

## Re: PCA – separation and variance

---

*Source:* <http://sci.tech-archive.net/Archive/sci.stat.math/2006-07/msg00085.html>

---

- *From:* Antti Penttilä <[Antti.I.Penttila@xxxxxxxxxxxxxxxxxxxxxx](mailto:Antti.I.Penttila@xxxxxxxxxxxxxxxxxxxxxx)>
  - *Date:* Thu, 06 Jul 2006 14:02:13 +0300
- 

shay wrote:

I have a multi-dimensional data set containing two different groups of data I want to separate utilizing PCA. I observed (by eye) that a certain set of variables gives me a good separation using the first three PCs.

What is the correlation, if it at all exists, between the separation in 3D to the actual separation in the higher dimensioned plane? How does the variance of the solution come into play? Can I say that the separation on the higher dimensioned plane is as good as or perhaps better/worse than the one in the 3D plane?

The more PCs you have, the more information you have. Therefore, in principle, it should always be possible to construct a better separation with more PCs. However in practice this might not be the case.

You could check the amount of variance that is loaded to the first PCs. That can be found comparing the squared eigenvalues of PCs against the sum of all squared eigenvalues. If, for example, the squared eigenvalues of the first three PCs sum up to more than, say 95% of all the squared eigenvalues you can be pretty sure that adding more PCs only adds noise.

Another suggestion is to do cross-validation. More PCs could produce better separation with the data in hand, but then the separation could be overestimated and would not be as good for future observations. Leave a part of the data out before PCA and try to separate them.

Furthermore, if you know the group labels (at least for some training set), then the linear discriminant analysis (LDA) could be even better choice. It is quite similar to PCA but intended to separate groups.

Regards,  
Antti

.