Guillaume Pagnier

April 4th, 2019

Multivariate Midterm

1. Linear combinations.
2. Linear combinations represent summing up different variables (or vectors) to end up with an aggregated variable. This is done by assigning a weight (a constant) to each variable and then combining it with other variables to create a ‘weighted combination’ of the original variables. If A and B are two vectors, then 4A + 3B would be a new vector, i.e. a linear combination of A and B.
3. Conceptually, a linear combination is a type of addition/combination as you can take many different vectors or variables, assign an appropriate weight to each of them and essentially reduce those variables to a single vector.
4. In multiple regression, the regression model you’re generating to predict your DV is actually just a linear combination. For instance, the generic multiple regression model DV=B0 + B1X1 + B2X2 + … + error is a linear combination of the beta terms (the x terms represent constant weights assigned to each beta). The weights assigned are specifically designed to maximize the explained variance.

In PCA, the generated orthogonal components are just linear combinations of the original variables. The weighting constants that are chosen (in PCA’s case, it’s a weighting vector) are specifically selected to generate components that maximize the variance of the data (i.e. be able to maximally distinguish participants from one another).

Contrasts are cases of linear combinations of experimental (i.e. factor) means. An additional restriction here is that the assigned weight constants have to sum up to 0. As a result, the resulting linear combination expresses the sum of squared differences between the means of different factors and allows you to test if there is a significant difference between these factor means.

1. Two types of reliability would be 1) Cronbach’s alpha (describing internal consistency) and 2) interrater agreement (describing external consistency). Cronbach’s alpha represents how consistent items in a group are. This is especially used to measure scale reliability. Interrater agreement quantifies individual differences between humans when classifying subjective behavior.
2. I would use Cronbach’s alpha if I was designing a survey that gauged an individual’s extraversion using a battery of questionnaires. A high Cronbach’s alpha would indicate that my questionnaires were consistent with one another and I could trust that my battery of surveys is reliable. I would look at interrater agreement if I was scoring a video for facial expressions of fear. This is a relatively subjective analysis and there may be individual differences in my RAs in classifying what a fear expression looks like. In this case, if the scores my RAs give the same trial are similar, then I have high interrater agreement and the data would be reliable.
3. Two types of validity would be 1) content validity and 2) predictive validity. Content validity is the degree to which you’re capturing what you think you’re actually capturing i.e. is the latent construct you’re measuring the one that you’re really measuring? Predictive validity represents how valid your predictions that you’re hoping to make actually are in the real world.
4. An example of content validity would be providing an extraversion questionnaire and having that questionnaire actually be capturing extraversion, and not agreeableness or something else. In a different example, If I designed questionnaire ZZ that was meant to predict an individual’s financial success in 5 years, ZZ would have high predictive validity if it actually DID measure an individual’s success in 5 years using real world data.
5. It is true that EDA could theoretically enable p-hacking. There WILL be random spurious patterns in your collected data that will appear interesting and could certainly bias the researcher towards conducting a hypothesis test they would not have conducted otherwise. In such cases, the researcher is indeed at a greater risk of making type 1 mistakes since they’re capitalizing on spurious patterns and either consciously or unconsciously driving their conclusions from these spurious results. However, this can very simply be avoided by knowing beforehand exactly by the researcher knowing what their questions are and how to answer them. If I look at some pilot data, and already know what tests I want to run and what patterns I want to look for, then it is impossible for EDA to have a negative effect since my strategy for analysis wouldn’t change when analyzing that particular set of data.
6. The strengths of EDA far, far outweigh the potential pitfall mentioned above. EDA is powerful for many reasons, one of which it TELLS you what you’re looking at. In the case of missing data or miscoded data, appropriate EDA will quickly alert you to these errors. EDA gives the researcher an idea of the distribution of the data and alerts them to any outliers or atypical responses, knowledge which is crucial when they are trying to understand what their data means. EDA allows confirms that the assumptions the researcher has in order to conduct a hypothesis test are actually true (failed assumptions – which would be invisible without EDA – can muddy up hypothesis testing). Most importantly, it allows for patterns in the data to be revealed. These patterns could be the source of future questions (to reduce the risk of a type 1 error, the researcher should replicate unexpected patterns in a new data set) or reveal different confounds that the researcher was unware of before.
7. 𝚲 is a diagonal variance-covariance matrix of the eigenvalues. All of the off diagonals are always 0.
8. Each of the entries in the side diagonal of 𝚲 represents an eigenvalue, or the variance of a newly generated primary component. These eigenvalues are ordered such that the maximum variance is the first, the second most variance is the second and so on. All of the off diagonals, which normally represent covariance, are 0 because the generated primary components are all orthogonal to one another so there is no covariance.
9. Starting with Rxx which is a correlation matrix, SPSS conducts spectral decomposition to reconfigure R into a new ordered variance-covariance matrix, 𝚲 that consist of linear combinations of the variables of R. Spectral decomposition represents this transformation and is done by multiplying R by eigen vectors (weights) to get the new linear combinations (which end up being our generated components). How does SPSS know what eigen vectors to use? This is the crux of PCA/spectral decomposition: SPSS uses weights that will subsequently generate components that account for the maximal variance of the original data. Said another way, SPSS essentially creates a new component (axis) by using the covariances in R (off diagonals of R) to stretch the variance (diagonals of R) in a way that maximally differentiates between people. This would then represent one generated component. SPSS repeats the process for any variance that couldn’t be explained by the first component to create a new, second component (this is why the generated components are orthogonal to one another since all of the generated components are explaining ‘different’ variances). SPSS repeats this process until all of the variance explained by R is parsed out onto the new generated components. The maximized variances of the calculated linear combinations (i.e. newly generated components) are called eigenvalues and compose the diagonal of 𝚲 in an ordered fashion with the highest number being the top number in the diagonal. As mentioned above, by nature of systematically maximizing the variance of the data with the new components, all of the new components will be orthogonal to one another and the off diagonals of 𝚲 will be 0.

<Overheard at the annual Data Reduction For Everyone meeting>

PCAFan (PCAF): Can you believe that people still opt to use barbaric subjective techniques like factor analysis? Objectivity is the only correct way to analyze data: if I create new components, I want to be sure someone else can replicate my work and calculate the components that accomplishes the goal I set for it: maximizing the variance of the original data.

FAFan (FAF): Well that’s not really a problem is it? Your generated components may be replicable but they won’t be very interpretable. You would be better off in using a more complete and sophisticated data reduction method like factor analysis that is really just PCA plus a few additional parameters. With factor analysis, you can generate underlying factors that account for common variance in the data and are a lot more useful to interpret. As long as I’m transparent about my methodology, anyone can replicate those same factors and interpret them.

PCAF: I can’t really trust any methodology that different researchers may argue as to what is actually real or not. PCA is a pure transformation of the data and is thus the always correct way to analyze high dimensional data. If I’m conducting PCA I specifically WANT to find the components that maximize variance of the data, altering the PCA technique in any way is just muddying it up. Factor analysis will always account for less variance than PCA.

FAFan: But factor analysis gets at the *reason* for why we’re running PCA in the first place. You shouldn’t care about the variance not captured in a factor analysis anyways. We would (should) have reason to believe that underlying constructs exist and any uncorrelated noise between variables is just measurement error. Factor analysis allows us to remove that error (which you shouldn’t care about anyways) and then conduct PCA to reveal the interesting latent driving constructs of the data. Think of factor analysis as a cleaner version of PCA. AND factor analysis has multiple extensions such as confirmatory factor analysis or independent component analysis that can offer concrete answers to specific problems. PCA will always just be PCA.

I actually think the difference between FA and PCA is overstated. They both generally use covariances to generate new components that explain the most variance of the data, it’s just that they have different objectives/assumptions. Which one I use would be dictated by my question but I think I would generally use factor analysis since I would be creating a study that sought to capture some latent constructs. With this in mind, any uncorrelated noise should actually be error and thus, I would be comfortable in removing it and moving forward with factor analysis.

Multivariate Statistical Techniques

CLPS2908, Spring 2019

Prof. Bertram Malle

**Take-home Midterm Exam**

#### Rules of Engagement

**You may use any material you have available to answer the questions (including books, notes, web pages), but you must write your answers in your own words. If you use outside sources, reference them (on a separate page). Most important, you must not collaborate with any other person, inside or outside this class.**

**If you have questions of clarification, please address them to me (BFM).**

Each of your answers must be on 1 page or less (measured at 1.5 line spacing, 1" margins, 12-pt Times New Roman font). Exceeding the length restriction leads to point deductions.

Number the subparts of your answers (a, b, etc.) clearly, but do not include the question text.

The completed exam is due on Thursday, April 4, 2019, at 5:00 p.m. Please submit it through the assignment module in canvas as a .doc(x) or .pdf file.

**Before you turn in the exam, sign or type your name on the line below, which indicates that you have read the rules of engagement and have abided by them throughout your work on this exam. Then add this page to your submitted document.**

**Signature:**

Guillaume Pagnier