SESUG Paper HOW-211-2023

Introduction to Principal Components and Factor Analysis

Jason Brinkley, PhD, Abt Associates

ABSTRACT

Principal Components Analysis (PCA) and Principal Factor Analysis (PFA) are popular analytic tools that can help bring structure and understanding to datasets with many unique variables. PCA is the backbone of these methods and serves as a mechanism for reducing the dimensionality of a large dataset into a smaller set of variables or characteristics that retain information from the original data. PFA is a core component in classical testing theory and psychometric analyses for instrument creation. These methods are complex but have a rich history of use across the entire realm of data sciences. This slides parallel a hands-on workshop which provided a general overview of PCA and PFA. The focus and use cases are be two-fold. First, we will discuss PCA in the context of data reduction and insight generation. Second, we discussed the basics of creating composite measures via PFA and exploratory factor analysis. The workshop is hands on and designed for non-statisticians who have a background in both descriptive statistics and regression analysis.

INTRODUCTION

Principal Components Analysis (PCA) and Principal Factor Analysis (PFA) are popular analytic tools that can help bring structure and understanding to datasets with many unique variables. PCA is the backbone of these methods and serves as a mechanism for reducing the dimensionality of a large dataset into a smaller set of variables or characteristics that retain information from the original data. PFA is a core component in classical testing theory and psychometric analyses for instrument creation. These methods are complex but have a rich history of use across the entire realm of data sciences.

Principal Components Analysis (PCA) and Principal Factor Analysis (PFA) are popular analytic tools that can help bring structure and understanding to datasets with many unique variables. PCA is the backbone of these methods and serves as a mechanism for reducing the dimensionality of a large dataset into a smaller set of variables or characteristics that retain information from the original data. PFA is a core component in classical testing theory and psychometric analyses for instrument creation. These methods are complex but have a rich history of use across the entire realm of data sciences.

This paper is an extension of a hands-on workshop whose goals are to provide a general overview of PCA and PFA. The focus and use cases will be two-fold.

- First, we will discuss PCA in the context of data reduction and insight generation.
- Second, we will discussion the basics of creating composite measures via PFA and exploratory factor analysis.

The workshop is designed to be hands on and designed for non-statisticians who have a background in both descriptive statistics and regression analysis. As such not all of the concepts translated well into a working white paper. Please see the references for a more in-depth discussion on PCA and PFA.

CONCEPTUAL EXAMPLE – HEIGHT AND WEIGHT

PCA and FA serve as mechanisms to reduce a large quantity of quantitative data into a smaller set of variables that are specifically designed to ‘retain information’. The goal is to go from many numeric variables to a smaller and manageable number of summary variables that retain as many of the relationships and insights from the original data as possible.

Height and Weight are correlated variables in adult humans. In general, the taller a person is, the larger their frame is, the heavier that frame will be. There are tall people who are underweight and there are short people who are overweight, but in general there is a baseline association between these two

variables. Body Mass Index (BMI) is an attempt to make a surrogate for 'size' that represents an amalgamation of height and weight data. BMI creation and calculation was mostly clinically driven, with the focus on maintaining the relationships and having better explanations without having to resort to using the raw height and weight data. BMI's purpose is to give a single summary measure that can be used for other analyses.

BACKGROUND ON PCA AND PFA/FA

PCA and FA are techniques that do similar tasks but are data focused instead of clinical focused. They help determine how to make summary composites based on the data itself instead of based on clinical insights. How does all this work? Well we will start with PCA because it serves as the basis for Principal Factor Analysis (PFA) or sometimes just referred to as Factor Analysis (FA).

PCA is a unique and beautiful methodology in that it has an algebraic, matrix, and geometric set of definitions and interpretations. It is at core a modeling framework that is more akin to least squares (regression) than it is to other multivariate analyses.

Suppose I have a series of quantitative variables, denoted as a vector $\boldsymbol{X}$ which we can define as a series of p different numeric indicators:

$$\boldsymbol{X} = (X_1, \ldots, X_p)$$

Suppose I want to reduce that to a smaller set of variables, one way to do that is to combine variables and reduce redundancy. I know that I don't need to keep age in both years (with decimal precision) and age in months. They convey the same information. But how far can I systematize that? The first step of a PCA is to take linear combinations of variables to assist in that reduction.

So let's make a new indicator, call it C, and it is a weighted sum of the original data:

$$C = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p$$

So I want to choose some numeric coefficients so that this summation is 'useful'. That is to say that if I can come up with a clever way to summarize a couple of these variables then that sum may have all the important information, I need for analysis but will allow me to abandon the use of the original variables. BMI does something akin to this, but it takes a ratio (which can be written as the sum of two log transformed variables). But PCA's purpose is to take the 'best' weighted sum. We define the best solution to this problem as the one that maximizes the variance in C. That is to say that if we pick a weighted sum that maximizes the variance of C (or associated standard deviation), then we are representing information as 'variation' and our composite helps to showcase how much variation there was in the original data. Maximizing variance across data, that should sound a little like least squares regression which also models data around variation.

One cannot just select 'large' weights for this problem. Large weights just shift the data and don't actually change the scale of the variance. What I want is to create a weighted sum in such a way as to make that variance useful and to do that I need to consider information about covariance or correlation between variables. PCA looks for a solution that maximizes the variance of C with some constraints on the standard deviations of your variables in X and the correlations between them. So we define a principal component as a linear combination of optimally weighted observed variables. The weighting means that the resulting component accounts for the maximal amount of observed variance in the dataset.

It is mathematical fact that a PCA can't give you a solution whose variance is higher than the variance in the original data, but that doesn't mean that the highest variance in X is the upper bound on your component variance. But it's a good bellwether. The performance of your first PCA component is dependent on how much variation and correlation there is in your raw data. In almost all instances, we cannot combine two or more indicators into a single composite retaining all information unless the

variables are perfectly correlated. In which case we are back to the age by years or months, and it is trivial. So, in almost all instances, you will create more than 1 component. But what does that mean? To help make sense of how PCA is applied here are some important properties:

- You can't have more components than you have variables in your original data. If you take as many components as you have variables then you are just reconstructing the data and you have achieved no data reduction but you have retained 100% of the variation in your original data.
- Good components retain a high proportion of variation in your original data, we have a way to measure that.
- The first principal component is the most important, it is the best single weighted sum of your data.
- If you regress your indicators on the first PCA component, you get an R-square of 1.
- The first PCA component must share correlation with your input variables if that correlation exists.
- A second principal component picks up where the first one leaves off. If a first component explains 60% of the variation in your original data, then 40% of the variation is still remaining and a second component tries to create a second composite that eats away at that remaining 40%.
- By construction, PCA requires that each subsequent component explain more unexplained variation while also providing the best weighted sum of the indicators. Each subsequent component uses the remaining unexplained variation in a repeatable process.

To understand how PCA components are created let's start with some terms and notation. We started with a vector of unknown variables which we could extend to a full matrix of individuals with each row representing $X_1, ..., X_p$ for a sample of data. So, let's extend $\boldsymbol{X}$ into a full matrix of data where we observe all p variables on a sample of n individuals.

Now rewrite the problem as:

$$\boldsymbol{C=\alpha X}$$

Where $\boldsymbol{C}$ is a fixed number of chosen components and $\boldsymbol{\alpha}$ represents the weights we must apply to our data to get those components out. It turns out that to maximize the variance of C that we must maximize the following:

$$\boldsymbol{\alpha^T Var(X)\alpha}$$

Where $\boldsymbol{Var(X)}$ is the variance/covariance matrix of $\boldsymbol{X}$ (or it's rewritten counterpart, the correlation matrix).

From this perspective, we can note that PCA has a lot in common so far with least square regression. Both want to maximize some target, care about the appropriate use of variation, correlations impact results, create new variables that are weighted sums, and if you regress original data on PCA components you get an R-square of 1. To really see the differences between what ordinary least square regression and PCA, consider the two visuals in Figures 1 and 2 below. Least Squares Regression finds the best fitting line by looking at the squared distance from the individual points to the line. Principal Components Analysis finds a best fitting line by looking at the shortest distance from the point to the line which is also perpendicular to the raw points. Note these are NOT the same lines and come from are different frameworks.

PCA is created to be an algebraic solution to a layered problem that has a matrix representation but a solution within an existing geography. PCA is akin to regression so has the same sensitivities as regression: scale, correlation, outliers, and fit are all important considerations for PCA. Just as you get a different regression for a different dataset of the same type, you might also get a different set of PCA results depending on dataset.

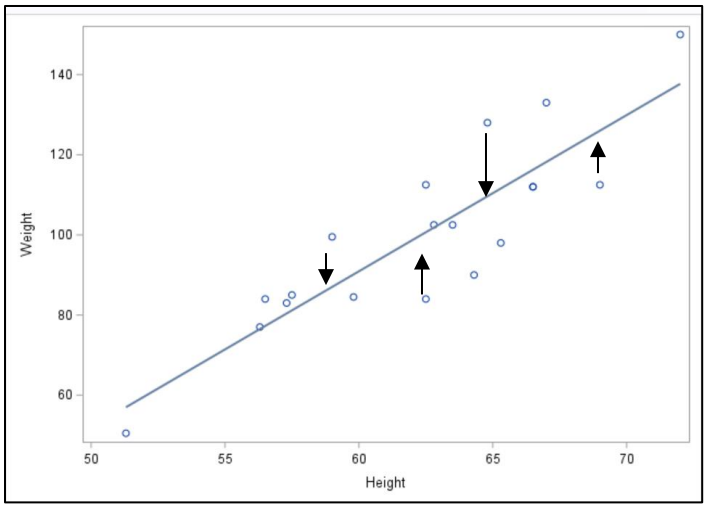

Figure 3. Example of Regression Based Framework

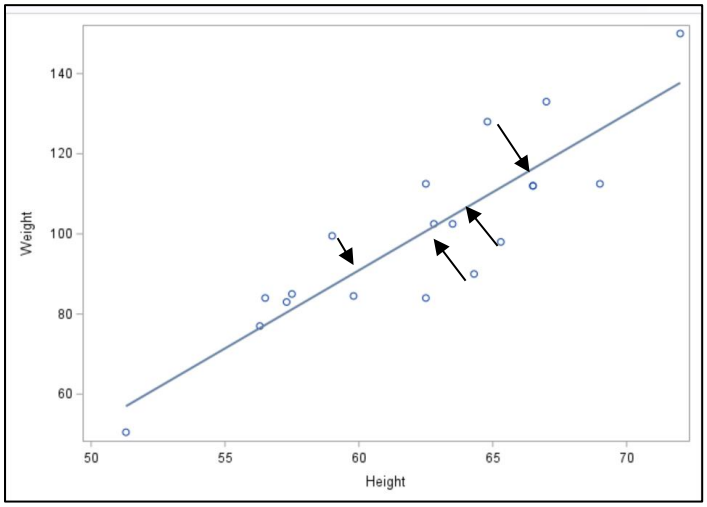

Figure 2. Example of PCA Based Framework

EXTENDING INTO FACTOR ANALYSIS

The leap from PCA to FA is not as distinct. But FA is not simply an extension of PCA. In order to go from PCA to FA we make two important considerations:

1. We assume there are underlying latent characteristics within the data and the observed variables have a causal relationship with those latent quantities. If we wanted to do a factor analysis of height/weight data we would assume that height and weight represent aspects of a dimension on 'size' and that they have some association that is measurable and predictable.
2. We believe these latent characteristics are measurable and we have to make an a priori assumption whether those characteristics are correlated or not. More on this later.

There are two types of factor analysis. Here we discuss exploratory factor analysis which is tied to PCA. An exploratory factor analysis attempts to use (or modify) a PCA to create specific components that represent the latent characteristics that are assumed to exist within the data. This is best understood via an example.

PERFORMING PCA IN SAS

Understanding PCA and FA is the hard part here. Doing PCA is actually straightforward. There are three procedures for PCA in SAS®:

1. Proc Princomp – Does basic PCA on quant data. Default gives results where variance of components is equal to the eigenvalue of that component.
2. Proc Factor – Most commonly used procedure for PCA and Factor Analysis. Default gives results where variance of components is equal to1.
3. Proc Calis – Advanced procedure for PCA, Factor Analysis, and Structural Equation Modeling.

Align these in complexity and utility like you would Proc Reg, GLM, and Mixed. Proc Calis is the most sophisticated of the bunch and will not be covered in today's workshop. We start with a toy example to help understand all that we have discussed regarding PCA. We will stick with data around height and weight and use SAS Help data for implementation. The first set of examples uses the SAS Help file 'Classfit'. Figure 3 below is a snapshot of the data. The data consists of 19 students with their age, sex, height, and weight. We will focus PCA on Height/Weight

	Name	Sex	Age	Height	Weight
1	Joyce	F	11	51.3	50.5
2	Louise	F	12	56.3	77
3	Alice	F	13	56.5	84
4	James	M	12	57.3	83
5	Thomas	M	11	57.5	85
6	John	M	12	59	99.5
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84
10	Carol	F	14	62.8	102.5
11	Henry	M	14	63.5	102.5
12	Judy	F	14	64.3	90
13	Robert	M	12	64.8	128
14	Barbara	F	13	65.3	98
15	Mary	F	15	66.5	112
16	William	M	15	66.5	112
17	Ronald	M	15	67	133
18	Alfred	M	14	69	112.5
19	Philip	M	16	72	150

Figure 3. SAS Help Classfit Dataset

All of the SAS code used to generate all the visuals in remaining document are listed in the appendix unless there is a need for a specific code callout here. We start by fitting a scatterplot to the data along with an OLS fit, those results are in Figure 4 below. We see a strong linear relationship between height and weight. We have specific students that we can easily see as 'larger' and 'smaller' (e.g. Philip and Joyce) but not everyone follows a perfect fit. How would you decide what to prioritize in a composite? PCA compares the spread of the y-axis against the spread on the x-axis and looks for compromise.

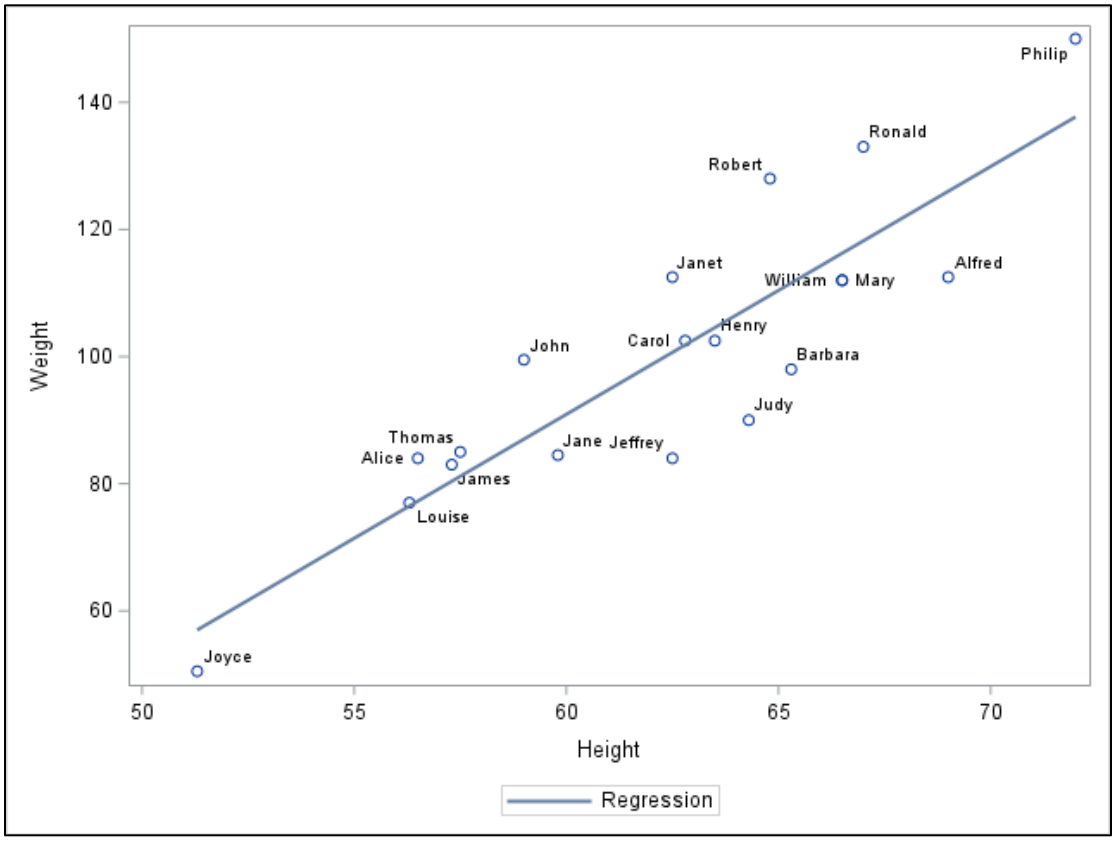

Figure 4. Scatterplot of Height Versus Weight Along with Regression Line

The first iteration of PCA will be done via Proc Princomp with example code seen in Figure 5. Note that the 'id' is not natively recognized in Proc Princomp but will execute nonetheless.

```
*Principal Components Analysis on Height, Weight;
Proc Princomp data=sample out=Output plots=score(ellipse);
var height weight;
id name;
run;
```

Figure 5. Example Proc Princomp Code

The important output from this code comes from the Eigenvector/Eigenvalue analysis (shown in Figure 6) and the visual showcase of PCA components (shown in Figure 7). Princomp gives summary statistics as well as a correlation matrix. Next, we see the Eigenvalue/Eigenvector analyses. We can create a single PCA composite from these two variables that explains 93.89% of the variance in the original data. Figure

7 is especially illustrative as we can see how individual data points contribute to the creation of PCA dimensions. Note that Joyce and Philip are still the extremes on the x-axis. A lot of what we see in these visual parallels the previous scatterplot. This is related to the geometry of PCA. The goal is to retain relationships while 'rotating' and 'rescaling' data so that differences are easier to measure. Very easy to see this in two dimensions, impossible to see in 20-30 dimensions.

The PRINCOMP Procedure

Observations	19
Variables	2

Simple Statistics

	Height	Weight
Mean	62.33684211	100.0263158
StD	5.12707525	22.7739335

Correlation Matrix

	Height	Weight
Height	1.0000	0.8778
Weight	0.8778	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.87778519	1.75557038	0.9389	0.9389
2	0.12221481		0.0611	1.0000

Figure 6. Eigenvalue Output

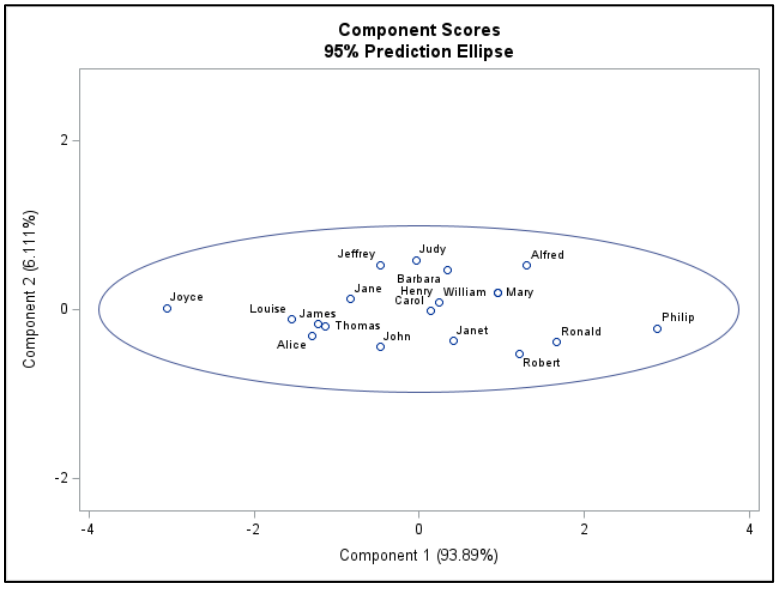

Figure 7. Visualizing PCA

Note that the appendix contains additional code for the reader to review that would standardize variables into z-scores and recalculates PCA and showcases the solutions are identical. The reason is that PCA prioritizes variance/covariance (or correlation) matrices and a correlation between two variables involves standardization. Similarly, we rant a Proc Factor on data with the exact same results. See the appendix code that includes these additional runs.

Not all output from Proc Princomp and Proc Factor are not exact though. Proc Princomp makes components whose variance are equal to eigenvalues while Proc Factor makes components that have variance 1. In cases where you are making many component variables and you need to triage the clinical importance of those variables, stick with eigenvalue-based scaling. If the components are of equal clinical importance, then standardize. The math behind PCA is not impacted by transformations but the interpretation can be. Recall that when selecting components that each component after the first builds upon the information left behind from the earlier components. There are mechanisms to get a better interpretation of multiple components by applying transformations to the data that do not change the underlying correlation matrix.

PERFORMING FA IN SAS BY EXAMPLE

The move from PCA to FA is both *clinical* and *statistical*. We have to make two critical assumptions in performing a factor analysis:

1. We make clinical assumptions that the data have one or more underlying latent characteristics and that the numbers in our data are reflective of a causal relationship between our data and those latent characteristics.
2. We make decisions on standardizations that might violate the traditional PCA assumptions. But when replaced with the above clinical assumptions still gives us something statistically valid.

A PCA assumes the components are uncorrelated and are associative (as opposed to causal).
Since FA assumes the relationships are causal in nature then we often find ourselves in conflict with the uncorrelated component assumption.
Suppose you had many physical measurements of a person including much more than height and weight. We could easily imagine 2 or 3 components that make up different dimensions of 'size'. It would be clinically unreasonable to think that those dimensions are uncorrelated. So FA can use oblique rotations to standardize the data for better interpretation at the expense of certain PCA assumptions and that tradeoff creates the need for an exploratory framework to assess fit. Therefore, FA uses a lot more output to make decisions than PCA. There are two types of 'rotations' that can be applied in this setting.

- Orthogonal rotations: standardizations that do not change the PCA assumptions or underlying correlation structure. Varimax is one of the most popular sets of rotations.
- Oblique rotations: standardizations that DO have the potential to change the underlying correlation structure. Oblique rotations move an analyst away from doing traditional PCA but can be used for interpretation. Oblique rotations are a hallmark of FA.

FA processes generally explores alternative correlation assumptions related to matrix decomposition so the idea of association is slightly perturbed. FA tends to use the terms 'communality' and 'loadings' to talk about specific types of correlations since FA is a multistage process with correlations built. Once you establish the first component then the second component (and subsequent others) build off the remaining information. Variables tend to 'load' on components to showcase how much information/correlation those variables have in specific components. A good factor analysis solution is one that explains a high proportion of initial data, reasonably distributes variables into components (called factors in a FA) and has a set of weighted summations that can be explained clinically.

EXAMPLE-BIG FIVE PERSONALITY TEST

For our use case in FA we turn to the Big Five Personality Test dataset which is described from the resourced link as:

“The Big Five personality traits, also known as the five-factor model (FFM) and the OCEAN model, is a taxonomy, or grouping, for personality traits. When factor analysis (a statistical technique) is applied to personality survey data, some words used to describe aspects of personality are often applied to the same person. For example, someone described as conscientious is more likely to be described as "always prepared" rather than "messy". This theory is based therefore on the association between words but not on neuropsychological experiments. This theory uses descriptors of common language and therefore suggests five broad dimensions commonly used to describe the human personality and psyche.”

The data contains over one million answers to this personality test, which has been developed and validated to have a 5-factor solution. In general, there is no way to know the exact number of factors a given problem can and should have. This toy example is useful in that the instrument is already designed and since it has already been validated, the best solution should be close to what investigators have already validated.

The SAS Code in the Appendix goes through a few iterations of fit. For this paper, only the final result will be shown, and the reader can run the iterations to ascertain how different options in Proc Factor contribute to the output. What we are looking for is a solution that fits what we already know clinically about the data

Start with Figure 8 below that has the basic code for creating components. Moving to Proc Factor and creating an option with rotations and prior communalities the output is labeled as ‘Factors’ instead of components, even when a basic PCA is applied in Proc Factor.

```
proc factor data=BigFive plots=scree priors=smc n=5
round flag=0.30 simple out=BigFiveScoredV rotate=varimax;
var EXT1--OPN10;
run;
```

Figure 8. Proc Factor Final Solution for Big Five Personality Test

Walking through this code step by step is critical to understanding which particular options are being used. These are the major components of the program:

- Input data is ‘*data=BigFive’* and an output data is ‘*out=BigFiveScoredV’* ; note that if you want to have actual component/factor scores AND the table of scoring coefficients, then you need to specify an output dataset.
- The round option multiplies loadings by 100 and rounds to a whole number, it flags any value greater than 0.30 for ease of interpretation.
- The ‘rotate=varimax’ performs a varimax rotation of the data. See the code for an oblique transformation version. The decision to NOT use an oblique transformation can be made after looking at the correlations between components.
- The ‘n=5’ pushes out a 5-factor solution. Note that the ‘plots=scree’ creates a scree plot of the data which helps folks review and select the number of components by visualizing inspection.
- The ‘priors=smc’ option is critical to the final solution output. Note that rotations do not impact Eigenvalue and Eigenvectors themselves. They are based on decomposing correlation matrices. However, a base assumption sometimes made is that all items used as inputs have the same basic pairwise contribution to the PCA analysis. One can make an alternative assumption that the contribution to the PCA should be related instead to the degree of pairwise correlation that exists between individual items. In those cases, one can use this option to allow for that basic correlation to drive the PCA derivation instead of a uniform or flat initial contribution by all items. Stated another way, PCA is an iterative process that selects subsequent components based on information left over from the previous Eigenvalues. The initial Eigenvalue and associate Eigenvector come from a set up where the initial assumption is that all items may contribute

equally OR an assumption that items will contribute according to the strength of their pairwise correlation. The latter assumption uses the Square Multiple Correlations (SMC) as prior communality estimates.

Figure 9 below has the Eigenvalues for an initial PCA with SMC as an initial assumption. Take care as the rescaling of input correlations has an impact on measuring percent variance explained. We do say that a 5-factor solution here does explain 95% of the variance in the original data. But we would not say that a 6-factor solution explains 102% of the variance.

Eigenvalues of the Reduced Correlation Matrix: Total = 20.5511721 Average = 0.41102344

	Eigenvalue	Difference	Proportion	Cumulative
1	6.65111849	2.18693947	0.3236	0.3236
2	4.46417902	1.10886438	0.2172	0.5409
3	3.35531463	0.45041849	0.1633	0.7041
4	2.90489614	0.71431440	0.1413	0.8455
5	2.19058174	0.79359592	0.1066	0.9521
6	1.39698582	0.67392323	0.0680	1.0200
7	0.72306259	0.25206882	0.0352	1.0552
8	0.47099377	0.15856722	0.0229	1.0781

Figure 9. Proc Factor Eigenvalue Output

So how can we be sure that this is the 'final' solution? Figures 10-14 contribute to this discussion by showing the Varimax rotated factor loadings from the data. We see that the items on each scoring domain load uniquely on one of the 5 factors. This solution aligns factor loadings with clinical insights and suggests there is a path forward where individual items most commonly load onto a specific factor. Low potential for cross-loadings. And (as a teaching example) we see our solution lines up with coding. Note that some of the loadings are especially weak but the pattern shows consistency and that each item has the most association on a single factor.

Rotated Factor Pattern

		Factor1		Factor2	Factor3	Factor4	Factor5
EXT1	EXT1-I am the life of the party	69	*	0	7	5	3
EXT2	EXT2-I dont talk a lot	-68	*	3	-12	0	6
EXT3	EXT3-I feel comfortable around people	63	*	-19	25	3	14
EXT4	EXT4-I keep in the background	-71	*	16	-4	5	2
EXT5	EXT5-I start conversations	69	*	-3	21	11	12
EXT6	EXT6-I have little to say	-53	*	13	-13	-19	3
EXT7	EXT7-I talk to a lot of different people at parties	70	*	-4	16	6	7
EXT8	EXT8-I dont like to draw attention to myself	-57	*	8	6	1	11
EXT9	EXT9-I dont mind being the center of attention	63	*	0	-2	17	0
EXT10	EXT10-I am quiet around strangers	-65	*	21	-5	4	2

Figure 10. Factor 1 Loadings

EST1	EST1-I get stressed out easily	-13		67	*	11		-7		1	
EST2	EST2-I am relaxed most of the time	12		-43	*	0		11		2	
EST3	EST3-I worry about things	-15		61	*	20		3		7	
EST4	EST4-I seldom feel blue	14		-28		-3		-1		15	
EST5	EST5-I am easily disturbed	-4		53	*	0		-7		-4	
EST6	EST6-I get upset easily	-4		72	*	3		-7		-3	
EST7	EST7-I change my mood a lot	1		73	*	-2		1		-12	
EST8	EST8-I have frequent mood swings	0		75	*	-4		0		-13	
EST9	EST9-I get irritated easily	-3		69	*	-16		-2		0	
EST10	EST10-I often feel blue	-25		60	*	-1		10		-17	

Figure 11. Factor 2 Loadings

AGR1	AGR1-I feel little concern for others	-1		8		-47	*	-4		2	
AGR2	AGR2-I am interested in people	33	*	0		53	*	13		3	
AGR3	AGR3-I insult people	12		26		-39	*	9		-15	
AGR4	AGR4-I sympathize with others feelings	4		12		75	*	7		7	
AGR5	AGR5-I am not interested in other peoples problems	-13		6		-63	*	3		5	
AGR6	AGR6-I have a soft heart	-1		21		57	*	0		7	
AGR7	AGR7-I am not really interested in others	-29		14		-61	*	0		4	
AGR8	AGR8-I take time out for others	14		5		54	*	9		14	
AGR9	AGR9-I feel others emotions	9		17		68	*	10		9	
AGR10	AGR10-I make people feel at ease	30		-5		39	*	15		18	

Figure 12. Factor 3 Loadings

CSN1	CSN1-I am always prepared	2		-8		3		10		61	*
CSN2	CSN2-I leave my belongings around	7		17		4		17		-51	*
CSN3	CSN3-I pay attention to details	-4		6		10		28		42	*
CSN4	CSN4-I make a mess of things	-3		41	*	-3		5		-51	*
CSN5	CSN5-I get chores done right away	7		-7		5		-5		62	*
CSN6	CSN6-I often forget to put things back in their proper place	3		25		1		11		-54	*
CSN7	CSN7-I like order	-5		11		5		8		56	*
CSN8	CSN8-I shirk my duties	-3		28		-12		1		-40	*
CSN9	CSN9-I follow a schedule	5		4		11		-2		62	*
CSN10	CSN10-I am exacting in my work	3		2		6		28		46	*

Figure 13. Factor 4 Loadings

OPN1	OPN1-I have a rich vocabulary	3		-2		-3		60	*	5	
OPN2	OPN2-I have difficulty understanding abstract ideas	0		26		-2		-51	*	6	
OPN3	OPN3-I have a vivid imagination	3		14		10		56	*	-6	
OPN4	OPN4-I am not interested in abstract ideas	3		17		-11		-44	*	14	
OPN5	OPN5-I have excellent ideas	20		-3		-1		59	*	17	
OPN6	OPN6-I do not have a good imagination	-6		11		-9		-44	*	9	
OPN7	OPN7-I am quick to understand things	7		-8		0		50	*	21	
OPN8	OPN8-I use difficult words	3		8		-10		58	*	-1	
OPN9	OPN9-I spend time reflecting on things	-13		16		19		41	*	8	
OPN10	OPN10-I am full of ideas	18		3		5		67	*	5	

Figure 14. Factor 5 Loadings

Considering all of Figures 11-14 together, the pattern emerges that these items fit together well to create a single set of factor scores that represent 5 latent dimensions of personality types. Classical Testing Theory encourages a clinical practice where once a 'good' Factor Analysis solution is found, the clinician uses a pseudo-version of it for scoring that is 'clean' in that the algorithm is well defined and items with poor loadings are removed from the weighted summation of composites. That is to say that our 5 factor Varimax solution with SMCs as prior communality estimates creates something so close to a solution where each 10 items loads onto a unique factor and we should just take a version of those 10 factors as our clinical score for that domain/trait.

To close the loop on the discussion around Factor Analysis, consider a confirmatory factor analysis (CFA). CFA starts by assuming that there are one or more latent characteristics in your data and that you put forth a hypothesized relationship. CFA then determines whether that hypothesized relationship is based on tenable assumptions by exploring how well the data fit to your model. It's akin to regression by suggesting a regression model for your data and then seeing whether your data significantly deviate from your chosen regression model. It's about measuring goodness of fit and consistency. You are confirming what someone else has already hypothesized about the relationships in your data.

You can do a PCA on just correlation matrices, which is one way to do a PCA with missing or incomplete data. You can do a PCA on all binary data but you run into some statistical issues. You can use polychoric or tetrachoric correlations to measure correlation between ordinal or binary variables and do PCA on those values. You should avoid PCA for correlated data. PCA assumes that all of your variation is measured as variance, covariance, and correlation. If you have geospatial correlation or temporal correlation or familial correlations then PCA can't account for that in base work. So avoid using THIS version on genomic work or geospatial data or time series data.

CONCLUSION

PCA and FA are powerful and useful techniques for reducing data that work by creating powerful and useful linear combinations that retain a large amount of variation from the input data. They are useful tools in any analysts toolkit as they do not require an inferential or statistical framework to begin implementation. Though not dependent on statistical inference, there are other aspects that one must consider in doing a successful PCA and/or FA. For example, in consideration of sample size – there are few statistical power discussions here so experts do not fully agree. Some have said a couple of hundred to start with while others have provided 'rules of thumb' like a 5 or 10 observations per variable under consideration.

When to stop? Hard to tell. Many folks explore multiple solutions and rotations in a FA until they get one that has a clinical interpretation that is useful for discussion. This can sometimes make FA results hard to validate in subsequent studies. Reproducibility is an issue and should be part of the overall discussion.

This presentation is only designed to scratch the surface and use example based training to understand how these methods perform with actual data in SAS. The readers should see Hatcher and O'Rourke (2013) for more on implementing PCA and FA in SAS and see Joliffe (2002) for a more complete understanding of PCA.

REFERENCES

Goldberg, Lewis R. "The development of markers for the Big-Five factor structure." Psychological assessment 4.1 (1992): 26.

Hatcher, Larry, and Norm O'Rourke. A step-by-step approach to using SAS for factor analysis and structural equation modeling. SAS Institute, 2013.

Jolliffe, Ian T. Principal component analysis for special types of data. Springer New York, 2002.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jason Brinkley, PhD
Principal Data Scientist
Abt Associates
Jason_Brinkley@abtassoc.com
https://twitter.com/DrJasonBrinkley

APPENDIX – SAS CODE FOR WORKSHOP

```
*Example 1 - SAS Help Class Data;

*load the data;
Data Sample;
set sashelp.classfit;
run;

*Simple Scatterplot of Height versus Weight;
Proc sgplot;
reg x=height y=weight/ datalabel=name;
run;

*Principal Components Analysis on Height, Weight;
Proc Princomp data=sample out=Output plots=score(ellipse);
var height weight;
id name;
run;

*Standardize Data;
Data Sample2;
set output;
z_height = (height - 62.33684211)/5.12707525;
z_weight = (weight - 100.0263158)/22.7739335;
Original_Prin = Prin1;
drop prin1 prin2;
run;

*Standardized PCA;
Proc Princomp data=sample2 out=Output plots=score(ellipse);
var z_height z_weight;
id name;
run;

*Proc Factor;
proc factor data=sample2 nfactors=1 out=output2;
var height weight;
run;

*Compare Principal Components, differenced due to scaling/standardization;
proc sgplot;
scatter x=original_prin y=factor1;
run;

proc corr;
var original_prin factor1;
run;

*Example 2 - The Big Five Personality Test;
*Data available at https://wwwkagglecom/datasets/tunguz/big-five-personality-
test/data;
*You must get rid of the 'NA' variables before trying to load or variables don't
come in as correct type;
*Data was imported into Microsoft Excel and resaved due to read-in issues See
weblink for details;

Proc Import datafile = "INSERT YOUR FILE PATH HERE”
Out=BigFive dbms = csv replace;
getnames=yes;
run;
```

```
Proc Import datafile = "FOLDERPATH\Factor Analysis Example Data.csv"
Out=BigFive dbms = csv replace;
getnames=yes;
run;

*labels;
Data BigFive;
set BigFive;
label EXT1 = 'EXT1-I am the life of the party';
label EXT2 = 'EXT2-I dont talk a lot';
label EXT3 = 'EXT3-I feel comfortable around people ';
label EXT4 = 'EXT4-I keep in the background ';
label EXT5 = 'EXT5-I start conversations ';
label EXT6 = 'EXT6-I have little to say ';
label EXT7 = 'EXT7-I talk to a lot of different people at parties ';
label EXT8 = 'EXT8-I dont like to draw attention to myself ';
label EXT9 = 'EXT9-I dont mind being the center of attention ';
label EXT10 = 'EXT10-I am quiet around strangers ';
label EST1 = 'EST1-I get stressed out easily ';
label EST2 = 'EST2-I am relaxed most of the time ';
label EST3 = 'EST3-I worry about things ';
label EST4 = 'EST4-I seldom feel blue ';
label EST5 = 'EST5-I am easily disturbed ';
label EST6 = 'EST6-I get upset easily ';
label EST7 = 'EST7-I change my mood a lot ';
label EST8 = 'EST8-I have frequent mood swings ';
label EST9 = 'EST9-I get irritated easily ';
label EST10 = 'EST10-I often feel blue ';
label AGR1 = 'AGR1-I feel little concern for others ';
label AGR2 = 'AGR2-I am interested in people ';
label AGR3 = 'AGR3-I insult people ';
label AGR4 = 'AGR4-I sympathize with others feelings ';
label AGR5 = 'AGR5-I am not interested in other peoples problems ';
label AGR6 = 'AGR6-I have a soft heart ';
label AGR7 = 'AGR7-I am not really interested in others ';
label AGR8 = 'AGR8-I take time out for others ';
label AGR9 = 'AGR9-I feel others emotions ';
label AGR10 = 'AGR10-I make people feel at ease ';
label CSN1 = 'CSN1-I am always prepared ';
label CSN2 = 'CSN2-I leave my belongings around ';
label CSN3 = 'CSN3-I pay attention to details ';
label CSN4 = 'CSN4-I make a mess of things ';
label CSN5 = 'CSN5-I get chores done right away ';
label CSN6 = 'CSN6-I often forget to put things back in their proper place ';
label CSN7 = 'CSN7-I like order ';
label CSN8 = 'CSN8-I shirk my duties ';
label CSN9 = 'CSN9-I follow a schedule ';
label CSN10 = 'CSN10-I am exacting in my work ';
label OPN1 = 'OPN1-I have a rich vocabulary ';
label OPN2 = 'OPN2-I have difficulty understanding abstract ideas ';
label OPN3 = 'OPN3-I have a vivid imagination ';
label OPN4 = 'OPN4-I am not interested in abstract ideas ';
label OPN5 = 'OPN5-I have excellent ideas ';
label OPN6 = 'OPN6-I do not have a good imagination ';
label OPN7 = 'OPN7-I am quick to understand things ';
label OPN8 = 'OPN8-I use difficult words ';
label OPN9 = 'OPN9-I spend time reflecting on things ';
label OPN10 = 'OPN10-I am full of ideas ';
run;

proc factor data=BigFive;
var EXT1--OPN10;
run;
```

```
proc factor data=BigFive plots=scree n=6
round flag=0.30 simple out=BigFiveScored;
var EXT1--OPN10;
run;


proc factor data=BigFive plots=scree n=6
round flag=0.30 simple out=BigFiveScoredV rotate=varimax;
var EXT1--OPN10;
run;

proc factor data=BigFive plots=scree n=6
round flag=0.30 simple out=BigFiveScoredP rotate=promax;
var EXT1--OPN10;
run;

proc factor data=BigFive plots=scree priors=smc n=5
round flag=0.30 simple out=BigFiveScoredV rotate=varimax;
var EXT1--OPN10;
run;
```