question

Monday, May 17, 2021 4:09 PM

考试tasks

w1 turt example///

```
> outlier=identify(pca.turt$scores)
warning: no point within 0.25 inches
```

why I cannot run outlier test use identify copied from the material?
it says: warning: no point within 0.25 inches

1.why I cannot run an outlier test use identify copied from the material?

I am not exactly sure what might cause the problem unless seeing more codes. Some possibilities are: 1) you click somewhere which is too far away from the point; 2) the arguments in the plots are not same as those in the identify function. Please follow the code strictly and see if it works. Check the post below.

https://www.reddit.com/r/RStudio/comments/7xso1l/getting_warning_when_trying_to_identify_a_point/

why the components are in *decreasing* order of eigenvalue? cannot understand

The first PC is found to maximise the variance. Following on the supplementary material on Page 6,

$$\mathrm{Var}(Y_1) = \sum_{i=1}^{r} \lambda_i b_i^2 \leq \lambda_1 \sum_{i=1}^{r} b_i^2$$

The maximum value equals to λ_1, which is obtained when the first element of the vector b is 1 and the remaining elements are 0.

For the second PC, again we have the expression

$$\mathrm{Var}(Y_1) = \sum_{i=1}^{r} \lambda_i b_i^2$$

Meanwhile, there is an additional constraint that $a_1^T a_1 = 0$. Under this constraint, the first element of the vector b is zero (see "A" below for a proper proof). So

$$\mathrm{Var}(Y_1) = \sum_{i=1}^{r} \lambda_i b_i^2 = \sum_{i=2}^{r} \lambda_i b_i^2 \leq \lambda_2 \sum_{i=1}^{r} b_i^2$$

The maximum value equals λ_2, obtained when the second element of the vector b is 1 and the remaining elements are 0.

Proof of statement (A): Since $b = U^T a$ and U is an orthogonal matrix, we have $a = Ub = b_1 e_1 + b_2 e_2 + \ldots + b_p e_p$. Under the constraint $a_1^T a_1 = 0$, (temporarily drop the subindex 2 in a_1 for clarity)

$$0 = a^T a = e_1^T (b_1 e_1 + b_2 e_2 + \ldots + b_p e_p) = b_1$$

Hope this is clear. If not, try to read P101 (maximization of quadratic forms for points on the unit sphere) and Result 8.1 from the book Applied Multivariate Statistical Analysis, or speak to me in the next tutorial.

////////

The question aboue Example2

4 days ago

1.It's said that "Because we used the correlation matrix, we know the average should be 1 but let's check"

Could you please give more explanation about it?

1. The principal components obtained from the correlation matrix, i.e. eigenvectors of the correlation matrix of the original variables, are same as the principal components for standardised variables, i.e. eigenvectors of the covariance matrix of standardised variables.

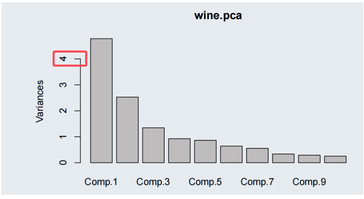
This will be discussed in the tutorial (conceptual question 3).

2. sum of eigenvalues from decomposing the correlation matrix = total variance of standardised variables = number of variables * variance of each variable (equals one for standardised variable).

As the number of eigenvalues is same as the number of variables, the average of eigevanlue equals one.

/////////

It's said that the pic is about "the proportion of variance versus component number", and we can see that Comp.1's Proportion of Variance is 0.3676196, why the value on the y-axis of Comp.1 is over 4?



The y-axis in this plot is the variance; their values can be seen by using `wine.pca$sdev^2`.

As you said, the scree plot is about plotting the proportion of variance versus the component number. As the total variance is fixed, the decision based on the variance is the same as the decision based on the proportion of variance. You can check it by using the command below:
`plot(wine.pca$sdev^2 / sum(wine.pca$sdev^2))`

///

regarding outlier test

3 days ago

In the notes, we used identify function to remove the outliers manually. Is there any default R program to perform the outlier test?

As mentioned in the tutorial, there are a few more quantitative ways and advanced methods to detect outliers. You may find R codes once you decided the method.
1. A relatively simple rule-of-thumb method
https://en.wikipedia.org/wiki/68–95–99.7_rule
2. Outlier detection based on statistical tests
https://statsandr.com/blog/outliers-detection-in-r/
3. Outlier detection based on Gaussian mixture models
https://rpubs.com/JayAhn/650433

//////

Interpretation of loadings

We look at the loadings column-wise. For example, "wine.pca$loadings[,1]" gives the first loading vector, a_1, which can be interpreted as how much each variable contributes to the first loading. For example, by looking at their absolute values, we could say Flavanoids, Total.phenols and OD280.OD315 are more important to the first PC than Ash and Colour.intensity. Looking at the signs, we could say the first PC is a *difference* between the *average* of Malic.acid, Alcalinity.of.ash, Nonfl.phenols (group A), and the *average* of Alcohol, Magnesium, Total.phenols, Flavanoids, Proanthocynanins, Hue, OD280.OD315, and Proline (group B). Our comment on average is made as variables in group A have relatively similar magnitude, similarly for variables in group B. The comment on difference is made as variables in group A have different signs with those in group B.

Prediction

PCA automatically centre the data. So when making a prediction, we need to first subtract the sample mean from the new observation and then multiply it by the first loading vector. The code is given below.

```
new.data <- c(4.8,4.7,3.9)
(new.data-pca.turt.new$center) %*% pca.turt.new$loadings[,1]
```

W1

```
> knit_print(my_skim(wine))
[1] "\nTable: Data summary\n\n|                         |     |\n|:------------------------|:----|\n|Nam
e                      |wine |\n|Number of rows           |178  |\n|Number of columns        |14   |\n|_
_______________________ |     |\n|Column type frequency:   |     |\n|numeric                  |14   |\n|_
_______________________ |     |\n|Group variables          |None |\n\n\n**Variable type: numeric**\n\n|s
kim_variable           | n|   mean|    sd|   p25|    p50|    p75|\n|:-----------------------
--|--:|-----:|-----:|-----:|------:|-----:|\n|Class                    | 178|  1.94|  0.78|
 1.00|  2.00|  3.00|\n|Alcohol                  | 178|  13.00|  0.81|  12.36|  13.05|  13.68|\n|M
```

wine example, the output of knit_print is not readable, how should I fix?
Answer: The code is implemented using Rmarkdown. If you are using R script, try use 'skim(employ)' (That is, delete knit_print and no need to use my_skim)

Answer to Task 1.
Cannot understand the following sentence,
The points lying close to a diagonal line indicate strong correlation between the original variables and principal components have no correlation.

W2///

Metric or nonmetric MDS? Apply nonmetric scaling when the actual values of dissimilarities are not reliable, but their orders can be trusted.

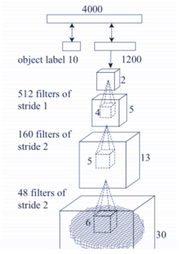
page 11, what is the order of dissimilarity? by its absolute value?
why want to change similarity into dissimilarity?

W6

Theorem 3.1. *The energy, E of a convolutional layer can be computed as*:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_f \sum_j \left(h_j^f \left(W^f * v \right)_j + c^f h_j^f \right) - \sum_l b_l v_l$$

where v_l denotes each visible unit, h_j^f denotes each hidden unit in a feature channel f, and W^f denotes the convolutional filter.



Hi, Could you go through tutorial 10 q6, tut9 concept. q1?