SOMETHING THAT IS NOT PCADSC

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

Abstract

abstract...

Key words: keywords...

1. Introduction

The origin stories of datasets have changed over time. While in the past, data was often collected with a specific scientific inquiry in mind, today, a lot of data is accumulated without such a specific purpose. This is due to vast amounts of data being registered online and due 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 e.g. differences in measurement instruments do not cause the data to be effectively divided into subsets that are simply not comparable?

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 - a very brief introduction

Principal component analysis (PCA) may be interpreted in terms of the *rank-q-reconstruction error*. To describe this consider n observations $x_1, \dots, x_n \in \mathbb{R}^d$ of d variables, and let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_n$ be their average. Suppose that we want to describe each of the observations by q numbers instead of the original d numbers. The 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

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

PCA ensures the existence of the 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 the empirical covariance $S = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top \in \mathbb{R}^d$. 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 $\eta_1, \dots, \eta_d \in \mathbb{R}^d$ of S in the columns. The eigenvalues are uniquely defined, and the eigenvectors are uniquely defined up to a sign whenever the eigenvalues are different. If some of the eigenvalue 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

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

and equals $\sum_{i=q+1}^d \lambda_i$. In the terminology of PCA the eigenvectors $\eta_j \in \mathbb{R}^d$ are called the *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 the *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

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

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 associated variation components λ_j describe the structure of the dataset simultaneously for all approximating dimension q. This implies that the loadings and variation components can be used to investigate the structure of the dataset 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 we can interpret the eigenvector/eigenvalue-pairs of the covariance matrix as tools for obtaining a different representation of the dataset. Specifically, we can interpret the PCA loadings (eigenvectors) as weights that determine the relative influence of each variable in each component. If two different data sets with the same variables, but different samples of observations, then have similar loading patterns, the two data sets will agree on which variables explain the most variance and also *how* this variance is explained. Looking at the PCA loadings and the cumulative explained variance thus provides a straightforward non-parametric graphical approach for assessing similarity between two datasets.

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

- PISA

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

Ackerman, T.A. (1992). EXAMPLE POST. A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.