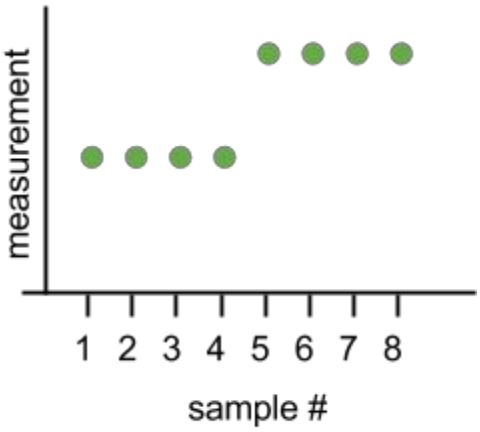
On Signal, Noise, and Variance

Zack Booth Simpson

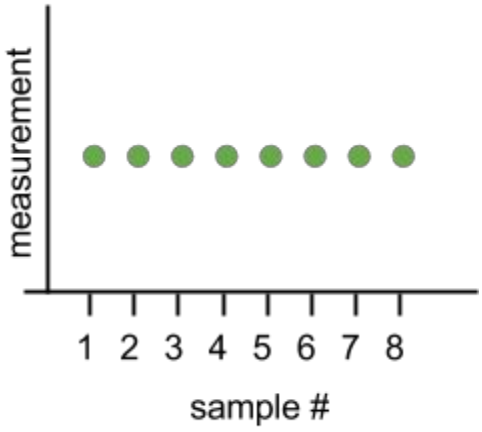
14 Oct 2014

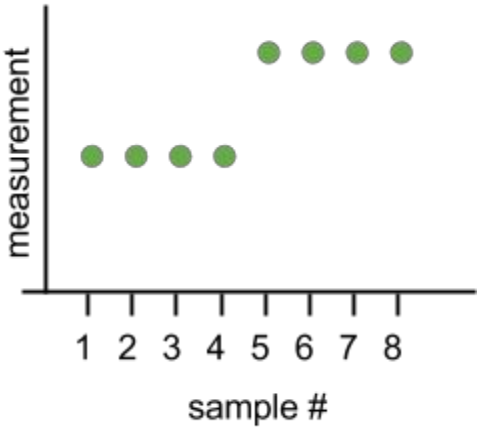
What is “signal”? What is “noise”? What are their relationships to “variation”?

Roughly: “signals” are variations in a measurement that we’re interested in and “noise” are variations that we’re not interested in. Inherent in these definitions is our knowledge about the measurement. We can not, for example, look at some measurement without context and say “That’s a good signal”. For example in the following graph there appears to be something different about samples #5, 6, 7 & 8. Is that signal?



When we measure something, we sample a “variable”. Which is exactly as it sounds: “something that varies”. If our measurement has no variability it is a very boring measurement. For a measurement to be interesting, be it made of signal or noise, it must vary.



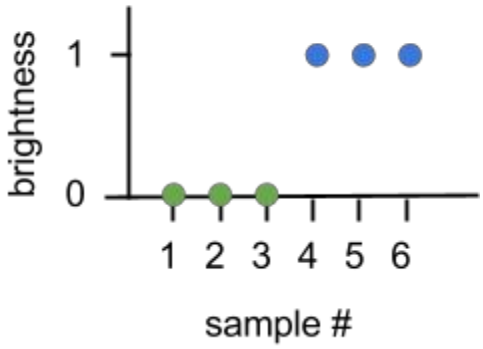


One of the most primitive things we ask about a measurement is, “What is its variance”? By this we mean, how much does it depart from a centerline? That question has a very clear mathematical answer which you may have learned, but qualitatively, the example on the left has no variance, the one on the right has a lot. Note, this is ***not*** a question about signal or noise -- it is a question of measurement; both signal and noise are by definition deviations from a centerline.

To understand why we care about variance, let’s consider an ideal experiment where we measure the brightness of something incredibly accurately.

replicate	condition 1 (control) “brightness”	condition 2 (experiment) “brightness”
1	0	1
2	0	1
3	0	1

In the above experiment, we took three replicate measurements of the system; with each replicate we sampled two conditions, one that we call “control” and the other that we call “experiment”.



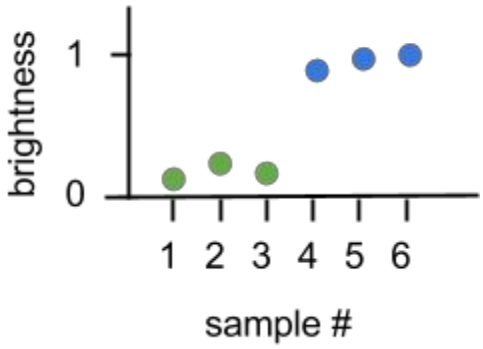
There is a clear gap between the two sets, they are well separated, and we say: “This is great! There is a big signal differentiating our control and experiment!” Not so fast.

How do we know that the measurement is really signal? Suppose that we took all the experiment samples on day 1 and all the controls on day 2? Maybe there was some ***bias*** on those days (maybe the temperature was different, maybe the media was a day older, etc.). Just because we see something that looks like “clean signal” doesn’t mean it is signal! If it is noise it

might be what we would call “structured noise” -- noise that mimics signal. The only way we can trust that something is true signal is by carefully convincing ourselves that we controlled for every conceivable factor! That might be an impossibly long list of things to control for -- but that’s **the art of science** -- we must do our best to eliminate all the sources of bias and challenge others to do the same. The critical point here is that just because something looks like signal doesn’t mean it is. Careers have been ended by confusing signal and noise!

Now, let’s imagine a situation with “uncorrelated noise”. In this case, there’s a little bit of error in our reading of our instruments and we thus see that our measurements are not quite as perfect as before.

replicate	condition 1 (control) “brightness”	condition 2 (experiment) “brightness”
1	0.1	0.95
2	0.2	0.97
3	0.15	1



This extra jitter -- that, you say, looks like noise. Right? Why? Because it is a variance that is uncorrelated to what you think is signal. That is, the extra variance doesn't seem to be related to what the signal did. The noise didn’t rise and fall along with the signal and it wasn’t stronger when the signal was larger. It was just jitter that seems to be independent of what the signal is doing. That’s the definition of “uncorrelated noise” but we must be careful because that’s not the only kind of noise. As described above, there might be bias which is just noise that looks like signal (structure noise) or there might be noise which is correlated to the signal.

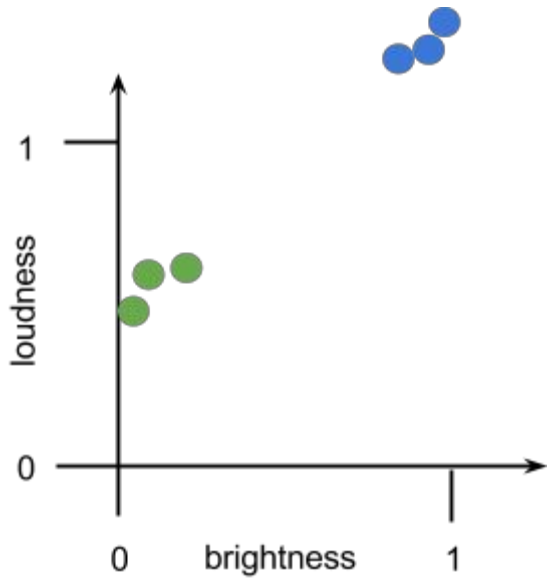
Noise, in general, is something we don't understand or can't control. If it were otherwise then we'd presumably have made an attempt to control for it and if we didn't then we're probably going to get called out for it in review.

Co-variance

So far we've considered only a single dimension "brightness" to our hypothetical measurement. Let's now say that we also measure "loudness" for each condition. We are said to have "two ***observables***" in this case.

We might think that this new measurement is independent. "Loudness", we might believe, has nothing at all to do with "brightness" in our system. But that's a guess on our part and we must remember that. It is easy to forget that just because we measure two things doesn't mean that there are two things to be measured!

replicate	condition 1 "brightness"	condition 1 "loudness"	condition 2 "brightness"	condition 2 "loudness"
1	0.1	0.5	0.8	1.2
2	0.2	0.65	0.9	1.25
3	0.3	0.7	1	1.4



Above we make a two dimensional scatter plot of our situation. We observe a few facts. The loudness measurements (vertical) are offset relative to the brightness measurements

(horizontal) and there is a correlation (diagonal trend) between the two. There is still good separation between the control and the experiment but both measurements seem to vary some and if we're lucky that's because of uncorrelated noise. If it isn't uncorrelated noise then it is one more thing we've got to go control for!

There's a lot of variance on the brightness axis and a lot on variance on the loudness axis, so we might say "There are two strong signals here!" But that's clearly wrong -- the two variables are ***correlated***; they are in effect just two aspects to the same thing. If we knew the value of brightness we could very reliably predict the value of loudness and vice versa. This is also called a high "***covariance***" between the measurements.

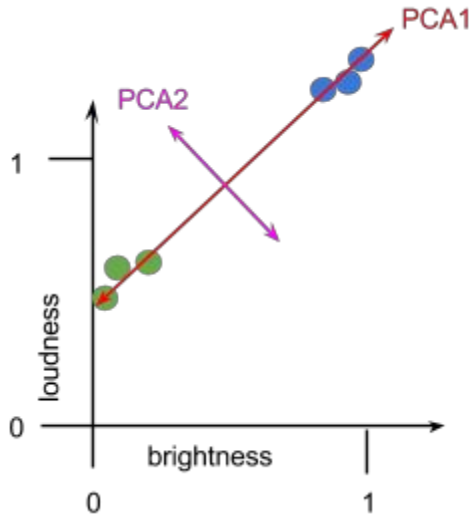
Sometimes correlation like this is unexpected and might be a novel discovery. You might, in that situation, go and apply some statistics to measure how confident you are that the correlation was not produced by chance. Other times the correlation might be completely obvious or be a well-known relationship. Either way, the one thing that is certain is that you ***do not have two independent variables*** in this situation, ***you have only one***.

A fancy mathematical way of saying the same thing is that the "dimensionality" in the direction of maximum variance (which we hope is signal) is lower than the dimensionality of the whole space. Specifically in the above example we have a two dimensional system (we measured two variables) but the variance is only in one dimension and the direction of that dimension happens to be diagonal -- it is not aligned with either axis.

This correlative situation is easy to see when there's just two measurements, but it is hopelessly difficult to see when there are more than three dimensions because you can't plot it. In other words, when there are a lot of dimensions you can't trivially see correlations.

What would be really helpful is if there was a mathematical tool which would find the direction of maximum variance in a high dimensional system. Then we'd be able to ask ourselves the question: "Is there any interesting variance in this system at all -- in any direction?" Maybe we can't see it because it is rotated in some way relative to the axes we actually measured. If there such is a direction, great, maybe that's our signal; then again, maybe not, it could be noise and as discussed previously only we can be the judges of that. **But if there is *no such direction*, that is, if all directions are more or less just scattered equally then we know that there's no hope of finding signal and we need to change our experiment.**

Fortunately, there is such a tool. It's called "Principal Component Analysis" and it does exactly what is described -- and more. After you push your measurements through the PCA process, you get back a direction (that is a combination of measurements) that is in the direction of maximum variance. It then goes on to give you more -- it gives you a second direction which is guaranteed orthogonal to the first direction which is the direction of the second-most variance. And so forth, one for each dimension of the original system it gives you mutually orthogonal directions.



In the above plot the first and second PCA directions are indicated. Note that the first direction is the direction of maximum variance and the second is orthogonal (perpendicular) to the first and in this case is probably an axis of noise.

To repeat: just because there's a direction of maximum variance does *not* mean that is the direction of signal! It means that it's the direction of maximum variance, nothing more. Only you can judge whether or not something is signal -- no fancy math will do that for you.

That said, if there is some signal in the system then you certainly hope that said signal is in the direction of maximum variance. It is an "ugly" experiment if the direction of maximum variance turns out to be artifact while the signal of interest is buried in some other direction. That said, sometimes that is exactly the situation whereby the first principal component(s) are known artifact(s) and your signal is buried deeper within. In that situation you should be prepared to get spanked for not controlling that source of noise or bias.

Example of PCA in Practice

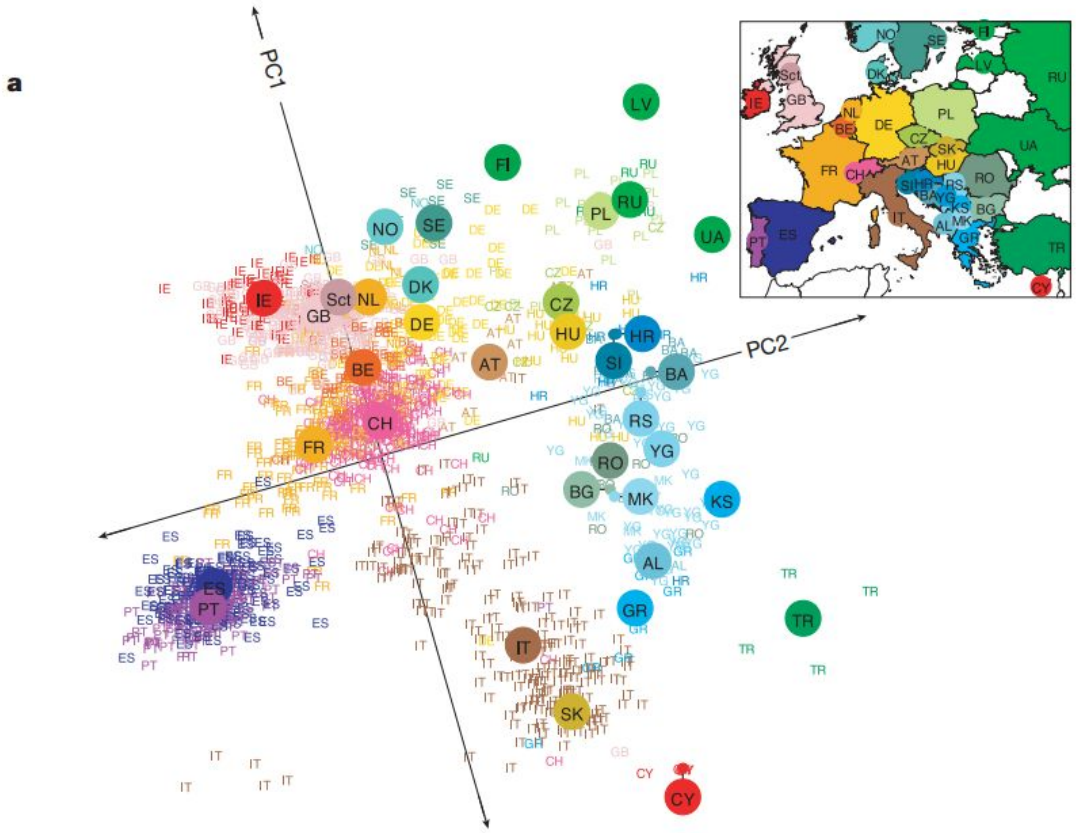
A lovely illustration of the power of PCA can be seen in the extraction of European ancestry from Single Nucleotide Polymorphisms (SNP) in a cohort of European men. (See a copy of the paper here: http://www.marcottelab.org/users/BIO337_2014/EuropeanGenesPCA.pdf).

The analysis begins with a table in the following form:

	Person 1	Person 2	...Person N
SNP 1	1	1	1
SNP 2	0	1	1
...SNP N	1	0	1

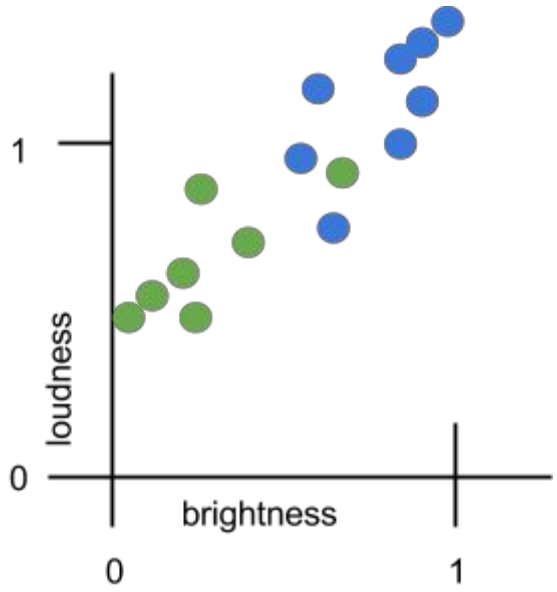
Each column is a person in the study and each row is a SNP position that each person either has (1) or doesn't have (0). (The actual analysis in the paper is slightly more complicated to deal with homo/hetero-zygous individuals, but this isn't germane.)

There are approximately 3000 people (columns) and 500,000 SNP loci (rows). This is a very high dimensional system indeed. Trying to stare at the data would get you nowhere. But, miraculously, PCA on the matrix turns out two very strong principal components that happen to characterize the coordinates of the origin of the individuals. In other words, when you project all the data points onto the plane formed by the first two axes of maximum variance that fell out of the PCA analysis then the data recapitulates a map of Europe as shown below!



Classification

It may sometimes be the case that the separation between your conditions is not as ideal as you'd like. Consider this case:



In the above situation the two conditions bleed into each other. We have a less clear signal than we had before and it is therefore harder to say where the line is that separates the two conditions. This is called the "classification" problem. Where do you draw the line between the conditions given the evidence? This is the kind of question that is answered by statistics and machine learning and beyond the scope of this paper. But, note that we'd probably want to perform PCA on this as a pre-step to those fancier techniques so that we can characterize what we believe is the correlation in the signal before we start classifying it.

Summary

Signal and noise are largely human concepts, not mathematical ones. Variance is a mathematical concept that is related to signal and noise in the sense that both have variance. Sometimes when you conduct more than one measurement in an experiment there is correlation between the observable variables -- they co-vary. When the dimensionality of the system is high, it is very difficult to visualize the directions of maximum covariance. The mathematical tool Principal Component Analysis (PCA) extracts the directions of maximum variance and is a helpful tool to find out if there is any possible signal in a set of measurements. But, we learned to always remember that directions of high variance are not necessarily signal!