Principal Component Regression. PLS.

Mathematical Tools for ITS (11MAI)

Mathematical tools, 2021

Jan Přikryl

(based on the book "Introduction to Statistical Learning", `https://www.statlearning.com/`)

11MAI, lecture 8

Monday, Novemeber 8, 2021

version: 2021-11-08 13:09

Department of Applied Mathematics, CTU FTS

Lectue Contents

Regression methods

- Principal Component Regression
- Matlab session **??**
- Matlab session 8.2
- Partial Least Squares
- Matlab session 8.3
- Matlab session 8.4

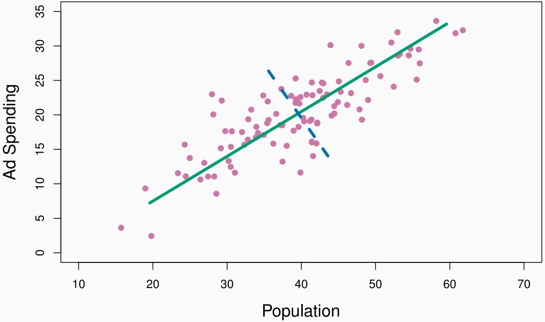
Assignment 4

Dimensionality reduction: Principal Components Regression

We said that a clever linear combination of original predictors may result in a set of synthetic predictors with lower variance.

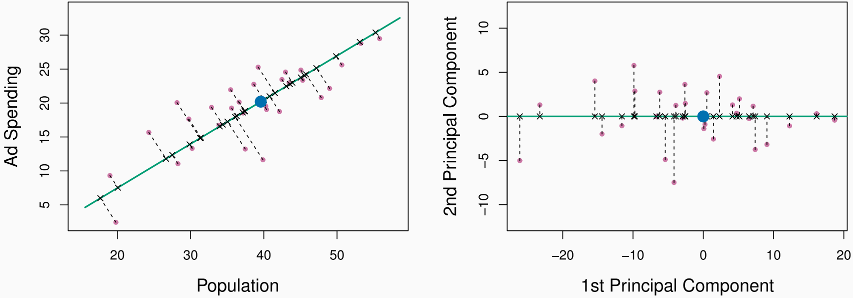
- Let's apply *principal components analysis (PCA)* to define the linear combinations of the predictors that we will use in our regression.
- The first principal component is the (normalized) linear combination of the predictors with the largest variance. The second principal component has the largest variance, subject to being uncorrelated with the first. And so on . . .
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

Pictures of PCA



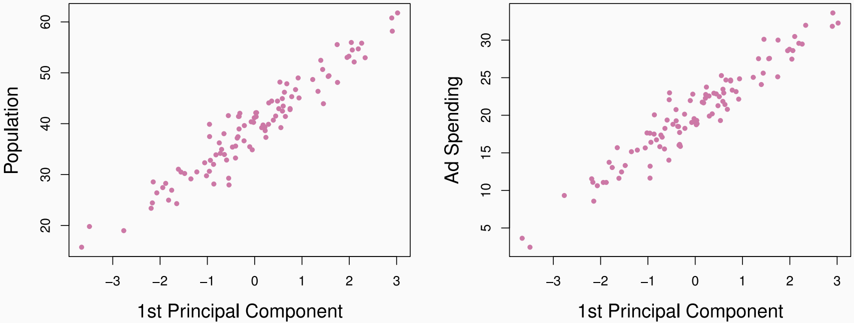
Advertising data: The population size (`pop`) *and ad spending* (`ad`) *for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Pictures of PCA: continued



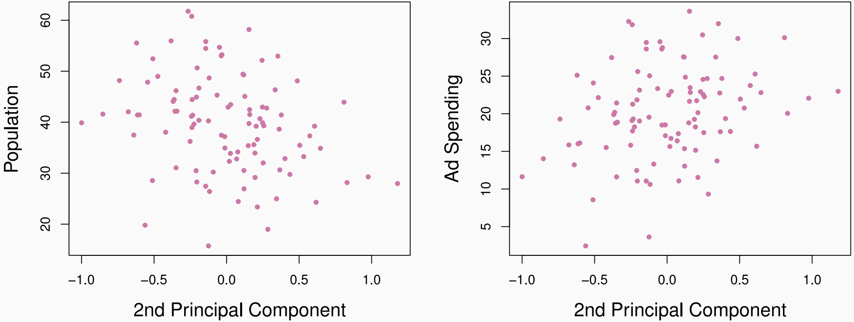
A subset of the advertising data. Left: The first principal component, minimizing the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments. Right: The left-hand panel has been rotated so that the first principal component lies on the x-axis.

Pictures of PCA: continued



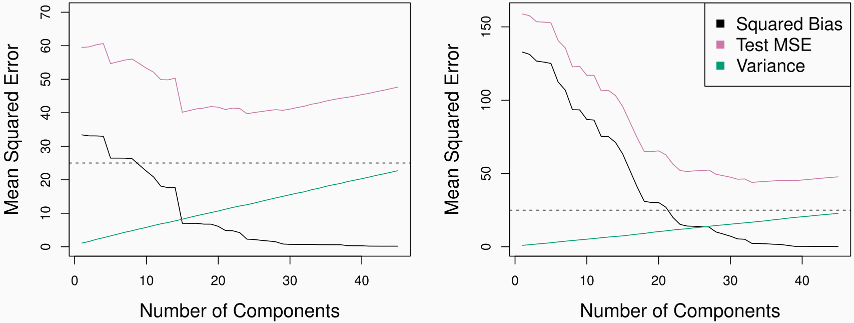
Plots of the first principal component scores z_{i1} versus `pop` *and* `ad`*. The relationships are strong.*

Pictures of PCA: continued



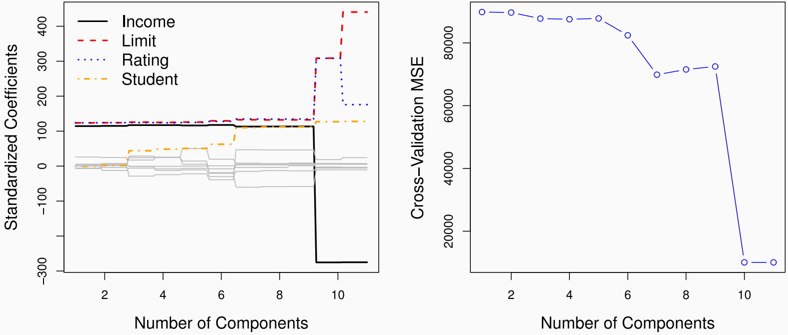
Plots of the first principal component scores z_{i2} versus `pop` *and* `ad`. *The relationships are weak.*

Application to Principal Components Regression



PCR was applied to two simulated data sets. The black, green, and purple lines correspond to squared bias, variance, and test mean squared error, respectively. Left: Simulated data from slide **??**. *Right: Simulated data from slide* **??**.

Choosing the number of directions M



Left: *PCR standardized coefficient estimates on the* `Credit` *data set for different values of M.* Right: *The 10-fold cross-validation MSE obtained using PCR, as a function of M.*

Important drawback of PCR

PCR reduces the number of "important" predictors.

- It identifies linear combinations, or *directions*, that best represent the predictors $X_1, \ldots, X_p$.
- These directions are identified in an *unsupervised* way, since the response Y is not used to help determine the principal component directions.
- That is, the response *does not supervise* the identification of the principal components.
- Consequently, PCR suffers from a potentially serious drawback:
  There is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Matlab Session 8.1

Matlab Session 8.1 — Principal Component Regression on Synthetic data

Matlab Session 8.2

Matlab Session 8.2 — Principal Component Regression on Hitters dataset

Partial Least Squares (PLS)

PLS is a dimension reduction method, which (like PCR) first identifies a new set of features $Z_1, \ldots, Z_M$ that are linear combinations of the original features $X_1, \ldots, X_p$, and then fits a linear model via OLS using these M new features.

- But unlike PCR, PLS identifies these new features in a *supervised* way – that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response.
- Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

Details of Partial Least Squares

After standardizing the p predictors, PLS computes the first direction Z_1 by setting each ϕ_{1j} in

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j$$

equal to the coefficient from the simple linear regression of Y onto X_j.

- One can show that this coefficient is proportional to the correlation between Y and X_j.
- Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$ PLS places the highest weight on the variables that are most strongly related to the response.
- Subsequent directions are found by taking residuals and then repeating the above prescription.

Matlab Session 8.3

Matlab Session 8.3 — Partial Least Squares on Synthetic data

Matlab implements partial least squares (PLS) using the `plsregress()` function from the Statistics and Machine Learning Toolbox.

```
num_components = size(data_x, 2);
num_cv_folds = 10;
[XL,yl,XS,ys,beta,varexpl,cv_mse] = plsregress(...
    data_x, data_y, num_components, ’cv’, num_cv_folds);
```

Matlab Session 8.4

Matlab Session 8.4 — Partial Least Squares Regression

Matlab implements partial least squares (PLS) using the `plsregress()` function from the Statistics and Machine Learning Toolbox.

```
num_components = size(data_x, 2);
num_cv_folds = 10;
[XL,yl,XS,ys,beta,varexpl,cv_mse] = plsregress(...
     data_x, data_y, num_components, ’cv’, num_cv_folds);
```

Lectue Contents

Regression methods

Assignment 4

Assignment 4.1 — What's the frequency, Kenneth?

a) Review the end of Lecture #4 and Matlab Sessions 5.1 and 5.2 to refresh your knowlege about spectrograms.

b) Download both recorder recordings from the lecture website:

```
[x1 sr1] = audioread('recorder1.wav');
[x2 sr2] = audioread('recorder2.wav');
```

c) Your task is related to `spectrogram` usage in non-visual mode,

```
[s,w,t] = spectrogram(...);
```

where `s` is a matrix of FFT coefficients, `w` vector of normalised frequencies, and `t` vector of time stamps where particular FFTs have been computed.

Assignment 4.1 — What's the frequency, Kenneth?

d) Both recording contain an octave played on soprano recorder (staccato, legato). Your task is to analyse `s` and use `w` to find out *particular tones* of the recording — you may verify your findings visually, but the core of this assignment is the computational procedure that will analyse the spectrogram matrices.

e) For each recording do the following:

   i. Identify particular base frequencies for each tone.

   ii. Identify the significant harmonics.

   iii. Estimate which tone is being played, i.e., C_2, Fis$_3$ etc.

   iv. Plot the spectrogram.

f) Summarise your findings from e) in a written report.

Homework rules

Submit your report by Friday, November 19, 2021 using the web page `http://zolotarev.fd.cvut.cz/mni`

Solution report should be formally correct (structuring, grammar).

Images should be vectors, not bitmaps.

Only `.pdf` files are acceptable. Handwritten solutions and `.doc` and `.docx` files will not be accepted.

Solutions written in TEX (using LyX, Overleaf, whatever) may receive small bonification.