

I (heart) factor analysis!

Eric Blair

28 August 2004

I've made the comment often enough that it's not fun anymore, but stats textbooks often blur the distinction between descriptive procedures and hypothesis testing procedures, and this creates endless woes in the world—notably, researchers are often taught to do exploratory stuff on data and then test the hypotheses that they'd just cooked up using the *same* data, which adds to human knowledge in some ways, but the hypothesis test is completely and totally invalid, with 100% certainty. Won't harp now; have blogged about this in entry #041.

There's something of a focus on hypothesis testing with most of this stuff since, again, it's the hypothesis test that gets you published. So this little column is to remind the reader of one of those things that gets left by the wayside: factor analysis (aka singular value decomposition or principal component analysis). It's more advanced than the basic descriptive stats (mean, variance), but not a hypothesis test, so in my own limited experience it seems to get left by the wayside.

Before I lose every last one of you, here are two super-hip examples of why factor analysis is fun. The first is Poole & Rosenthal's analysis of the U.S. Congress. They found that about 90% of the variance in voting patterns can be described in two dimensions, so that can be made into a graph; string them together, and you have a movie of the U.S.'s political history.

Second example: Eharmony.com has a system to match you with someone you're most compatible with, and how do they do it? Factor analysis, of course. They give potential love birds a survey of a few hundred questions (oh, the tedium!) and then map the test-taker in a few artificial dimensions. They find people who are close in the artificial space, and have them go out on dates. Romance blossoms.

[It would be rude of me to encourage the reader to use factor analysis without mentioning that Eharmony has patented their matchmaking algorithm. Therefore, when you do your factor analysis, you must be careful that you don't assign variable names that sound too much like 'sexual passion', 'spirituality', or 27 other such terms—if you do, then you are in violation of the patent and could be sued for damages. As of this writing, most other names one could assign to the dimensions resulting from a factor analysis are still in the public domain, and can be used without a licensing fee.]

Oh, and my other favorite is a paper by Moore et al (citation available on request) about color perception: for sighted people, factor analysis neatly puts the perception of words such as 'red', 'blue', and 'yellow' in two dimensions, in a circle—a color wheel. Factor analysis of responses to the same words by the blind fall on one dimension, basically ranging from bright to dark. And thus, factor analysis shows us what the blind see.

There are two ways you could go about it, I guess: the first is to say, ‘I have no idea how the data was generated, but darn it, I want a picture’, in which case a two-dimensional factor analysis fits the bill wonderfully, and once you get the picture, you just might learn something (even if it can’t be formally tested). The other is to say ‘I really think there are some latent variables driving the variables I’ve observed’, and then factor analysis may again save the day, by showing you the best linear combination of existing variables to suggest what those latent variables may be. Both of these sorts of behavior are exactly what statistics is really about, and think they’re a great thing to try on any given data set.

This all comes up because I have some data on how junior high kids smoke. I have zero information about the kids, not even their gender, since the guy who collected the data wants to play with it more and get some more publications out before he puts it out for others to use. So I’m not assuming that there are important latent variables underlying what data we have—I *know* that there are important missing variables underlying the data we have. The factor analysis did a very neat job of pulling out what I reasonably believe are the most important underlying characteristics, thus saving the day. [The hypothesis tests, by the way, are about ten steps further down the line in the project, so no, no worries about confounding the two. Won’t bore you with the details.]

So don’t forget to use factor analysis on your favorite data set, and maybe write a paper or two about the results, since good results from a good factor analysis really can teach the world stuff. Also, next time you’re asked to review a paper, please remember that good descriptive stats can stand by themselves without being bolstered by a table of statistically sound but procedurally bogus linear regressions. After all, the true level of explanatory power and persuasiveness of a paper just isn’t measured by confidence intervals.