

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 5 – Dimensionality Reduction
Wednesday 20th January 2016

1. Halfway Surveys
2. What is Dimensionality Reduction?
3. Why reduce dimensions?
4. What is Principal Component Analysis?
5. Lab
6. Talks
7. Discussion

DATA SCIENCE PART TIME COURSE

HALFWAY POINT SURVEYS

- › Help us continue to facilitate what is working throughout the course
- › We value student feedback
- › We're halfway through so we can correct anything that's not going well (or keep doing the things that are working)
- › We will sit down and discuss themes from the surveys and action points in review after surveys are completed

On the numerical scale questions, 10 is the highest and 1 is the lowest.



DATA SCIENCE PART TIME COURSE

DIMENSIONALITY REDUCTION

scikit-learn algorithm cheat-sheet

START

classification



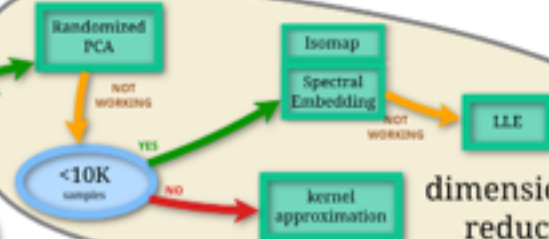
clustering



regression



dimensionality reduction



Back

scikit
learn

A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

In general, the idea is to regard the dataset as a matrix and to decompose the matrix into simpler, meaningful pieces.



- The number of features in our dataset can be difficult to manage, or even misleading (eg, if the relationships are actually simpler than they appear).
- Ideally, we would like to eliminate redundancy and consolidate the number of variables we're looking at.
- The complexity that comes with a large number of features is due in part to the curse of dimensionality.

- The complexity that comes with a large number of features is due in part to the curse of dimensionality.
- Namely, the sample size needed to accurately estimate a random variable taking values in a d -dimensional feature space grows exponentially with d (almost).
- (More precisely, the sample size grows exponentially with $1 \leq d$, the dimension of the manifold embedded in the feature space).

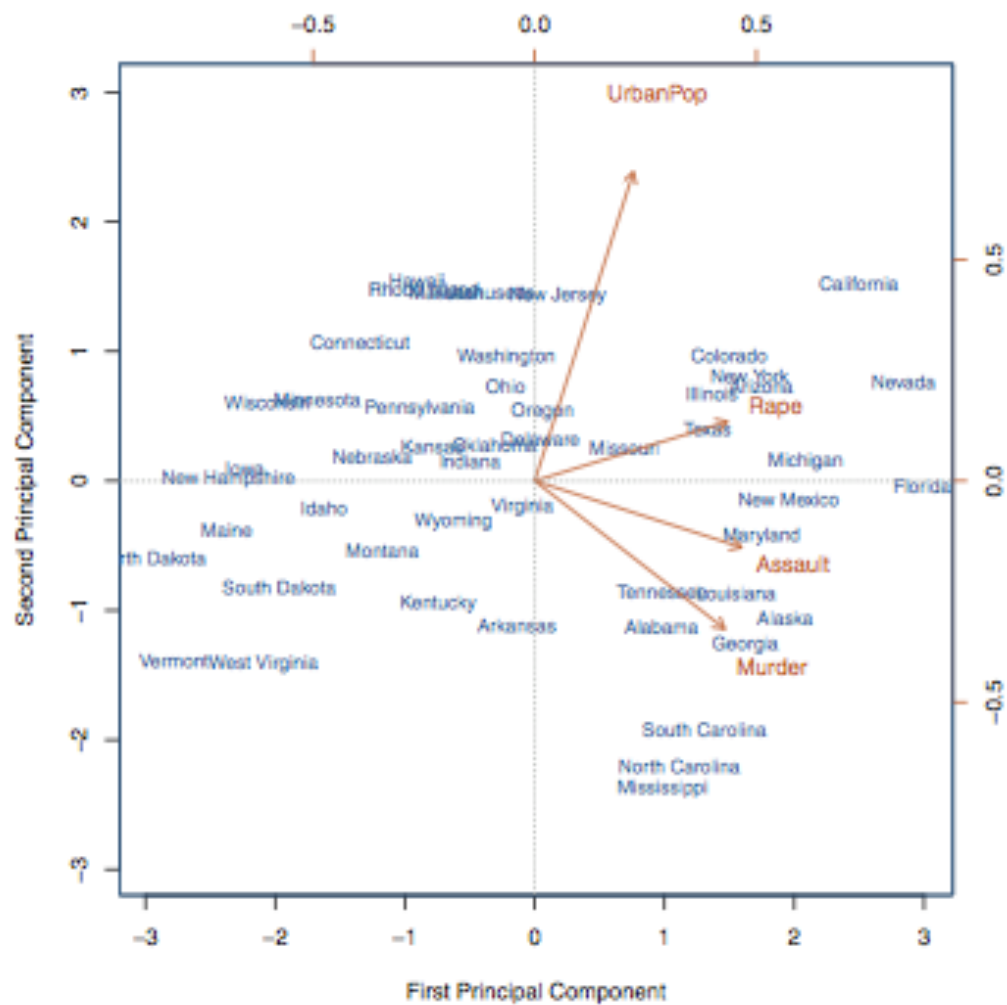
- We'd like to analyze the data using the most meaningful basis (or coordinates) possible.
- More precisely: given an $n \times d$ matrix X (encoding n observations of a d -dimensional random variable), we want to find a k -dimensional representation of X ($k < d$) that captures the information in the original data, according to some criterion.

DATA SCIENCE PART TIME COURSE

PRINCIPAL COMPONENTS

‘It finds a low-dimensional representation of a data set that contains as much as possible of the variation. The idea is that each of the n observations lives in p -dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the p features.’

- Introduction to Statistical Learning



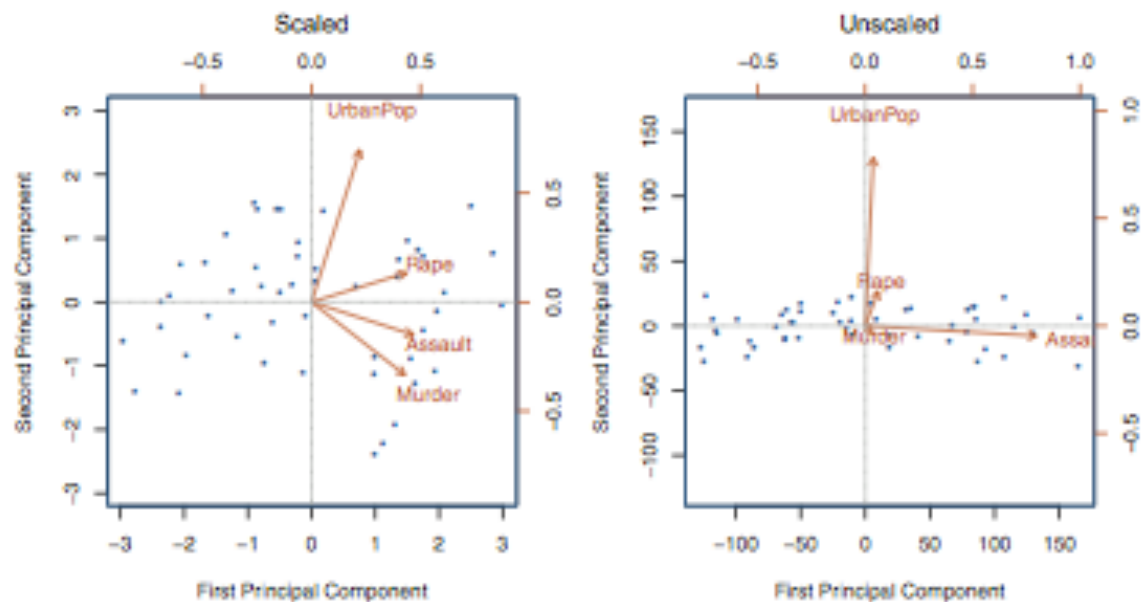
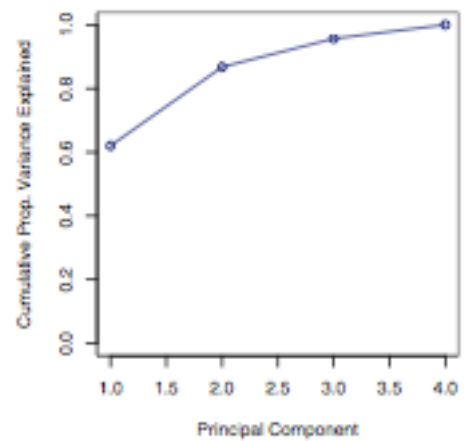
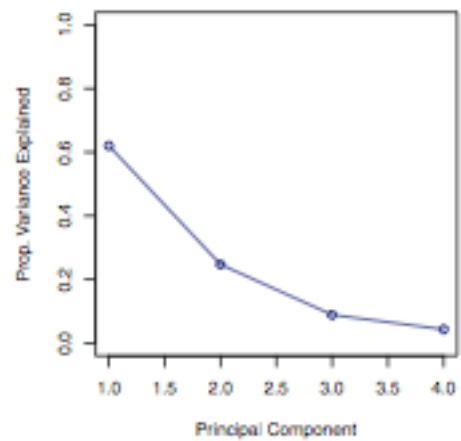


FIGURE 10.3. Two principal component biplots for the `USArrests` data. Left: the same as Figure 10.1, with the variables scaled to have unit standard deviations. Right: principal components using unscaled data. `Assault` has by far the largest loading on the first principal component because it has the highest variance among the four variables. In general, scaling the variables to have standard deviation one is recommended.



DATA SCIENCE PART TIME COURSE

LAB




```
git remote -v
```

```
git remote add upstream https://github.com/ihansel/SYD_DAT_3.git
```

```
git remote -v
```

```
git fetch upstream
```

```
git checkout master
```

```
git merge upstream/master
```

```
OR git reset --hard upstream/master
```



DATA SCIENCE - Week 5 Day 1

DISCUSSION TIME

- **Homework**
- **Project & Hackday**
- **(Another) Survey**

DATA SCIENCE - Week 4 Day 1



Just work on your project and read the following

- **Chapter 8 of Introduction to Statistical Learning, Tree Based Methods (30 pages)**

DATA SCIENCE – Week 5 Day 2

Project

- **I will be checking progress over the weekend and making some recommendations and asking questions.**
- **Have most recent progress up on github by Saturday**

Hackday

- **Opportunity to work on project, cover any material you want to review, & maybe a few bonus topics ?????**
- **Who's (really) keen ?**
- **When suits everyone ?**
- **Saturday 30th January ?**

