Principal Component Analysis, Principal Component and Partial Least Squares Regression - An overview

Ana Alina Tudoran

Assoc.Prof.PhD

Aarhus University, Denmark

anat@econ.au.dk

October 1, 2023

Learning objectives

- Understand the "Curse of Dimensionality"
- Understand how the Principal Components Analysis works
- Compare and contrasts Principal Components Regression (PCR) and Partial Least Squares Regression (PLS) as dimension reduction methods
- Implement, interpret and visualize the results obtained from PCA, PCR and PCR in R

Agenda for today

1. Principal Component Analysis (PCA)

2. How does PCA work?

3. Application PCA in R

Curse of Dimensionality

- The Curse of Dimensionality refers to the exponential increase in data volume as the number of dimensions (features) in a dataset grows
- *Consequences* of high dimensionality
  - Increased data sparsity: Data points become more scattered, making it harder to find meaningful patterns
  - Computational complexity: Analyzing high-dimensional data requires more time and resources
  - Overfitting: Models may perform well on training data but poorly on unseen data

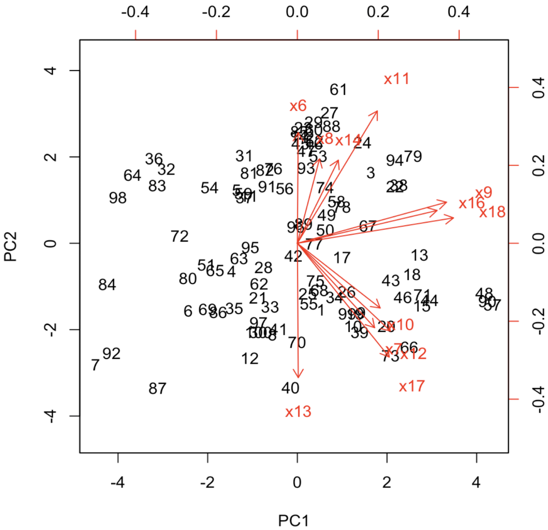
Main purpose PCA

When faced with a *large set of correlated variables*, PCA allows us *to visualize/summarize/reduce them* into a smaller number of representative variables (called "Principal Components") that collectively explain most of the variability in the original data set. Therefore it is called a *dimension reduction* method.

PCA for data visualization

- Consider a data set of p features, $X_1, X_2, ...X_p$
- To visualize these p features, we could draw two-dimensional scatter-plots
- Too many and not informative
- PCA aims to find a lower dimensional representation of the data set that captures most of the information

Example



PCA for variable grouping

- Consider a data set of p consumer perceptions about a company, $X_1, X_2, ...X_p$
- To better understand the relationships between these p features, we could look at their correlation matrix
- Bivariate correlation too many, not too informative overall
- PCA aims to find groups of highly correlated variables and assign them a name

PCA for dimension reduction in ML (example)

```
# blueprint
ames_recipe <- recipe(Sale_Price ~ ., data = ames_train) %>%
  step_nzv(all_nominal())  %>%
  step_integer(matches("Qual|Cond|QC|Qu")) %>%
  step_center(all_numeric(), -all_outcomes()) %>%
  step_scale(all_numeric(), -all_outcomes()) %>%
  step_pca(all_numeric(), -all_outcomes())
ames_recipe
# prepare
prepare <- prep(ames_recipe, training = ames_train)
prepare
```

PCA key characteristics

- Exploratory: no generally accepted mechanism to validate the model
- Subjective: interpreting the results of such analyses can be problematic
- PCA **looks similar** to Exploratory Factor Analysis (EFA)
- PCA and EFA are applied to discover which variables in the set form *coherent subsets relatively independent of one another*

Agenda

1. Principal Component Analysis (PCA)

2. How does PCA work?

3. Application PCA in R

How PCA works?

- Consider a data set of p observable variables, $X_1, X_2, ...X_p$ in a dataset with n observations
- Let $Z_1, Z_2, ...Z_M$, principal components, representing linear combinations of the original variables

$$Z_m = \sum_{j=1}^{p} \Phi_{jm} X_j$$

- Φ are called **loadings**

Loadings in R

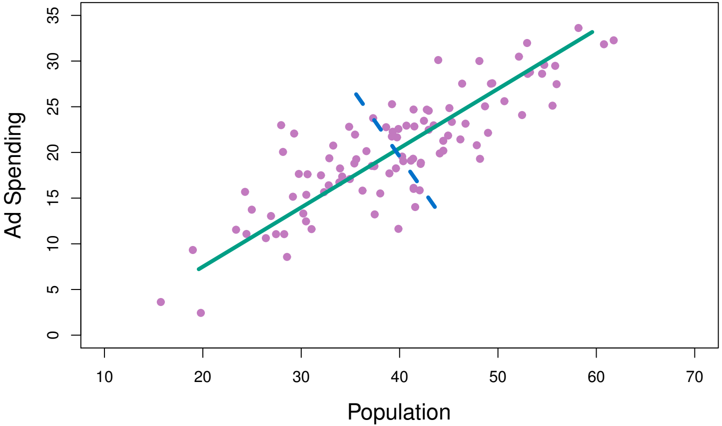
Matrix rotation representing the loadings of x6-x18 original variables on the first five PCs:

```
Rotation (n x k) = (12 x 12):
            PC1         PC2         PC3         PC4         PC5
x6  0.002611283  0.35500899 -0.03639396  0.47488094  0.05809925
x7  0.238813037 -0.27073718  0.35271763  0.33589229  0.39575861
x8  0.068434884  0.27050596  0.52702093 -0.36346781 -0.05088079
x9  0.461195683  0.13192382 -0.20484606 -0.09528049  0.05182484
x10 0.256010190 -0.20764489  0.19079590  0.33909954 -0.79274678
x11 0.247041802  0.42445152 -0.09615687  0.23256221  0.15788424
x12 0.287599154 -0.28132461  0.35505139  0.28248482  0.20450204
x13 0.002933242 -0.43089779 -0.01489212 -0.15317167  0.32819325
x14 0.126661159  0.26654967  0.53237332 -0.31928271 -0.02410275
x16 0.431889142  0.10522420 -0.16340420 -0.15343830  0.07379335
x17 0.284074668 -0.36502029 -0.16307116 -0.34329733 -0.16015346
x18 0.482715527  0.08148635 -0.21425640 -0.06932900  0.01194169
```

Interpretation of components

- The *1st* principal component, Z_1, is that along which the observations *vary most*
  - If we project the data on the line represented by Z1 (i.e. PC1), then the resulting projected observations would have the largest possible variance
- The *2nd* principal component, Z_2 is that along which the observations vary second most, subject to the *constraint* that Z_1 uncorrelated with Z_2.
- so on..

Example with 2 variables and 2PCs extracted



PC1

- The 1st PC mathematically is:

$$Z_1 = 0.839 \times (pop - \overline{pop}) + 0.544 \times (ad - \overline{ad})$$

imposing:

$$\Phi_{11}^2 + \Phi_{21}^2 = 1$$

- This particular linear combination also defines the line that is closest to all n of the observations.

PC2

- The 2st PC mathematically is:

$$Z_2 = 0.544 \times (pop - \overline{pop}) - 0.839 \times (ad - \overline{ad})$$

imposing:

$$\Phi_{1}{}_1^2 + \Phi_{2}{}_1^2 = 1$$

- This particular linear combination is that along which the observations vary second most, subject to the *constraint* that Z_1 uncorrelated with Z_2.

What PC scores are?

- New variables, called **scores**, are constructed as weighted averages of the original variables
- In our example:

$$z_i = 0.839 \times (pop_i - \overline{pop}) + 0.544 \times (ad_i - \overline{ad})$$

$$z_i = 0.544 \times (pop_i - \overline{pop}) - 0.839 \times (ad_i - \overline{ad})$$

for all i=1:N in the dataset.

- In R: **matrix x**

How important is each PC?

- Check the proportion of variance explained by each PC

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.8721 1.7399 1.3178 1.1319 0.78636
Proportion of Variance 0.2921 0.2523 0.1447 0.1068 0.05153
Cumulative Proportion  0.2921 0.5443 0.6891 0.7958 0.84734
                          PC6     PC7     PC8     PC9    PC10
Standard deviation     0.7454 0.67782 0.53969 0.46246 0.41499
Proportion of Variance 0.0463 0.03829 0.02427 0.01782 0.01435
Cumulative Proportion  0.8936 0.93193 0.95620 0.97402 0.98837
                          PC11    PC12
Standard deviation     0.36269 0.08930
Proportion of Variance 0.01096 0.00066
Cumulative Proportion  0.99934 1.00000
```

Proportion of variance explained (mathematically)

- Total Variance present in the data set (assuming the variables were standardized):

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

- Variance Explained (VE) by each component, m:

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{jm} x_{ij} \right)^2$$

- The proportion of VE of any component is:

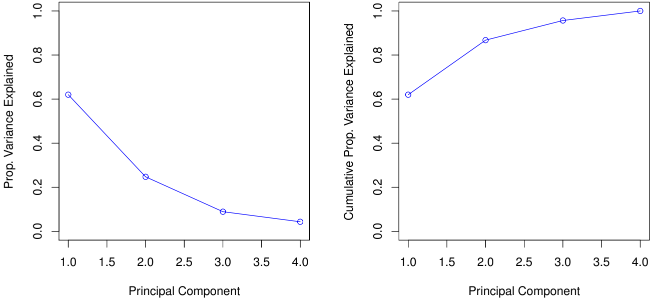
$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

How many PCs to select?

Criteria:

1. Scree Test criterion
   - Elbow rule
   - Eigenvalues >1
2. Cumulative proportion of variance explained is more then .60
3. A-priori: based on previous studies
4. In combination with supervised techniques using cross-validation (e.g. Section 6.3.1 ISL)

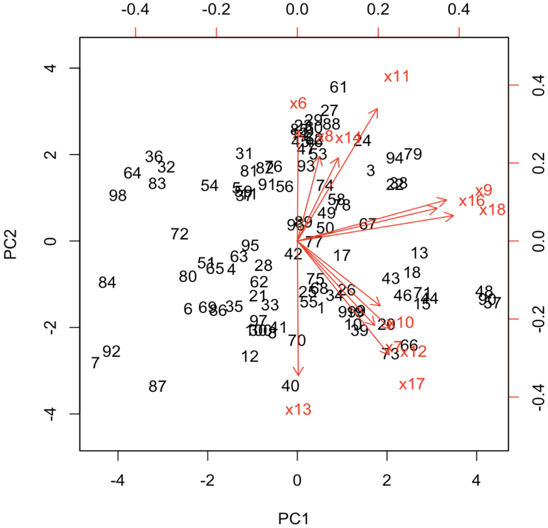
Screeplot in R



Interpret PCA visually

- **Biplot**: to see visually how data and variables are mapped in regard to the principal components discovered

Example



Assumptions behind standard PCA

- Measurement scales: continuous
- Sample size: at least 5 observations (ideally 10-20) per observed variable and at least 100 observations overall
- Linear relationship between observed variables
- Normal distribution for each observed variable

Other dimension reduction techniques

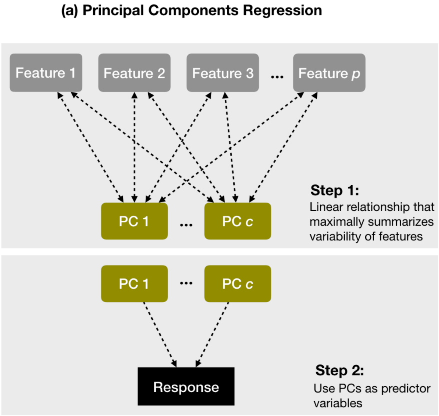
- A plethora of procedures to combine the original variables for producing new ones exist:
  - PLS: as in PCA, PC are **linear** combinations and require continuous scales
  - Discriminant Analysis: PC are **linear** combinations but require discrete scales
  - NLPCA - Nonlinear PCA: PC are **nonlinear** combinations with different types of measurement levels
  - Autoencoders (AEs) - have emerged as an alternative for conducting **nonlinear** feature fusion)

Application PCA in R

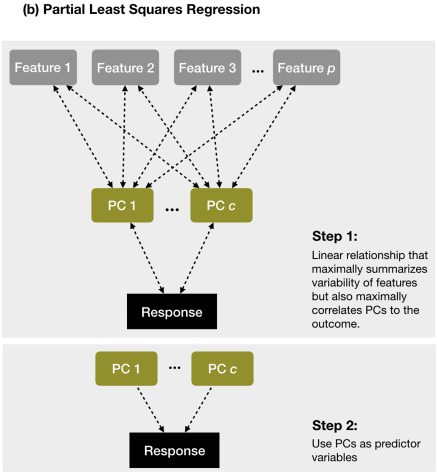
1. Check if PCA applies to your data
2. Run PCA model with **prcomp(dataset, scale=TRUE)**
3. Evaluate the rotation matrix (loadings)
4. summary() to evaluate the variance explained by each PCs
5. screeplot() to decide how many PCs to use
6. biplot() to visualize and interpret

Go to ISL (2013) Ch. 10, 10.4 Lab 1, Applied Ex. 10(a,b) in R file or ISL (2021) Ch 12, Lab 12.5.1., Applied Ex. 10(a,b) in R file

Principal Components Regression



Partial Least Squares Regression



Main references

- James et al. (2013, 2021)
  An Introduction to Statistical Learning, Ch. 6.3.1., Ch.10.1, 10.2 or Ch. 12 *Springer Texts in Statistics*

- Boehmke, B. and Greenwell, B. (2020). Hands-On Machine Learning with R, Ch. 4.6 and 4.7 *CRC Press*