Principal Component Analysis

Jeffrey Woo

MSDS, University of Virginia

1. Principal Component Analysis

2. Worked Example

3. Choosing Number of PCs

Principal Component Analysis

PCA is a dimension reduction method, as it seeks to produce a low(er)-dimensional representation of the data.

- Take a large set of correlated variables and replace them with a smaller number of (uncorrelated) variables that collectively can explain **most of the variability** in the variables.
- The uncorrelated variables are the principal components (PCs), which are directions in the variable space along which the original data are most variable.

PCA seeks a small number of variables that can explain most of the variation in the variables. Each dimension will be a linear combination of the k variables.

Finding the first PC

The first PC of a set of variables $X_1, \cdots, X_k$ is the **normalized linear combination** of the variables

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{k1}X_k \tag{1}$$

that has the largest variance, subject to the constraint $\sum_{j=1}^{k} \phi_{j1}^2 = 1$.

Finding the First PC

- The elements $\phi_{11}, \cdots, \phi_{k1}$ in (1) are called the **loadings** of the first PC.
- The loadings make up the principal component loading **vector** $\phi_1 = (\phi_{11}, \phi_{21}, \cdots, \phi_{k1})^T$.

Computing the First PC

- Let **X** denote a $n \times k$ set of k variables, where each variable has been centered to have mean 0.
- Look for the linear combination of the variables of the form

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{k1} x_{ik} \tag{2}$$

  for $i = 1, \cdots, n$ that has the largest sample variance, subject to the constraint $\sum_{j=1}^{k} \phi_{j1}^2 = 1$.
- Since the variables are centered, z_{i1} has mean 0, for any values of ϕ_{j1}.
- Hence, the sample variance of z_{i1} can be written as $\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$.

Computing the First PC

Thus, the first principal component loading vector solves the following optimization problem: maximize

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{k}\phi_{j1}x_{ij}\right)^2 \tag{3}$$

subject to $\sum_{j=1}^{k}\phi_{j1}^2 = 1$.

Computing the First PC

The optimization problem can be solved using standard techniques in linear algebra (singular-value decomposition).

- Z_1 is called the first principal component,
- $z_{11}, \cdots, z_{n1}$ are scores of the first PC for each observation.

Geometric Interpretation of the First PC

- The loading vector ϕ_1 defines a direction in the variable space along which the variables vary the most.
- If we project the n observations onto this direction, the projected values are the scores $z_{11}, \cdots, z_{n1}$.

Geometric Interpretation of the First PC

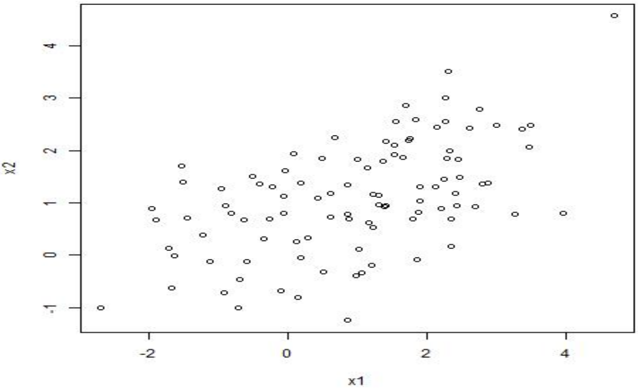

Figure 1: Data Simulated from Bivariate Normal

Other PCs

- The second PC is the linear combination of $X_1, \cdots, X_k$ that has maximum variance among all linear combinations that are **uncorrelated** with Z_1.
- The 2nd PC takes the form

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{k2}x_{ik}$$

where ϕ_2 is the second principal component loading vector and $\sum_{j=1}^{k} \phi_{j2}^2 = 1$.

Other PCs

- Constraining Z_2 to be uncorrelated to Z_1 is equivalent to constraining the direction ϕ_2 to be **orthogonal** to the direction of ϕ_1.
- Subsequent PCs are constructed similarly: they are orthogonal to all previous PCs and the sum of the squared loadings equal to 1.
- There are at most $min(n-1, k)$ PCs.

Geometric Interpretation of the PCs

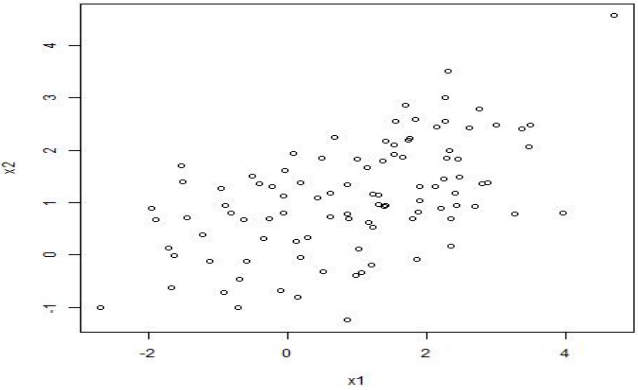

Figure 2: Data Simulated from Bivariate Normal

Alternate Interpretation of PCs

- The first principal component loading vector has another nice property: it defines the line in a *k*-dimensional space that is **closest** to the *n* observations (using average squared Euclidean distance).

Question: Does this sound a little like the least squares criterion in ordinary least squares regression?

Geometric Interpretation of the First PC

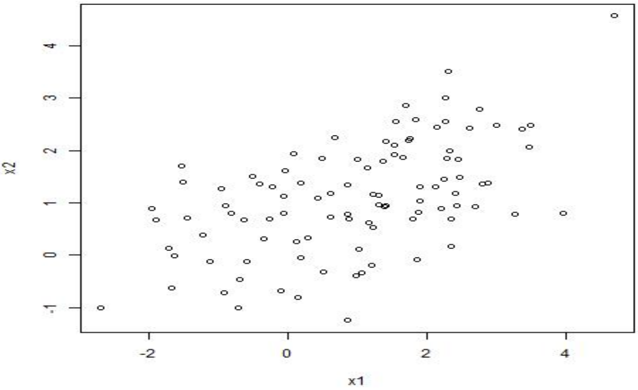

Figure 3: Data Simulated from Bivariate Normal

Practical Considerations

- Each variable should be **standardized** so they each have mean 0 and standard deviation 1. This is to prevent any variable from dominating the direction of the loading vectors.
- If the variables are measured on the same units, standardizing may not be needed.
- PCA should only be applied to **quantitative** variables.

1 Principal Component Analysis

2 Worked Example

3 Choosing Number of PCs

Worked Example: Gas Mileage

We will use the mtcars data set that is available in R. The data set contains information about fuel consumption and 10 aspects of automobile design and performance for 32 classic vehicles.

Worked Example: Gas Mileage

```
> head(mtcars, 8)
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
```

Worked Example: Gas Mileage

mpg:	Miles/(US) gallon
cyl:	Number of cylinders
disp:	Displacement (cu.in.)
hp:	Gross horsepower
drat:	Rear axle ratio
wt:	Weight (1000 lbs)
qsec:	1/4 mile time
vs:	Engine (0 = V-shaped, 1 = straight)
am:	Transmission (0 = automatic, 1 = manual)
gear:	Number of forward gears
carb:	Number of carburetors

Worked Example: mtcars

Below, we have the loadings for the PCs.

```
> round(pr.out$rotation,3)
        PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9
mpg  -0.393  0.028 -0.221 -0.006 -0.321  0.720 -0.381 -0.125  0.115
cyl   0.403  0.016 -0.252  0.041  0.117  0.224 -0.159  0.810  0.163
disp  0.397 -0.089 -0.078  0.339 -0.487 -0.020 -0.182 -0.064 -0.662
hp    0.367  0.269 -0.017  0.068 -0.295  0.354  0.696 -0.166  0.252
drat -0.312  0.342  0.150  0.846  0.162 -0.015  0.048  0.135  0.038
wt    0.373 -0.172  0.454  0.191 -0.187 -0.084 -0.428 -0.198  0.569
qsec -0.224 -0.484  0.628 -0.030 -0.148  0.258  0.276  0.356 -0.169
gear -0.209  0.551  0.207 -0.282 -0.562 -0.323 -0.086  0.316  0.047
carb  0.245  0.484  0.464 -0.214  0.400  0.357 -0.206 -0.108 -0.320
```

Worked Example: Interpreting Loading Vectors

- The first loading vector places **slightly higher weights** on mpg, cyl, disp, hp, wt.
- PC1 roughly corresponds to a measure of size of engines.
- The 2nd loading vector places more weight on drat, qsec, gear, carb.
- PC2 roughly corresponds to how powerful a car is.
- Interpreting later PCs is often **challenging**.

Worked Example: Biplot

A biplot displays the principal component scores and principal component loadings.

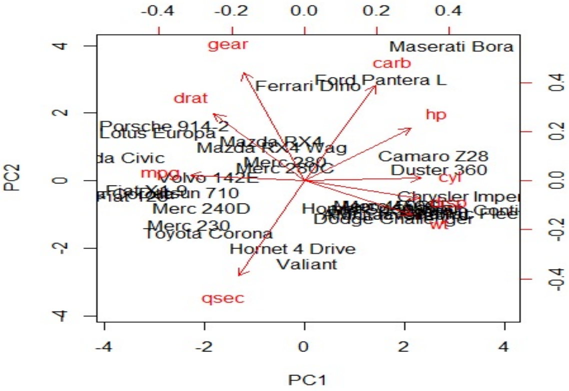

Figure 4: Biplot of mtcars Data

Worked Example: Interpreting Biplot

- The center of the car's name corresponds to the scores of PC1 and PC2 for the car. For example, the Maserati Bora has a score of 2.9 and 4.00 for PC1 and PC2 respectively, indicating a large engine and a powerful car.
- Scores of 0 correspond to the **average**.
- By contrast, the Valiant has a score of 0.00 and -2.50 for PC1 and PC2 respectively, indicating average size of engine and having small engine power.

Worked Example: mtcars

Below, we have the scores for the Valiant and the Maserati Bora

```
> rbind(round(pr.out$x[6,],3), round(pr.out$x[31,],3))
       PC1    PC2   PC3    PC4    PC5    PC6   PC7    PC8    PC9
[1,] 0.050 -2.447 0.112 -0.872 -0.126 -0.230 0.225  0.099 -0.004
[2,] 2.963  3.999 0.703 -0.730 -0.228  0.656 0.494 -0.082 -0.053
```

Worked Example: Interpreting Biplot

Notice the values on the axes opposite the horizontal and vertical axes. These give the loadings each feature has on PC1 and PC2 respectively.

- For example, the loading of Carb on PC1 and PC2 is about 0.2 and 0.5 respectively (actually 0.245 and 0.484).

1. Principal Component Analysis

2. Worked Example

3. Choosing Number of PCs

Proportion of Variance Explained

Note how each PC is found by finding the linear combination of features that have the largest variance, subject to

- the constraint $\sum_{j=1}^{k} \phi_{jm}^2 = 1$ for $m = 1, 2, \cdots, min(n-1, k)$
- every PC being orthogonal to each other.

To evaluate the strength of each PC, we can compute the **proportion of variance** in the variable space that is explained by each PC, called PVE.

Proportion of Variance Explained

The total variance in the variable space (assuming variables have been centered to have mean 0) is

$$\sum_{j=1}^{k} Var(X_j) = \sum_{j=1}^{k} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2. \tag{4}$$

The variance explained by the *m*th PC is

$$Var(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2 \tag{5}$$

Proportion of Variance Explained

Therefore, the PVE is just simply (5) divided by (4):

$$PVE(m) = \frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{k} \sum_{i=1}^{n} x_{ij}^2} \tag{6}$$

Typically, a scree plot is produced. A scree plot has the PVE on the vertical axis and the number of PCs on the horizontal axis.

Worked Example

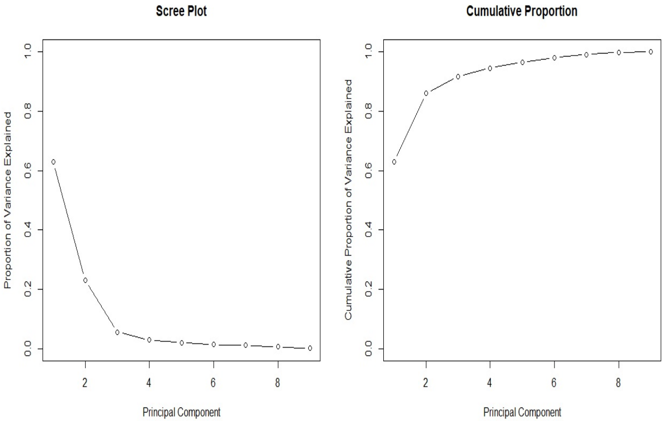

Figure 5: Scree Plot mtcars Data