What does a non positive definite covariance matrix tell me about my data?

Asked 11 years, 3 months ago Modified 11 months ago Viewed 72k times

31

I have a number of multivariate observations and would like to evaluate the probability density across all variables. It is assumed that the data is normally distributed. At low numbers of variables everything works as I would expect, but moving to greater numbers results in the covariance matrix becoming non positive definite.

I have reduced the problem in Matlab to:

```
load raw_data.mat; % matrix number-of-values x number of variables
Sigma = cov(data);
[R,err] = cholcov(Sigma, 0); % Test for pos-def done in mvnpdf.
```

If err>0 then Sigma is not positive definite.

Is there anything that I can do in order to evaluate my experimental data at higher dimensions? Does it tell me anything useful about my data?

I'm somewhat of a beginner in this area so apologies if I've missed out something obvious.

normal-distribution multivariate-analysis covariance

Share Cite Improve this question Follow

asked Jun 14, 2012 at 16:53

ralight

421 1 4 5

It sounds like your data are too sparse for the high-dimension representations. Are you planning on running regression models with this data? – Jonathan Thiele Jun 14, 2012 at 16:58

I found it highly useful to understand how euler rotations apply to this question: robotics.stackexchange.com/questions/2556/... – D A Nov 5, 2021 at 20:35

2 Answers

Sorted by: Highest score (default)

43

The covariance matrix is not positive definite because it is singular. That means that at least one of your variables can be expressed as a linear combination of the others. You do not need all the variables as the value of at least one can be determined from a subset of the others. I would suggest adding variables sequentially and checking the covariance matrix at each step. If a new variable creates a singularity drop it and go on the the next one. Eventually you should have a subset of variables with a postive definite covariance matrix.

Share Cite Improve this answer Follow

answered Jun 14, 2012 at 17:04

Michael R. Chernick

41.7k 28 80 151

28 +1. It's also worth noting that all covariance matrices are positive definite and *all* positive definite matrices are the covariance matrix of some multivariate distribution. Therefore, saying "non-positive definite covariance matrix" is a bit of an oxymoron. It appears the OP was really just saying that the *sample* covariance matrix was singular which can happen from exactly collinearity (as you've said) or when the **number of observations is less than the number of variables**. – Macro Jun 14, 2012 at 17:23

5 Some stats software can be induced to correct this problem automatically. E.g., Stata's `regress` command will automatically drop extra variables when some are collinear (and its output can be saved in a form that identifies these variables and marks a non-collinear subset for future use). A likely complication, though, is that the variables might not necessarily be collinear, but they may be close enough that propagation of floating point error in the Cholesky decomposition produces negative eigenvalue estimates, making the variables collinear for all practical purposes. – whuber ♦ Jun 14, 2012 at 17:55

1 @whuber, there is similar functionality in `R` as well - regression models automatically drop variables from the linear predictor if there is exact collinearity. – Macro Jun 14, 2012 at 18:35

2 @whuber, it's a bit hacky but you can do a similar trick. If `g` is your linear model, then `colnames(model.matrix(g))[-which(is.na(coef(g))==TRUE)][-1]` should return the

names of the predictors entered into the model that are not exactly collinear. This works by checking which coefficients were `NA` (that's how `R` indicates a variable was dropped), and finding the corresponding column names of the model matrix (deleting the intercept column). By the way, that won't work if there are no collinear terms so an if statement to check that `which(is.na(coef(g))==TRUE)` isn't empty is required :) – Macro Jun 14, 2012 at 18:59

10 @Macro All covariance matrices are positive semi-definite. When they are singular they are not positive definite because x'Ax>0 for all vectors x for the matrix A to be positive definite. In the singular case x'Ax=0 occurs for some x. – Michael R. Chernick Jun 14, 2012 at 19:22

2

One point that I don't think is addressed above is that it IS possible to calculate a non-positive definite covariance matrix from empirical data even if your variables are not perfectly linearly related. If you don't have sufficient data (particularly if you are trying to construct a high-dimensional covariance matrix from a bunch of pairwise comparisons) or if your data don't follow a multivariate normal distribution, then you can end up with paradoxical relationships among variables, such as cov(A,B)>0; cov(A,C)>0; cov(B,C)<0.

In such a case, one cannot fit a multivariate normal PDF, as there is no multivariate normal distribution that meets these criteria - cov(A,B)>0 and cov(A,C)>0 necessarily implies that cov(B,C)>0.

All this is to say, a non-positive definite matrix does not always mean that you are including collinear variables. It could also suggest that you are trying to model a relationship which is impossible given the parametric structure that you have chosen.

Share Cite Improve this answer Follow

answered May 5, 2016 at 16:13

Adam Clark
39 1

2 Your answer is wrong on so many levels. Anyhow, consider a covariance matrix with 1's on the diagonal, and 1/2 for cov(1st and 2nd components), 1/2 for cov(1st and 3rd components), and -1/2 for cov(2nd and 3d components). The covariance matrix has eigenvalues approximately 0.15, 1.35, 1.50, providing a counterexample to the assertion in your 2nd paragraph. – Mark L. Stone May 5, 2016 at 16:31

1 @MarkL.Stone, you're right about the 2nd paragraph, but I wonder if some of this is ambiguous & could be saved under a generous interpretation. Eg, I wonder if, in paragraph 1, "don't have sufficient data... trying to construct a high-dimensional covariance matrix from a bunch of pairwise comparisons" refers to having a lot of missing data & using the pairwise complete observations to compute each element in the covariance matrix. – gung - Reinstate Monica May 5, 2016 at 17:04