

Some comments on assignment #2 (hmw F1).

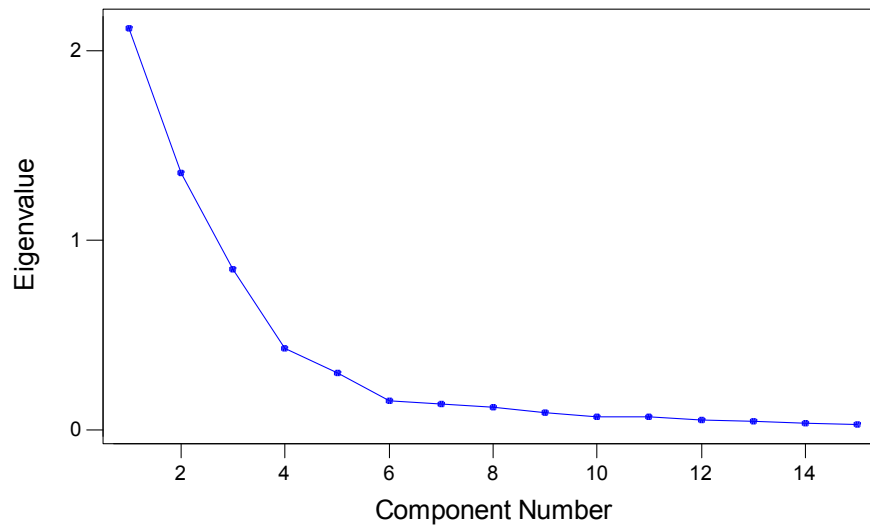
Here, just as an example to bring up some issues:

- 15 time points (but there are good reasons to drop some)
- 678 profiles without missing values (but you did a nice job with trying to rescue information with ad-hoc imputation)
- no row centering nor standardization (but there are good reasons to do it; making sure that profiles are considered only in terms of their shape, and that their overall location and/or amplitude does not affect the analysis)
- PCA (but it's fine to use FA, MLE extraction might have normality trouble...)
- Covariance matrix (keeping within what we consider as “structural” the differences in variability across time points that DO occur, but an argument can be made against this, which gives more “weight” to time points/original variables with larger variability)

Eigenanalysis of the Covariance Matrix

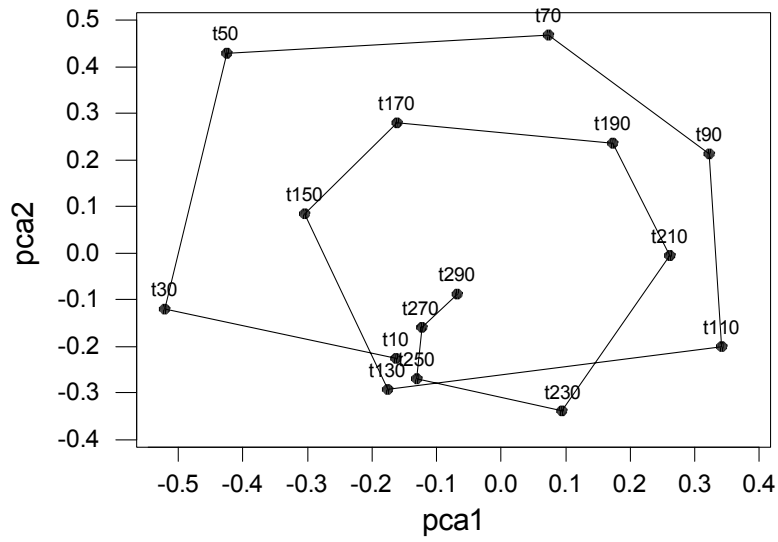
Eigenvalue	2.1206	1.3541	0.8472	0.4311	0.2984	0.1536
Proportion	0.364	0.233	0.146	0.074	0.051	0.026
Cumulative	0.364	0.597	0.743	0.817	0.868	0.895
Eigenvalue	0.1318	0.1154	0.0866	0.0661	0.0637	0.0484
Proportion	0.023	0.020	0.015	0.011	0.011	0.008
Cumulative	0.917	0.937	0.952	0.963	0.974	0.983
Eigenvalue	0.0444	0.0303	0.0270			
Proportion	0.008	0.005	0.005			
Cumulative	0.990	0.995	1.000			

Scree Plot of cdc15-10-cdc15-29

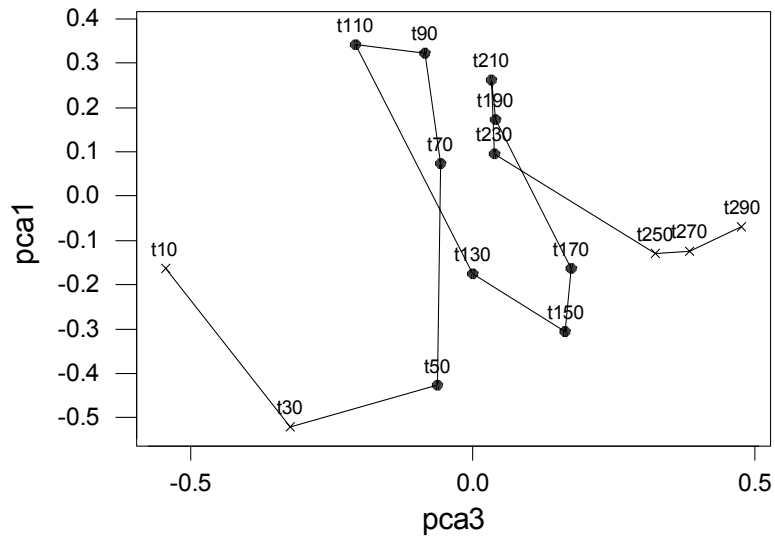


Would consider **at least three...**

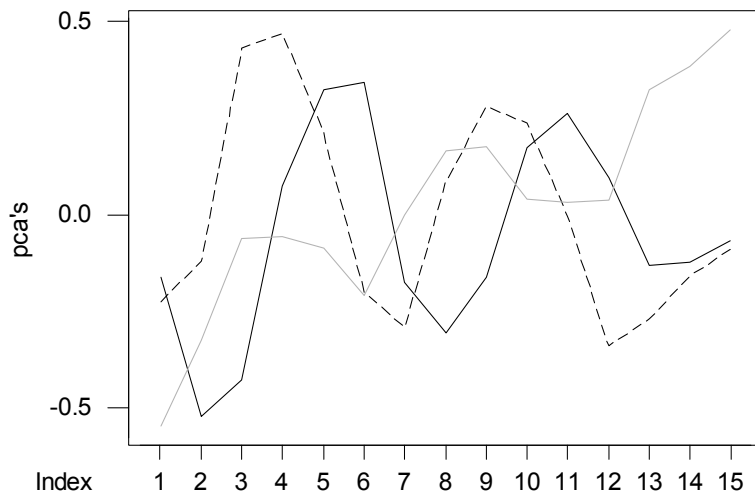
Fist, let us analyze the *relationships between original variables and principal components, and try to interpret the latter.*



- I: Oscillating behavior in pca1 and 2
- II: Upward trend with oscillations in pca3



x = drug? loss of synchronization?



solid=pca1, dash=pca2, gray=pca3

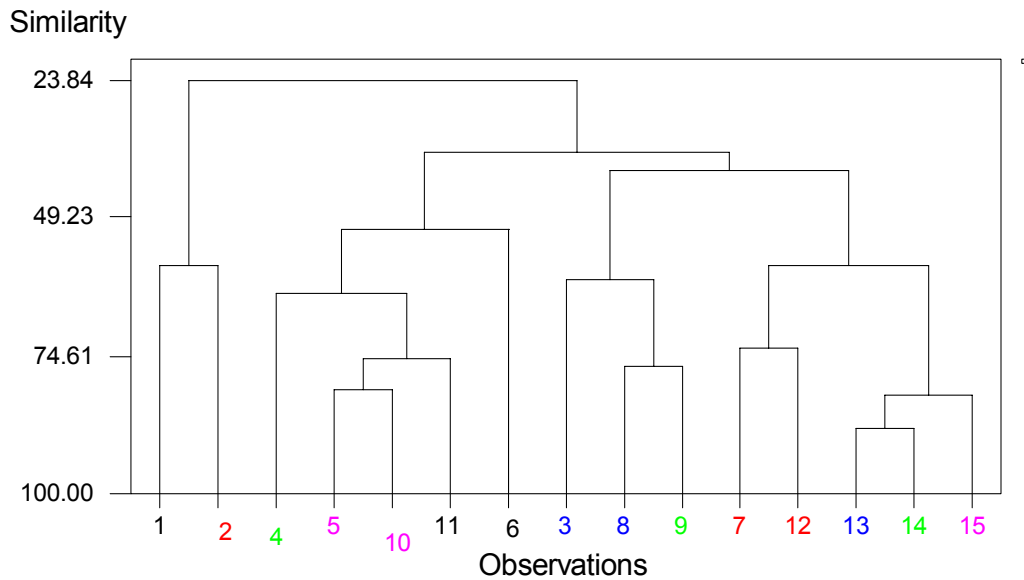
First and second principal component capture “oscillating behavior”, third captures an oscillating upward trend, and accounts for 15% of the variability; certainly not a negligible one. Some of you made the comment that the third direction ought to be ignored when investigating cell-cycle related genes and expression behavior. But this third directions is telling us about something NON negligible that is going on in the data. It clearly allows us to isolate the two first points and the three last points.

orig_vbls	Mean	StDev	R-sq, pca1,2	R-sq, pca1,2,3
<u>t10 drug</u>	-0.2293	0.7234	24.00%	72.10%
<u>t30 drug</u>	-0.0479	0.8622	80.30%	92.20%
t50	-0.0488	0.8449	88.70%	89.20%
t70	0.0087	0.6263	78.50%	79.20%
t90	-0.0954	0.6228	72.60%	74.20%
t110	-0.0998	0.662	68.80%	77.10%
t130	0.1318	0.553	59.20%	59.20%
t150	0.2593	0.5609	65.70%	72.90%
t170	0.1353	0.5233	58.80%	68.30%
t190	-0.0731	0.5195	51.30%	51.80%
t210	-0.0885	0.4739	64.60%	65.00%
t230	-0.1223	0.6111	46.80%	47.20%
<u>t250 unsyn</u>	0.0279	0.5553	43.50%	72.20%
<u>t270 unsyn</u>	-0.0379	0.5011	26.60%	76.20%
<u>t290 unsyn</u>	-0.0252	0.5463	6.90%	71.40%

Drop last three? Maybe also drop first two? Then repeat the analysis. What happens to the principal components? Do we still need the third? Do the first two (and possibly the third) resemble the previous ones?

Dropping last three, third component explains much less... and “trends downward!”

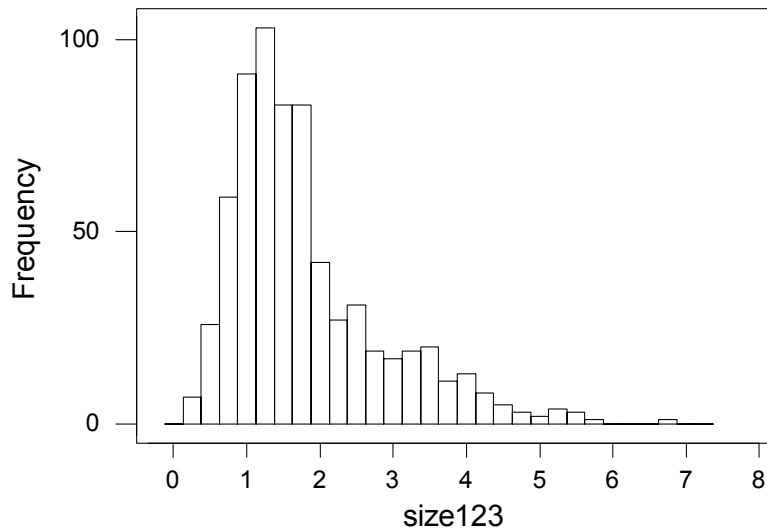
Dendrogram for original variables, pca1,2 and 3 (ave. link, Euclidean dist.)



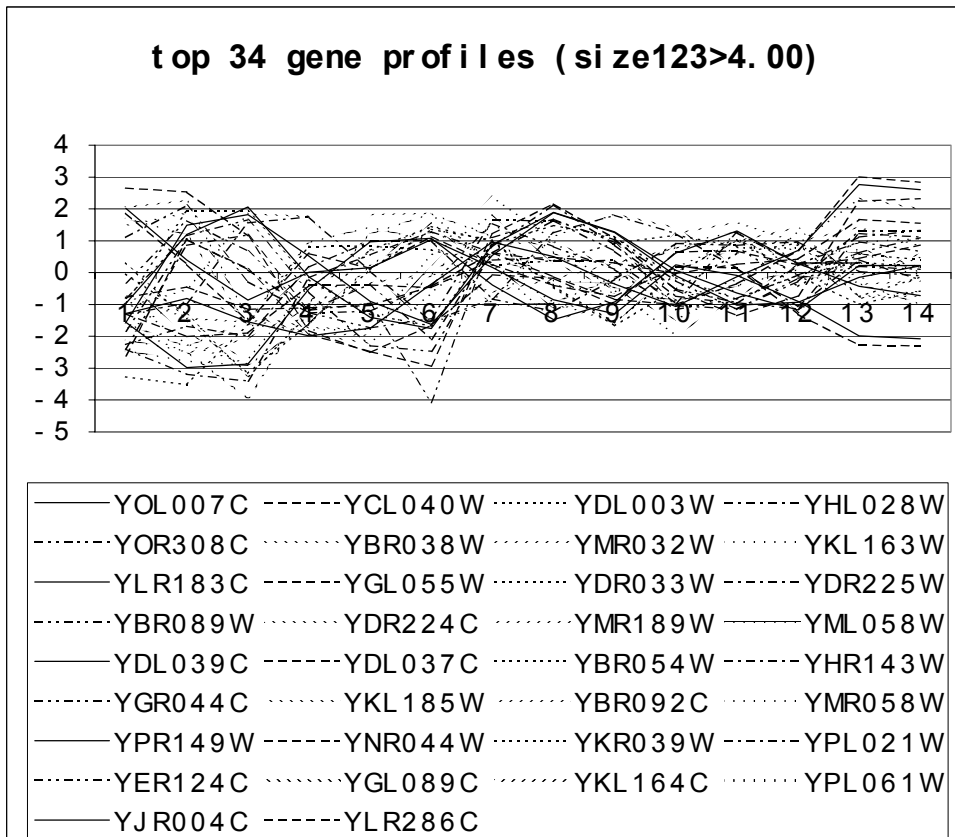
t10=1,t110=6,t210=11
t30=2,t130=7,t230=12
t50=3,t150=8,t250=13
t70=4,t170=9,t270=14
t90=5,t190=10,t290=15

What is the distinction in cell-cycle phases? Are we catching it with pca1,2 and 3? Some of you tried (FA, more than three factors though, use/no use of rotations... do “natural clusters” of time points appear?)

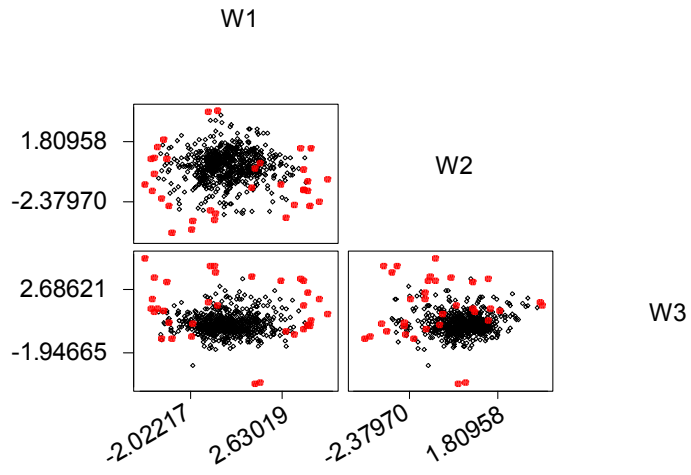
Now, let us try to *use principal components to isolate groups of genes.* Also here you tried various things.



$$\text{size123}(i) = \sqrt{W_i^1{}^2 + W_i^2{}^2 + W_i^3{}^2}, i=1 \dots N$$

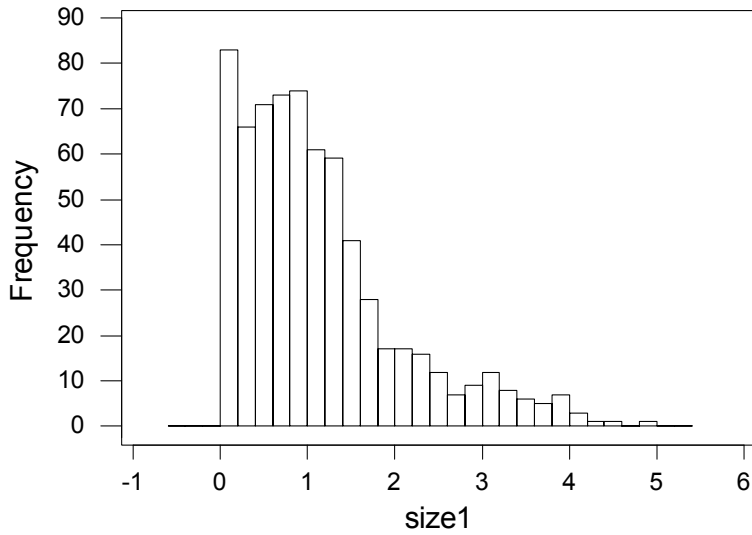


Largest projection on the space of pca1,2 and 3. Closest? Should check how long they were to start with. I haven't standardized by row.

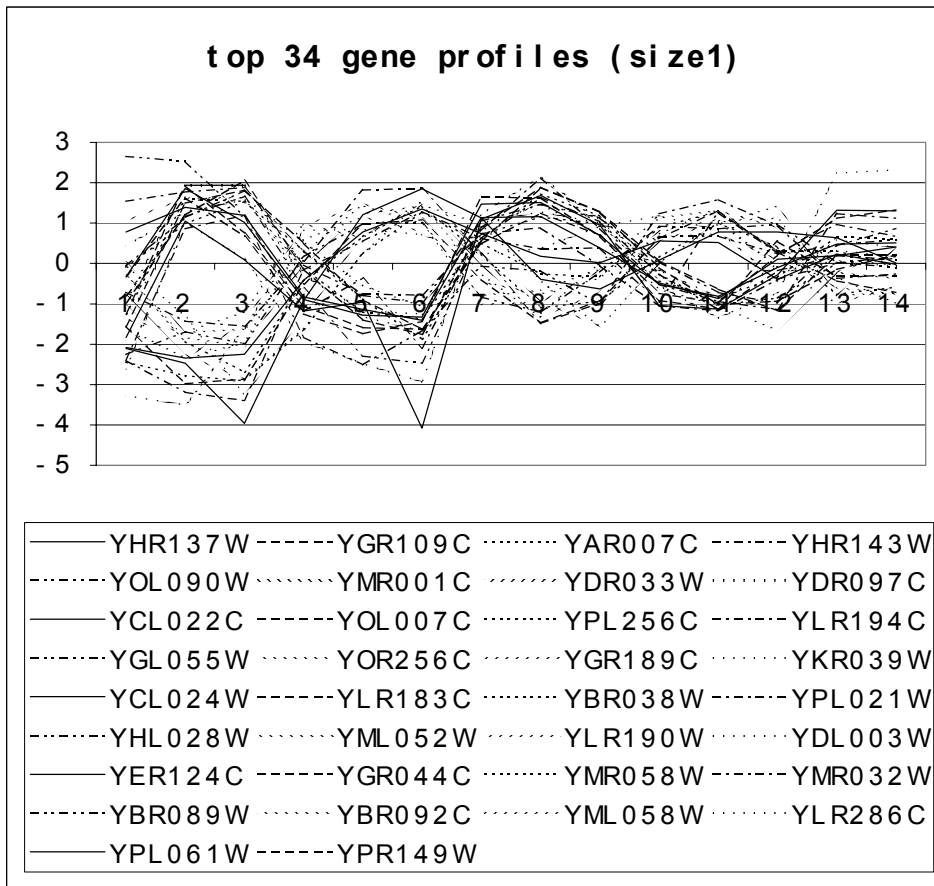


red = top 34 (size123>4)

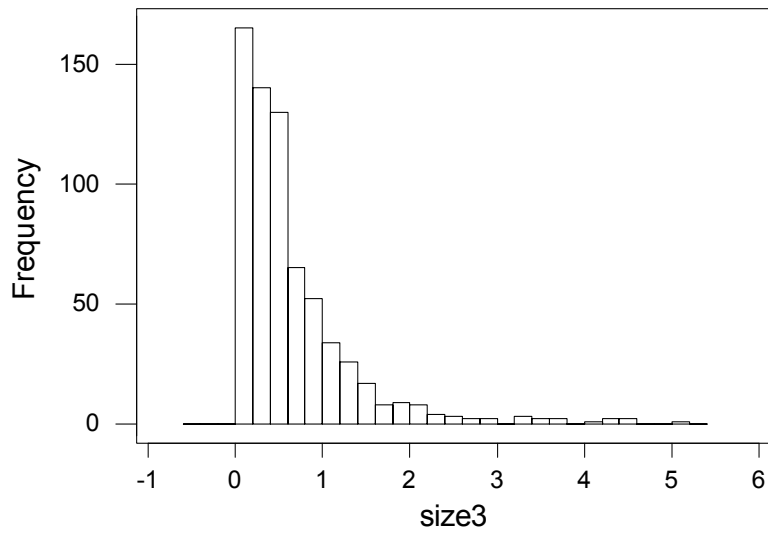
Now, using one component at a time



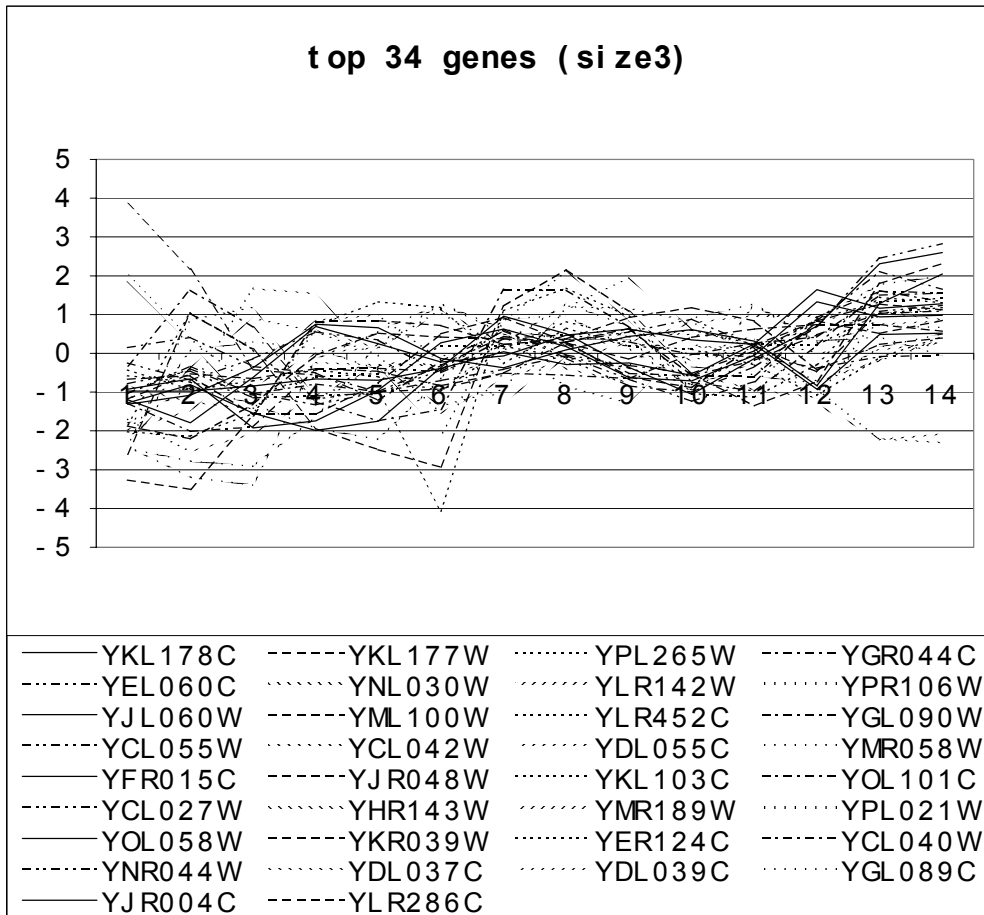
$size1(i) = \sqrt{W_i^2}$, $i=1 \dots N$



Largest projection on the direction of $pc1$ (Closest? Rescale...)
 Note sign does not matter! (close to $pc1$... or to $-pc1$).



size3(i)=sqrt(Wi3^2), i=1...N



Now, one could try removing genes that have the largest projection on pca_3 , and repeat the spectral decomposition to see

- If we still need the third component
- If and how the first two components have changed

Other ways could be used to identify a group of genes, to then investigate whether their removal would allow us to go down to 2 components, and whether the first two did change.