Cross Validated

Can I use the CLR (centered log-ratio transformation) to prepare data for PCA?

Asked 2 years, 2 months ago Active 1 month ago Viewed 6k times

13

4

I am using a script. It is for core records. I have a dataframe which shows the different elemental compositions in the columns over a given depth (in the first column). I want to perform a PCA with it and I am confused about the standardization method I have to choose.

Has anyone of you used the `clr()` to prepare your data for the `prcomp()` ? Or does it adulterate my solutions. I have tried using the `clr()` on the data before using the `prcomp()` function in addition to using the attribute scale in `prcomp()` .

```
data_f_clr<- clr(data_f)
data_pca <- prcomp(data_f, center = TRUE, scale. = TRUE)
```

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html

scale is described to scale the data, so they have unit variance. Since my data have a very different scale that is what i wanted, I think. The problem is, that I receive a different solution, when I use the code above or when I skip the `clr()` (which makes the more wanted result). But I want to know why is the `clr()` disturbing in that case?

r pca normalization compositional-data

edited Oct 5 '17 at 12:35
amoeba says Reinstate Monica
75.3k 19 233 281

asked Oct 2 '17 at 18:30
T.rex
131 1 3

2 For non-R users like me, it might be helpful to clarify what `clr` does.... – Dougal Oct 2 '17 at 18:57

3 Of course the CLR changes the solutions--why else would you use this procedure? Perhaps you should be asking how to determine which approach is better. There are useful posts to be found by searching our site for CLR. In an answer to a related question I provided some illustrations that might help you. – whuber ♦ Oct 2 '17 at 18:57

1 The quick answer is that you can do anything you want with data prior to PCA. There are no edicts, laws or recipes governing this. Some contend that PCA (without rotation) is scale invariant while others contend that the results of a PCA are highly sensitive to scale. But if you rotate the results of PCA then the rules of thumb mandate pre-PCA normalizing such as *CLR* or standardizing to mean=0 and SD=1. A great discussion of *CLR* is in Lee Cooper's book *Market Share Analysis* (anderson.ucla.edu/faculty/lee.cooper/MCI_Book/BOOKI2010.pdf) linking it to component analysis. – Mike Hunter Mar 1 '18 at 13:23

2 @DJohnson I searched the pdf linked for various words in CLR and centered log-ratio transformation but could find nothing. What did I do wrong? There is no index in that version, but the section headings don't

1 As already mentioned, there is no index in the version you've linked to, so forgive me for not being to consult it. Thanks for the keyword "log-centering" from which I find discussions of a different beast, not the **centered log-ratio transformation**, which this thread is all about. @whuber already gave a link to a discussion on this site. The key is that for compositional data with proportions adding to 1, there is need and scope for collective transformation to a different space. You missed the word "ratio" as pointing to a different idea from the one you know. – Nick Cox Mar 2 '18 at 12:47

2 Answers

5

Yes you can, and in fact you should, when your data is compositional.

A review from the field of microbiology can be found here, which motivates to use the CLR-transformation followed by PCA to analyze microbiome datasets (which are per definition compositional): https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full.

answered Mar 1 '18 at 11:24
Archie
470 6 15

Quite unfortunately, that paper is terribly wrong in many cases, which is a pity, considering that two coauthors are champions of compositional data analysis. – Eli Korvigo Jun 18 '18 at 20:42

@EliKorvigo That comment may be well-founded but by itself it is not helpful. If you could point to a published or at least public critique then such a critique would change the picture. – Nick Cox Jun 19 '18 at 18:32

@NickCox sure, there is a paper by Filzmoser and Hron. It's not a direct critique of the aforementioned paper, but it argues against using CLR for correlation analysis, while the aforementioned paper recommends tools based on CLR. – Eli Korvigo Jun 19 '18 at 18:51

@NickCox I'd like to stress my deep respect for Dr. Pawlowsky-Glahn and Dr. Egozcue, who are the last two authors of the paper mentioned by Archie. In fact, they've introduced ILR to address CLR's shortcomings (Egozcue and Pawlowsky-Glahn, 2003). Referring to CLR they write: *"Nevertheless, orthogonal references in that subspace are not obtained in a straightforward manner"*. – Eli Korvigo Jun 19 '18 at 19:33

Pawlowsky-Glahn and Egozcue state in "Compositional data and their analysis: an introduction" (2006) that clr coefficients "have certain advantages: the expression is symmetric in the parts and these coordinates reduce the computation of Aitchison distances to ordinary distances. They are useful in the computation of bi-plots (...)" – jO. Oct 14 at 19:47

5

You might experience some issues with vanilla PCA on CLR coordinates. There are two major problems with compositional data:

- they are strictly non-negative
- they have a sum constraint

Various compositional transforms address one or both of these issues. In particular, CLR transforms your data by taking the log of the ratio between observed frequencies $\mathbf{x}$ and their geometric mean $G(\mathbf{x})$, i.e.

Now, consider that

$$\log(G(\mathbf{x})) = \log\left(\exp\left[\frac{1}{n} \sum_{i=1}^{n} \log(x_i) \right] \right) = \mathbb{E}\left[\log(\mathbf{x})\right]$$

This effectively means that

$$\sum \hat{\mathbf{x}} = \sum \left[\log(\mathbf{x}) - \mathbb{E}\left[\log(\mathbf{x})\right]\right] = 0$$

In other words CLR removes the value-range restriction (which is good for some applications), but does not remove the sum constraint, resulting in a singular covariance matrix, which effectively breaks (M)ANOVA/linear regression/... and makes PCA sensitive to outliers (because robust covariance estimation requires a full-rank matrix). As far as I know, of all compositional transforms only ILR addresses both issues without any major underlying assumptions. The situation is a bit more complicated, though. SVD of CLR coordinates gives you an orthogonal basis in the ILR space (ILR coordinates span a hyperplane in CLR), so your variance estimations will not differ between ILR and CLR (that is of course obvious, because both ILR and CLR are isometries on the simplex). There are, however, methods for robust covariance estimation on ILR coordinates [2].

Update I

Just to illustrate that CLR is not valid for correlation and location-dependant methods. Let's assume we sample a community of three linearly independent normally distributed components 100 times. For the sake of simplicity, let all components have equal expectations (100) and variances (100):

```
In [1]: import numpy as np

In [2]: from scipy.stats import linregress

In [3]: from scipy.stats.mstats import gmean

In [4]: def clr(x):
   ...:     return np.log(x) - np.log(gmean(x))
   ...:

In [5]: nsamples = 100

In [6]: samples = np.random.multivariate_normal(
   ...:     mean=[100]*3, cov=np.eye(3)*100, size=nsamples
   ...: ).T

In [7]: transformed = clr(samples)

In [8]: np.corrcoef(transformed)
Out[8]:
array([[ 1.        , -0.59365113, -0.49087714],
       [-0.59365113,  1.        , -0.40968767],
       [-0.49087714, -0.40968767,  1.        ]])

In [9]: linregress(transformed[0], transformed[1])
Out[9]: LinregressResult(
   ...:     slope=-0.5670, intercept=-0.0027, rvalue=-0.5936,
   ...:     pvalue=7.5398e-11, stderr=0.0776
   ...: )
```

Considering the responses I've received, I find it necessary to point out that at no point in my answer I've said that PCA doesn't work on CLR-transformed data. I've stated that CLR can break PCA in **subtle** ways, which might not be important for dimensionality reduction, but is important for exploratory data analysis. The paper cited by @Archie covers microbial ecology. In that field of computational biology PCA or PCoA on various distance matrices are used to explore sources of variation in the data. My answer should only be considered in this context. Moreover, this is highlighted in the paper itself:

> ... The compositional biplot *[note: referring to PCA]* has several advantages over the principal co-ordinate (PCoA) plots for β-diversity analysis. The results obtained are very stable when the data are subset (Bian et al., 2017), meaning that **exploratory analysis** is not driven simply by the presence absence relationships in the data nor by excessive sparsity (Wong et al., 2016; Morton et al., 2017).

Gloor et al., 2017

Update III

Additional references to published research (I thank @Nick Cox for the recommendation to add more references):

1. Arguments against using CLR for PCA
2. Arguments against using CLR for correlation-based methods
3. Introduction to ILR

edited Oct 15 at 19:51

answered Jun 18 '18 at 21:58

Eli Korvigo

511 1 5 13

2 A singular covariance matrix isnt a problem for pca! – kjetil b halvorsen Jun 18 '18 at 22:37

@kjetilbhalvorsen indeed, PCA per se doesn't not require the matrix to be full rank. Technically speaking, a singular covariance matrix will only result in one or more zero eigenvalues. Yet, people usually apply PCA to explore sources of variance, which is where compositionality kicks in. That's why I've been rather careful with my wording: *"... effectively breaks PCA/... in many* ***subtle*** *ways"* – Eli Korvigo Jun 18 '18 at 22:47

So you mean that due to the singularity one cannot calculate the amount of variance that is explained per component? Other then that, one can still perform PCA to perform dimensionality reduction. How does this then affect ANOVA/linear regression? – Archie Jun 19 '18 at 8:42

1 +1 because the answer is very interesting. It does not go without criticism, though. You seemingly (for me stupid) didn' explain precisely why doing PCA on compositional or clr-transformed data is improper "in subtle ways" (which? how?). Also, you are giving a python code but not its results. Can you display and comment its results? Finally, could you leave a link about ILR transfotm, to read about? – ttnphns Jun 20 '18 at 8:40

1 @ttnphns 1) as I've written in the comments, CLR doesn't not remove the distortion of variance-sources introduced by compositional closure, affecting exploratory data analysis: robust covariance estimation requires a full-rank matrix; 2) I'm not sure I follow, why you say there are no results: that's an interactive Python session with inputs and outputs (i.e. results); 3) I've added a reference for ILR. – Eli Korvigo Jun 29 '18 at 18:22