Opinionated Lessons in Statistics

by Bill Press

#48 Principal Component Analysis

Principal Component Analysis (PCA)

Note that the (sample) covariance of the experiments is:

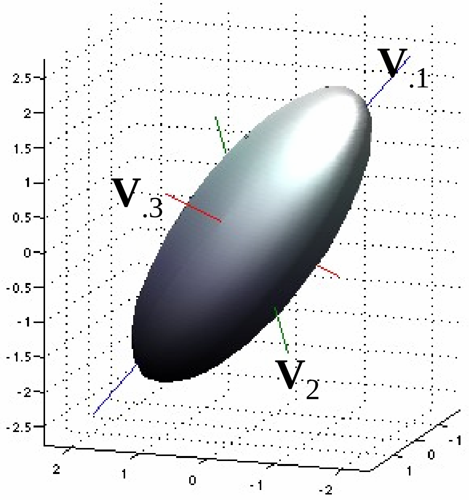
$$\mathrm{Cov}(\mathrm{Expt}_i, \mathrm{Expt}_j) = \Sigma_{ij} = \frac{1}{N}\sum_k X_{ki} X_{kj}$$

Uses fact that we subtracted the means!

$$N\boldsymbol{\Sigma} = \mathbf{X}^T\mathbf{X} = (\mathbf{V}\mathbf{S}^T\mathbf{U}^T)(\mathbf{U}\mathbf{S}\mathbf{V}^T) = \mathbf{V}(\mathbf{S}^2)\mathbf{V}^T$$

diagonal

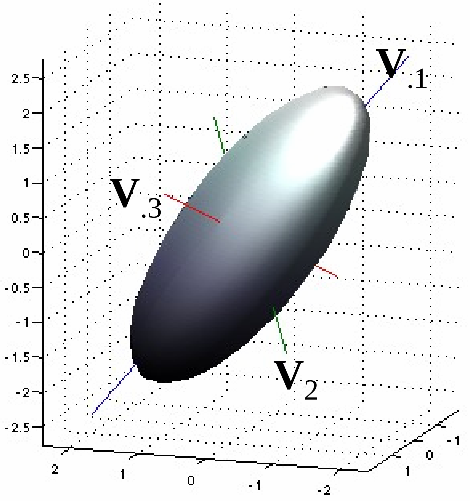
So V is a rotation matrix that diagonalizes the covariance matrix.



in our example, 300 dim space with 500 scattered points in it – and now a covariance ellipsoid

It follows that the data points in $\mathbf{X}$ have their largest variance in the $\mathbf{V}_{.1}$ direction.
Then, in the orthogonal hyperplane, the 2nd largest variance is in the $\mathbf{V}_{.2}$ direction.
And so forth.

So we might usefully coordinatize the data points by their M projections along the $\mathbf{V}_{.i}$ directions (instead of their M raw components). These projections are a matrix the same shape as $\mathbf{X}$. Since the directions are orthonormal columns, it is simply



$$\mathbf{XV} = \mathbf{US} \qquad (\text{using } \mathbf{X} = \mathbf{USV}^T)$$

rows are the data points,
column components are the principal coordinates of that row

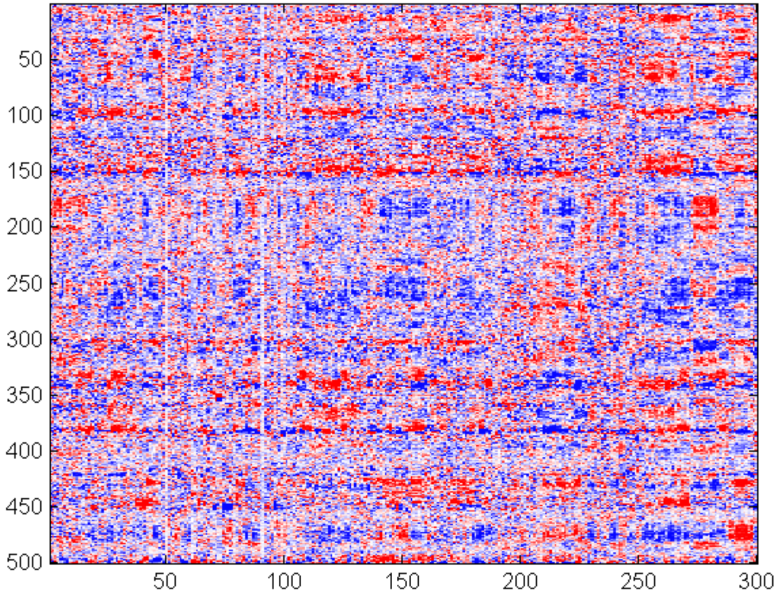
Also, it’s easy to see that (by construction) the principal components of the points are uncorrelated, that is, have a diagonal correlation matrix:

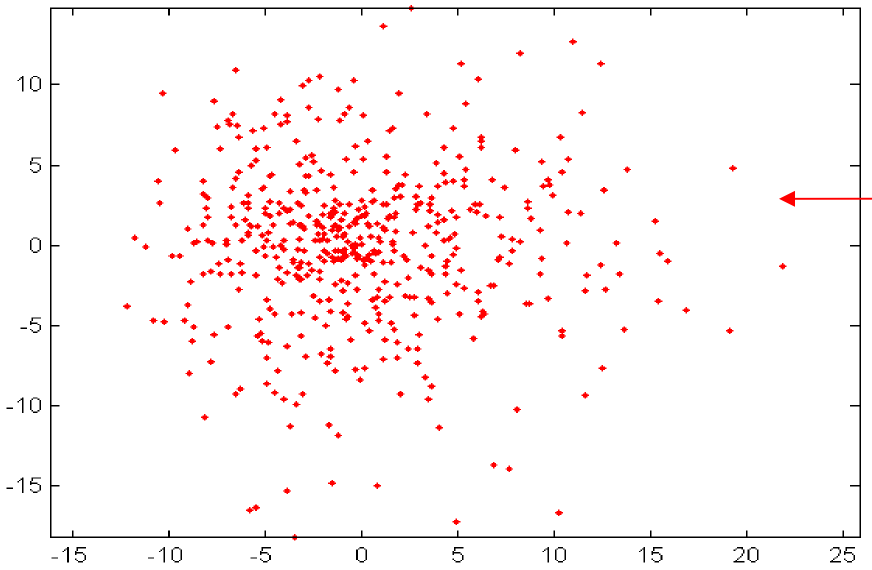
diagonal

$$(\mathbf{XV})^T(\mathbf{XV}) = (\mathbf{US})^T(\mathbf{US}) = \mathbf{S}^T(\mathbf{U}^T\mathbf{U})\mathbf{S} = \mathbf{S}^2$$

Lets plot our expression data in the plane of the top 2 principal components:

```
[U S V] = svd(data, 0);
pcacoords = U*S;
plot(pcacoords(:,1),pcacoords(:,2),'r.')
axis equal
```



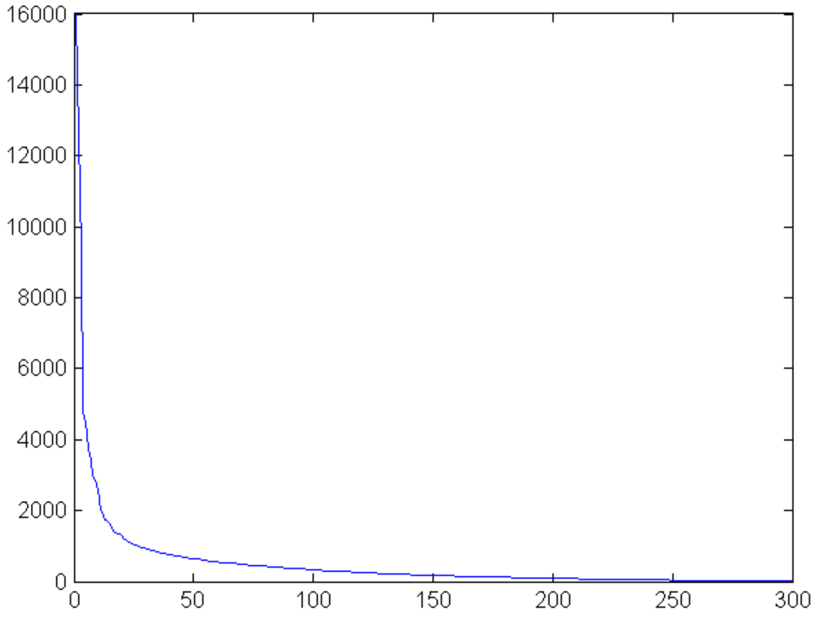


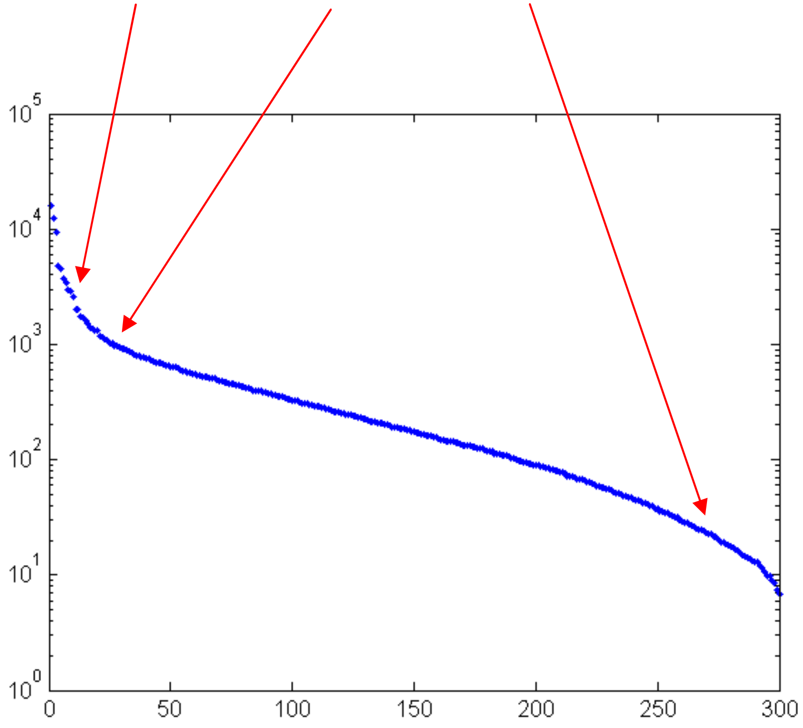
Direction 1 has larger variance than direction 2, and there is no correlation between the two, all as advertised.

As already shown, the squares of the SV’s are proportional to the portion of the total variance (L^2 norm of $\mathbf{X}$) that each accounts for.

```
ssq = diag(S).^2;
plot(ssq)
semilogy(ssq,'.b')
```

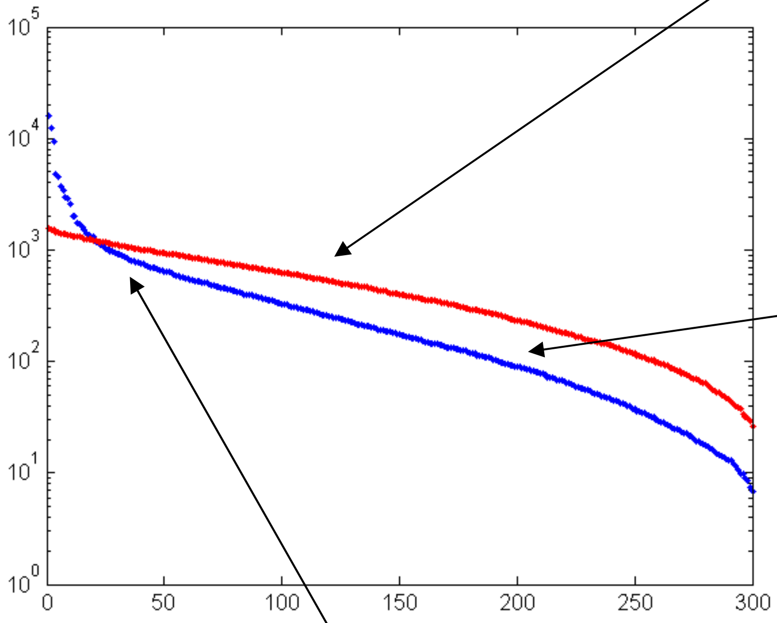
Where do these values start to be explainable simply as noise? Here? or here? or here?





One way to inform our guess as to what is signal (vs. noise) is to compare to a matrix of Gaussian random deviates:

```
fakedata = randn(500,300);
[Uf Sf Vf] = svd(fakedata,0);
sfsq = diag(Sf).^2;
semilogy(sfsq,'.r')
```



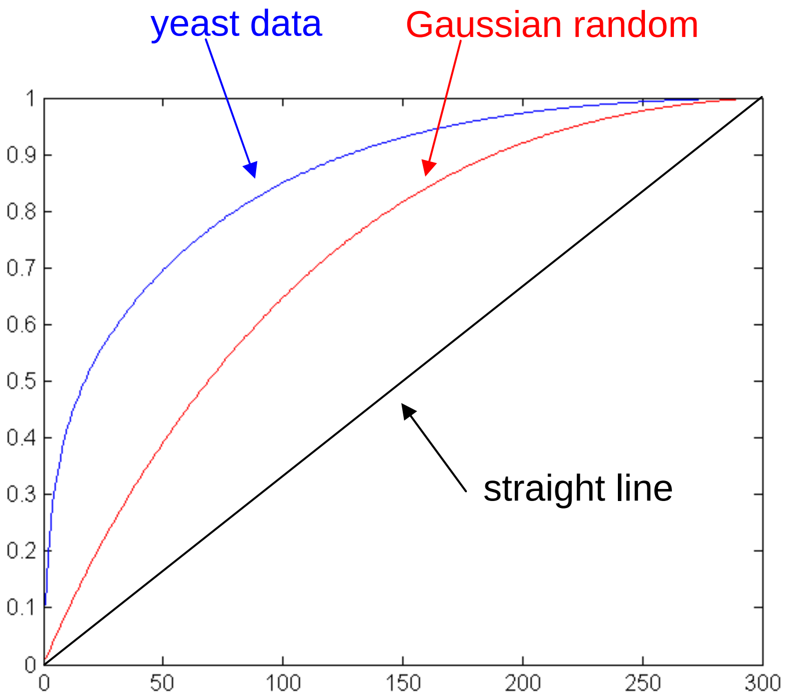
Why does fake show a trend at all? Because even random numbers are monotonic if you sort them! We are seeing the “order statistics” for SVs from a Gaussian random matrix.

Fake has to be higher than real here, because area under the curves (if they were plotted on a linear scale) has to be the same for the two curves (same total variance or L^2 norm)

I’d say it’s questionable that there’s much information in our real data set beyond around here

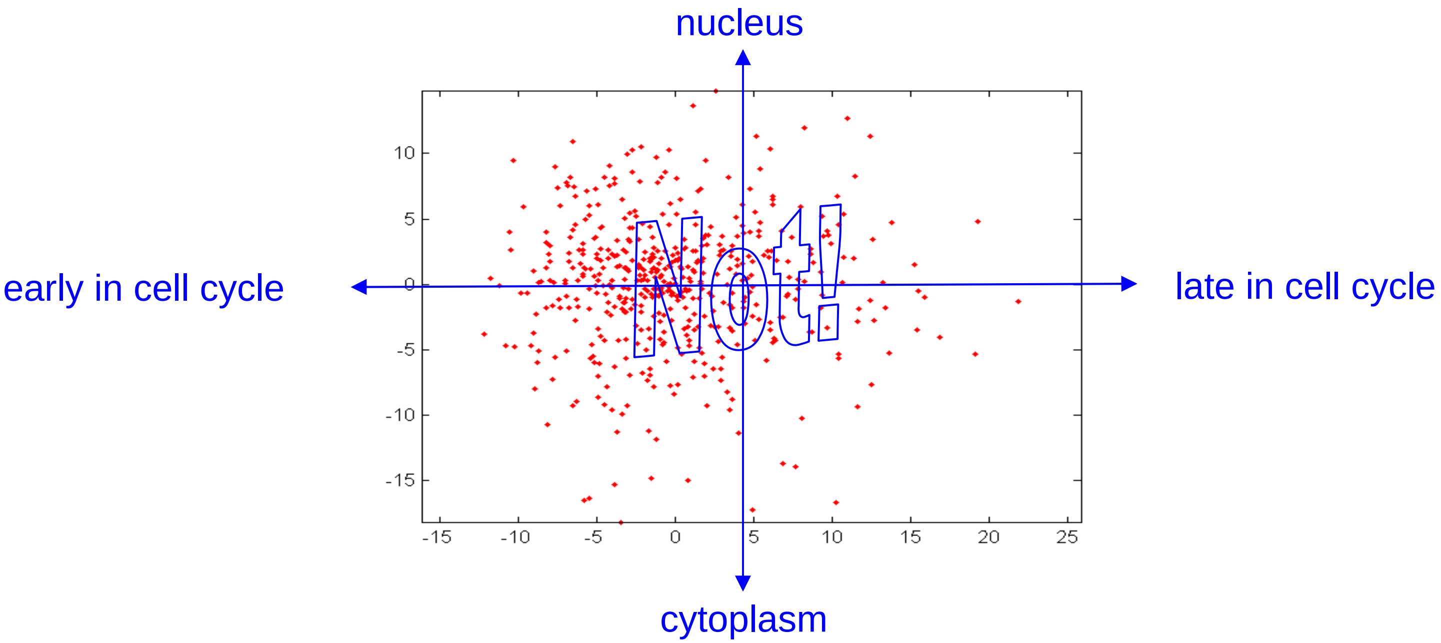
Sometimes, people plot the fractional variance as a function of number of SVs, which also shows how we converge to an exact SV decomposition:

```
ssqnorm = cumsum(ssq)/sum(ssq);
sfsqnorm = cumsum(sfsq)/sum(sfsq);
plot(ssqnorm,'b')
hold on
plot(sfsqnorm,'r')
hold off
```



You might have expected the Gaussian random to be close to the straight line, since each random SV should explain about the same amount of variance. But, as before, we're seeing the (sorted) order statistics effect. So it is actually rather hard to interpret from this plot (if you had only the blue curve) what is real versus noise and how impressed you should be by a rapid initial rise.

People who love PCA (I call them “linear thinkers”) always hope that the principal coordinates will magically correspond to distinct, real effects (“main effects”).



This is sometimes true for the 1st principal component, and rarely true after that. I think the reason is that orthogonality (in the mathematical sense of SVD) is rarely a useful decomposition of “distinct, main effects”, which tend to be highly correlated mathematically, even when they are “verbally orthogonal”.

However, it is often true that ~K main effects are captured (somewhere) in the subspace of the first ~K principal components.

So, PCA is a useful technique for dimensional reduction. Just don’t try to [over]interpret the meaning of individual coordinates!