

Re: Principal Component Analysis– Do I need to scale (i.e. normalize) my variables?

## Re: Principal Component Analysis– Do I need to scale (i.e. normalize) my variables?

---

*Source:* <http://sci.tech–archive.net/Archive/sci.stat.math/2009–05/msg00153.html>

---

- *From:* Ray Koopman <[koopman@xxxxxx](mailto:koopman@xxxxxx)>
  - *Date:* Sun, 17 May 2009 01:43:40 –0700 (PDT)
- 

On May 16, 1:07 pm, Kerry <[kbro...@xxxxxxxxxx](mailto:kbro...@xxxxxxxxxx)> wrote:

On May 16, 9:46 am, Art Kendall <[Arthur.Kend...@xxxxxxxxxxxx](mailto:Arthur.Kend...@xxxxxxxxxxxx)> wrote:

When you tell the software to use the covariance matrix in any type of factor analysis (principal components, principal factors, alpha, image, etc) the scales are important. In the more routine situation, where you tell the software to use the correlation matrix the variables are implicitly standardized.  $z = (\text{value of } x - \text{mean of } x) / (\text{standard deviation of } x)$ .

I don't know what the substantive meaning of x's would be if you used the absolute value of the z–scores aka standardized variables. This is a correlation matrix on 4 z–scores. (it is the same as for raw variables).

```
x1 x2 x3 x4
x1 1 .646 –.303 –.533
x2 .646 1 –.437 –.399
x3 –.303 –.437 1 .327
x4 –.533 –.399 .327 1
```

```
This is a correlation on the absolute value of the z–scores
abs1 abs2 abs3 abs4
abs1 1 .535 –.269 .125
abs2 .535 1 –.324 .117
abs3 –.269 –.324 1 .022
abs4 .125 .117 .022 1
```

Art Kendall  
Social Research Consultants

Kerry wrote:

Hi,

I need to perform PCA on 20 or so variables (ex. height of say a tree,

Re: Principal Component Analysis– Do I need to scale (i.e. normalize) my variables?

weight of a tree, age, etc), many with different units and/or value ranges. Will this bias my results? I noticed in a past PCA I did that I converted all of my values to z scores [i.e.  $((\text{abs}(\text{value}-\text{mean}))/\text{std dev})$ ], but not sure why or if this was even a correct way to normalize. If I do need to normalize my values, wouldn't it make more sense to convert them to value/mean? Or what about value/sum(values)?

To be clear, I am referring to making my values unitless prior to adding them all to my PCA.

Thanks,  
K

Thank you both Ray and Art, this is very clear. I'm using SPSS so I'd assume (but will still double-check) that it standardizes the values. I also now get why I was scaling using variance. Thanks for finding my abs value mistake too, that was a typo here that I fortunately was not putting in my analysis.

I am a bit confused about the log transformation suggested by Ray. You mention that standardization undoes any previous linear but not non-linear transformations. But if I allow SPSS to do its inherent standardization, then I won't need to be doing any standardization beforehand right? I'll just be putting my raw values for all my variables directly into the PCA. Should I still consider log transformation of certain variables prior to PCA implementation (i.e. SPSS standardization)? I have some variables that are integers, some that are real, natural, and some ratios. The distribution shape also varies greatly from variable to variable, and I'll be concerned if this is a potential confound that I need to correct for prior to PCA analysis.

Also, I'm not sure what the difference is in meaning for using correlation vs covariance, but I'm going to be spending some time looking into the additional PCA options SPSS provides so hopefully I'll know soon.

My purpose of the analysis is to explore which variables best define object shape which I'll then by throwing into a cluster analysis and possibly Kolmogorov–Smirnov testing to look for multimodality.

Thanks again,  
K

## Re: Principal Component Analysis– Do I need to scale (i.e. normalize) my variables?

Log transformations should be done to all variables that have a real natural zero–point, and intervals whose interpretive magnitude is inversely proportional to the scale values. The second part of that is most easily understood by way of an example: a one–foot difference ("interval") in the height of two trees is substantial when the trees are 2 and 3 feet tall, but negligible when the trees are 20 and 21 feet tall, so height should be logged. The larger the ratio of the maximum/minimum value, the more important it is that the variable be logged. This applies to all the analyses, not just the PCA. Failure to log means that the cases (trees, whatever) with larger values will dominate the analysis, essentially ignoring those with smaller values.

Logging changes multiplicative relations to additive relations. The log of a ratio equals the difference between the logs of the numerator and denominator. If you have variables that are ratios then you should consider including the logs of the numerator and denominator as separate variables, and omitting the log of the ratio.

Logging is done before standardizing, not after.

You should probably be doing the PCA on the correlation matrix, not the covariance matrix. That is, you should let the program do its default analysis. Usually, a PCA is done on the covariance matrix only when all the variables are in the same units, and any differences in variance are due to causes whose effects you want the analysis to be sensitive to.

.