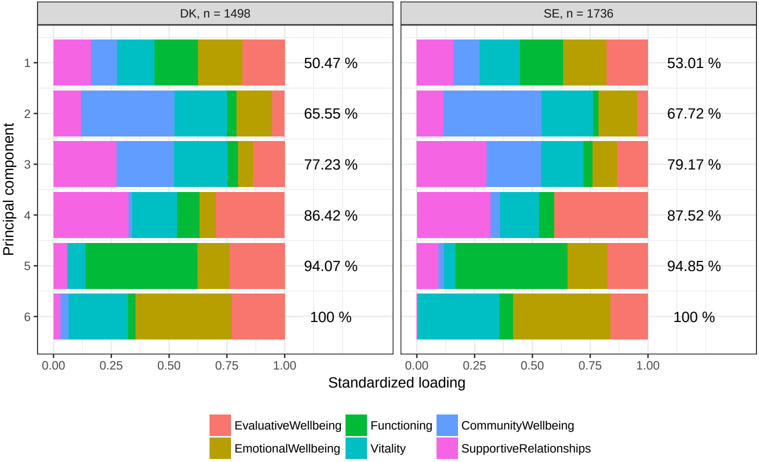
DK vs. SE: CE plot



DK vs. SE: hair plot



DK vs. SE: Pancake plot



DK vs. SE: Pancake Wally plot

Ideas for next steps

- Investigate sensitivity towards the sample sizes n_1 and n_2
- Limitations: What sorts of problems can never be found using PCADSC?
  - Differences in scaling, as we standardize all variables
  - More?
- Interpretation for binary variables?
- Any meaningful way to allow for nominal, categorical variables?
- ... ?