

lab PCA

The topic that are covered in this lab worksheet is:

- Principal Component Analysis (PCA)
- Dimension Reduction

1. Principal Component Analysis (PCA)

Recall that one primary purpose for performing PCA is dimension reduction.

Today we will work with the USArrests dataset.

| | | | | |
|---|--|--|--|--|
| nrow(USArrests) | | | | |
| ## [1] 50 | | | | |
| ncol(USArrests) | | | | |
| ## [1] 4 | | | | |
| dim(USArrests) | | | | |
| ## [1] 50 4 | | | | |
| names(USArrests) | | | | |
| ## [1] "Murder" "Assault" "UrbanPop" "Rape" | | | | |

There are 50 observations (which correspond to the 50 states in US), and each observation is of dimension 4.

The built-in function prcomp() is used for performing PCA.

Recall that pre-processing of the dataset is needed. First, we do centering but no normalization.

center=TRUE ==> means do centering but no normalization.

| | | | | |
|---|--|--|--|--|
| USArr.noscale <- prcomp(USArrests, center=TRUE) | | | | |
| USArr.noscale | | | | |
| ## Standard deviation (1, ..., p=4): | | | | |
| ## [1] 83.7324408 14.212642791 0.489426 2.482790 | | | | |
| ## | | | | |
| ## Rotation (n x k) = (4 x 4): | | | | |
| ## PC1 PC2 PC3 PC4 | | | | |
| ## Murder 0.8417847 -0.084482166 0.07989066 0.9949217 | | | | |
| ## Assault 0.99522128 -0.05876003 -0.06756974 0.03893830 | | | | |
| ## UrbanPop 0.04633575 0.97685748 -0.20054629 -0.05851914 | | | | |
| ## Rape 0.07515550 0.20071807 0.97408059 0.07232502 | | | | |

The first principal component (PC1) is dominated by Assault with weight more than 0.995, while the other three are less than 0.08.

But this happens not because Assault is not correlated with the other two crimes and UrbanPop (which means the percentage of urban population), but because the values of Assault are significantly larger than the values of the other three attributes, so the result of PCA is over-determined by Assault.

Use the code snippet below to see the ranges of number of different attributes.

| | | | | |
|--------------------------------|--|--|--|--|
| range(USArrests[, "Murder"]) | | | | |
| ## [1] 0.8 17.4 | | | | |
| range(USArrests[, "Assault"]) | | | | |
| ## [1] 45 337 | | | | |
| range(USArrests[, "UrbanPop"]) | | | | |
| ## [1] 32 91 | | | | |
| range(USArrests[, "Rape"]) | | | | |
| ## [1] 7.3 46.0 | | | | |

In this case, normalization is necessary, as is done below.

scale=TRUE ==> means do both centering and normalization.

| | | | | |
|---|--|--|--|--|
| USArr.scale <- prcomp(USArrests, scale=TRUE) | | | | |
| USArr.scale | | | | |
| ## Standard deviations (1, ..., p=4): | | | | |
| ## [1] 1.5748783 0.9948694 0.5971291 0.4164494 | | | | |
| ## | | | | |
| ## Rotation (n x k) = (4 x 4): | | | | |
| ## PC1 PC2 PC3 PC4 | | | | |
| ## Murder -0.5358995 -0.4301869 -0.3412327 0.64922780 | | | | |
| ## Assault -0.5891836 0.1878985 -0.2851484 -0.74349748 | | | | |
| ## UrbanPop -0.2781909 -0.8728962 -0.3780158 0.13887773 | | | | |
| ## Rape -0.5434321 -0.1673186 0.8177779 0.08962432 | | | | |

After performing PCA with normalization, we look at the first PC. Assault is positively correlated with all the other two crimes, which is expected. Assault is also positively correlated with the percentage of urban population.

It is worth noting that the standard deviation of the first PC are not much larger than that of the second PC, so we also need to consider the second PC. This is quite common with real-world datasets, as the relationship between different attributes are often not clear-cut.

2. Dimension Reduction

Before going into dimension reduction, we briefly discuss how R handles the expressions like (a matrix + a vector) and (a matrix * a vector). Understanding this is crucial for the rest of this section.

This is best illuminated by examples.

| | | | | |
|--|--|--|--|--|
| A <- matrix(c(1,2,3,4,5,6), ncol=3, byrow = TRUE) | | | | |
| A | | | | |
| ## [1,] [2,] [3,] | | | | |
| ## [1,] 1 2 3 | | | | |
| ## [2,] 4 5 6 | | | | |
| A + c(7,11) | | | | |
| ## [1,] [2,] [3,] | | | | |
| ## [1,] 8 9 10 | | | | |
| ## [2,] 15 16 17 | | | | |
| A * c(7,11,3) | | | | |
| ## [1,] [2,] [3,] | | | | |
| ## [1,] 8 5 14 | | | | |
| ## [2,] 15 12 9 | | | | |
| A + c(7,11,3,13) | | | | |
| ## Warning in A + c(7, 11, 3, 13): longer object length is not a multiple of | | | | |
| ## shorter object length | | | | |
| ## [1,] [2,] [3,] | | | | |
| ## [1,] 8 5 10 | | | | |
| ## [2,] 15 18 17 | | | | |

To understand why the outputs are as above, you may first hypothetically think the matrix A is treated as a vector c(1,4,2,5,3,6) in R (go down the first column, then go down the second column, and so on).

In A+v for any vector v, what R has done is to add to each entry of the hypothetical vector of A by the entry of v in a cyclic manner.

Analogous outputs are obtained when you replace addition by multiplication (or subtraction or division). Experiment yourself.

Now we come back to PCA.

Recall that one primary purpose for performing PCA is dimension reduction. prcomp has already done this for you.

You can retrieve the result as below.

| | | | | |
|---|--|--|--|--|
| USArr.scale\$center | | | | |
| ## Murder Assault UrbanPop Rape | | | | |
| ## 7.788 170.760 65.540 21.232 | | | | |
| USArr.scale\$scale | | | | |
| ## Murder Assault UrbanPop Rape | | | | |
| ## 4.355510 83.337661 14.474763 9.366385 | | | | |
| USArr.scale\$e | | | | |
| ## PC1 PC2 PC3 PC4 | | | | |
| ## Alabama -0.07566045 1.12200121 -0.43980366 0.15409581 | | | | |
| ## Alaska -1.03051708 1.06242092 0.03506927 0.43417544 | | | | |
| ## Arizona -1.74544285 -0.73945954 0.05423025 -0.82626420 | | | | |
| ## Arkansas -0.13999894 1.10854226 0.11342217 -0.18097355 | | | | |
| ## California -2.49854328 -1.02742672 0.59254100 -0.33895920 | | | | |
| ## Colorado -1.49934074 -0.97762966 1.08406162 0.80145654 | | | | |
| ## Connecticut 1.34492236 -1.07798362 -0.63679250 -0.11728736 | | | | |
| ## Delaware -0.04722081 -0.32208890 -0.71141632 -0.873113315 | | | | |
| ## Florida -2.98275067 0.03883425 -0.57103206 -0.095317042 | | | | |
| ## Georgia -1.62089742 -1.26008038 0.33091818 1.065974459 | | | | |
| ## Hawaii 0.90348444 -1.55487609 0.05027151 0.093731198 | | | | |
| ## Idaho 1.62331903 0.20885253 0.25719021 -0.494087852 | | | | |
| ## Illinois -1.36585197 -0.67498834 -0.67068647 -0.120794916 | | | | |
| ## Indiana 0.50038122 -0.15003928 0.22576277 0.420397595 | | | | |
| ## Iowa 2.23095979 -0.10300628 0.16291036 0.017379470 | | | | |
| ## Kansas 0.78887206 -0.26744041 0.02529648 0.20421034 | | | | |
| ## Kentucky 0.74331256 0.94880748 -0.02808429 0.603817237 | | | | |
| ## Louisiana -1.54909076 0.06230011 -0.77560598 0.495177791 | | | | |
| ## Maine 2.37274014 0.37260805 0.86502225 0.327138529 | | | | |
| ## Maryland -1.74564663 0.42335704 -0.15566968 -0.55340589 | | | | |
| ## Massachusetts 0.48128007 -1.45967706 -0.60337172 -0.177739802 | | | | |
| ## Michigan -2.08725025 -0.15938500 0.30100604 0.101343128 | | | | |
| ## Minnesota 1.67560851 -0.62590670 0.15153200 0.065640316 | | | | |
| ## Mississippi -0.98547919 2.36973712 -0.73385290 0.212342040 | | | | |
| ## Missouri -0.68978426 -0.26070794 0.37365033 0.223554851 | | | | |
| ## Montana 1.17353751 0.53147851 0.24440796 0.122498555 | | | | |
| ## Nebraska 0.54503180 -1.40833518 -0.48712332 0.636731051 | | | | |
| ## Nevada -2.84506942 -0.76780902 0.80070797 -0.143433097 | | | | |
| ## New Hampshire 2.35995585 -0.01790055 0.03648498 -0.82084291 | | | | |
| ## New Jersey -0.17974182 -1.43493745 -0.75677041 0.240936580 | | | | |
| ## New Mexico -1.96012351 0.14141308 0.18184598 -0.33621113 | | | | |
| ## New York -2.05821199 -0.60512507 0.63013102 0.023408044 | | | | |
| ## North Carolina -1.11208808 2.20561081 -0.85489245 -0.954749648 | | | | |
| ## North Dakota 2.96215223 0.59309738 0.29024930 -0.251434628 | | | | |
| ## Ohio 0.22369436 -0.73477837 -0.03082616 0.469152817 | | | | |
| ## Oklahoma 0.30804928 -0.20405113 0.01559212 0.327138529 | | | | |
| ## Oregon -0.05852787 -0.53596999 0.03038718 -0.235390872 | | | | |
| ## Pennsylvania 0.07840680 -0.56530650 -0.39606218 0.355452378 | | | | |
| ## Rhode Island 0.85509072 -1.47689328 -1.35617705 -0.607402746 | | | | |
| ## South Carolina -1.30744088 -1.91937397 -0.29751723 0.130145786 | | | | |
| ## South Dakota 1.96770669 0.83560522 0.30530073 -0.100470512 | | | | |
| ## Tennessee -0.98969737 0.85160534 0.18619262 0.640302674 | | | | |
| ## Texas -1.34151838 -0.40833518 -0.48712332 0.636731051 | | | | |
| ## Utah 0.54503180 -1.40833518 -0.48712332 0.636731051 | | | | |
| ## Vermont -2.77255123 1.38019435 0.86070797 -0.143433097 | | | | |
| ## Virginia 0.09536670 0.19772785 0.01559482 0.209246429 | | | | |
| ## Washington 0.21423739 -0.96037394 0.61859067 -0.218628161 | | | | |
| ## West Virginia 2.08468199 -0.60512507 0.10372163 0.13740523 | | | | |
| ## Wisconsin 2.05821199 -0.60512507 0.10372163 0.13740523 | | | | |
| ## Wyoming 0.62310061 0.31778662 -0.23824049 -0.164078666 | | | | |

To interpret the above tables, recall that in some code snippet before, prcomp has computed the first, second, third and fourth PCs for us.

Let's denote the PCs by u1, u2, u3, u4. Then the original observation of Alabama can be recovered as follows. First, compute $-0.97566045u_1 + 1.12200121u_2 - 0.43980366u_3 + 0.15409581u_4$ as below.

| | | | | |
|---|--|--|--|--|
| t(USArr.scale\$rotation %*% USArr.scale\$e["Alabama",]) | | | | |
| ## Murder Assault UrbanPop Rape | | | | |
| ## [1,] 1.242564 0.762093 -0.520906 -0.003416473 | | | | |

Recall that we had done normalization. Thus, to recover the original observation, we need to scale back, as below.

| | | | | |
|--|--|--|--|--|
| q <- t(USArr.scale\$rotation %*% USArr.scale\$e["Alabama",]) | | | | |
| q <- USArr.scale\$scale * q | | | | |
| q | | | | |
| ## Murder Assault UrbanPop Rape | | | | |
| ## [1,] 5.412 65.24 -7.54 -0.032 | | | | |

Recall that we had done centering. Thus, the final step to recover the original observation is to add the center, as below.

| | | | | |
|---------------------------------|--|--|--|--|
| q <- q + USArr.scale\$center | | | | |
| q | | | | |
| ## Murder Assault UrbanPop Rape | | | | |
| ## [1,] 13.2 236 58 21.2 | | | | |

| | | | | |
|---------------------------------|--|--|--|--|
| USArrests["Alabama",] | | | | |
| ## Murder Assault UrbanPop Rape | | | | |
| ## Alabama 13.2 236 58 21.2 | | | | |

What if we want to represent all the observations approximately by just using the first two PCs?

We can use the following short code snippet. This is the reason why we discuss (a matrix + a vector) and (a matrix * a vector) in the beginning of this section.

</