Pedagogical Preface

I want to explain how I like to develop and present statistical concepts and principles. I want you to understand the basic concepts and principles clearly and intuitively. It is much more important to understand statistics intuitively than to memorize formulas. If you know the intuition, the formulas will make sense. If you don't know the intuition, the formulas will not make sense. Besides, with the steady increase in computer power, we have put most of the formulas out of sight inside computer software. So you don't *have* to memorize formulas.

Experience shows that concepts and principles are most readily mastered in a two-step process:

1. First, begin with one or more very simple examples that have been stripped of complicated real-world context so that the concept or principle can be seen in pure form. These simple examples may not be very practical. But once the concept/principle has been grasped, then it will be much easier to apply to the real-world problems that we are actually interested in. I will often begin a topic by discussing a simulated data set. I can set up a simulation to reflect the concept or principle that I want to illustrate in pure form.
2. The second step is to present one or more real-world examples with more complicated contexts in which the basic concept/principle can now be recognized. Developing competence at statistics requires developing the ability to strip away the interesting but distracting real-world aspects of a problem to expose the underlying statistical principles. If you try to learn the concepts in complicated contexts, you can get distracted and confused by the interesting real-world detail and fail to recognize the underlying principles.

Principal Components Analysis
Part 1
Introduction

The objective of Principal Components Analysis (PCA) is to uncover structure in a dataset by replacing the old variables with new variables that have more desirable properties. The new variables are linear combinations of the old variables.

What are these more desirable properties?

- Information…………... The new variables capture most of the information (variability, explanatory power) of the old variables.
- Dimensionality reduction… There are fewer new variables than old variables (parsimony).
- Uncorrelated………….. The new variables are uncorrelated with each other.
- Insight………………… The new variables may reveal hidden structure and unsuspected interpretations.

A potential killer disadvantage: It may be hard to understand what the new variables mean.

So imagine a dataset arranged in a spreadsheet of *n* rows (observations) and *p* columns (variables). PCA seeks to find structure in the columns (variables), rather than in the rows (observations). Other techniques, like cluster analysis and sufficiency analysis, seek to find structure and information in the rows. In PCA, there is no variable in the dataset than can be singled out for attention as a *Y* or dependent variable that is to be estimated or predicted. So the "information" or "explanatory power" mentioned above as one of the desirable outcomes of PCA is not the same as the explanatory power of regression, which is commonly assessed by R-square, which measures how much of *Y* is explained by the model. The "explanatory power" of PCA is different, although related, and is assessed differently. As a preview, we will measure the total variability of the dataset and ask what proportion of the total variability is accounted for by the individual principal components. The components that account for the greatest portion of the total variability contain the most information. They have the most explanatory power. They are the most important and the most worth using. This parallels regression, in which the most important predictor variables are those that account for the greatest proportion of the variability of *Y*. In PCA there is no *Y*. So we look for the components that account for the greatest proportion of the inherent variability of the whole dataset, appropriately measured.

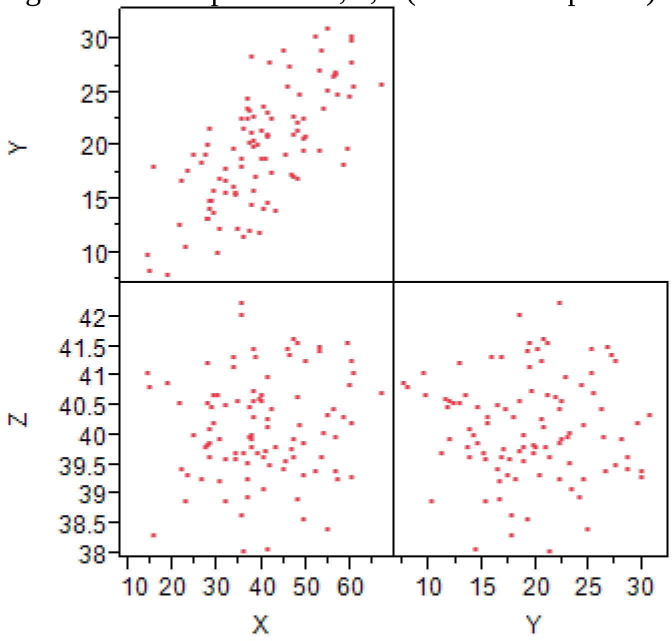
One common application of PCA is to simplify the predictor variables in a regression. In this application, one seeks to replace the *X* variables in the regression by a smaller set of their own principal components in order to profit from the desirable properties listed above. This procedure is called *principal components regression*. However, the *Y* variable in that regression plays no role whatsoever in the PCA to be done on the *X* variables. The *Y* variable is not used in the PCA. You could replace the *Y* in the regression with any other variable – even random numbers – and the PCA of the *X* variables would remain exactly the same. That is, PCA is an unsupervised statistical learning technique. *Unsupervised* means that there is no *Y* variable to direct or guide the analysis.

Rotations

In order to understand principal components, it is first necessary to understand rotations. Here is an example with simulated data on three variables (X, Y, Z) [n = 100]. [1]

Scatterplots of the three variables plotted against each other do not seem to show much going on:

Figure 1. Scatterplots of X,Y,Z (n=100 data points)



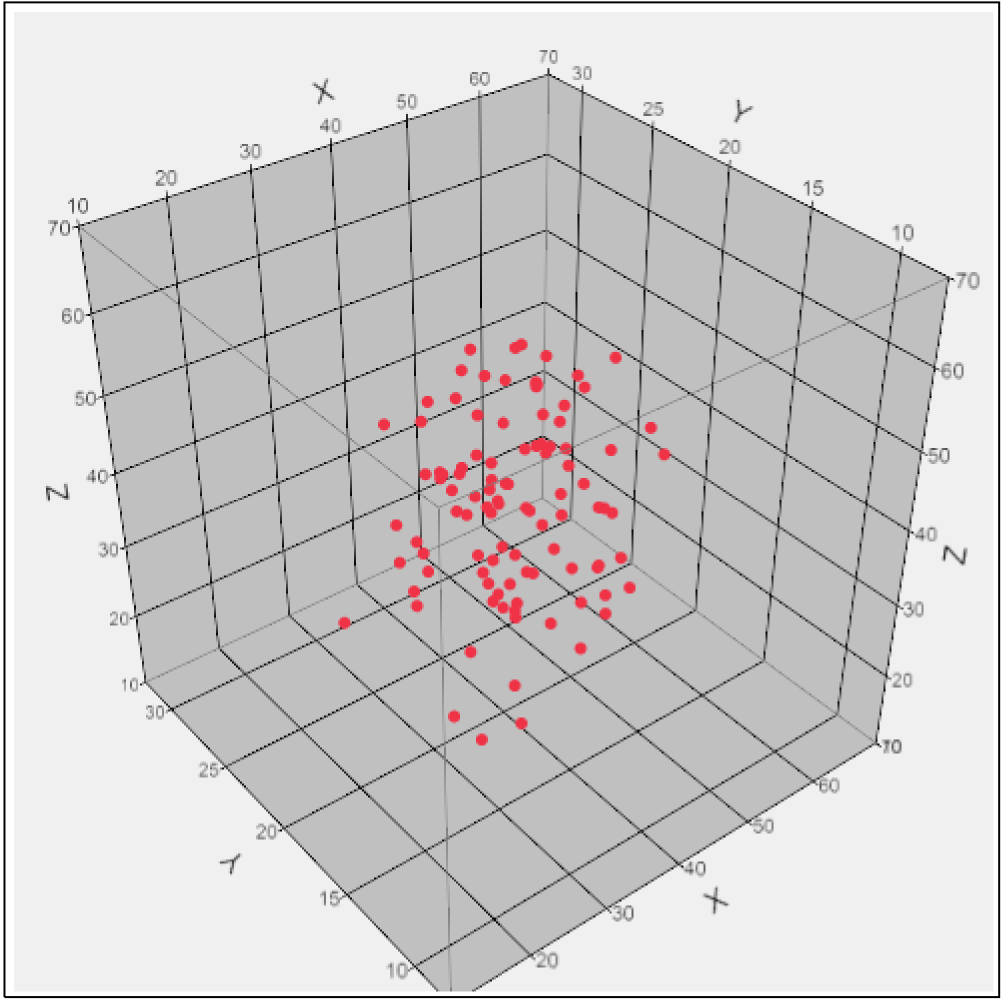
[Plots produced by Scatterplot Matrix feature of JMP.]

There is some positive relationship between X and Y (top left), but no discernible relationship between any other pair of variables (bottom right and left).

[1] The actual data may be found in the Excel file *PCA XYZ UVW ABC data and prin comps.xlsx*.

A 3D view also does not show much structure:

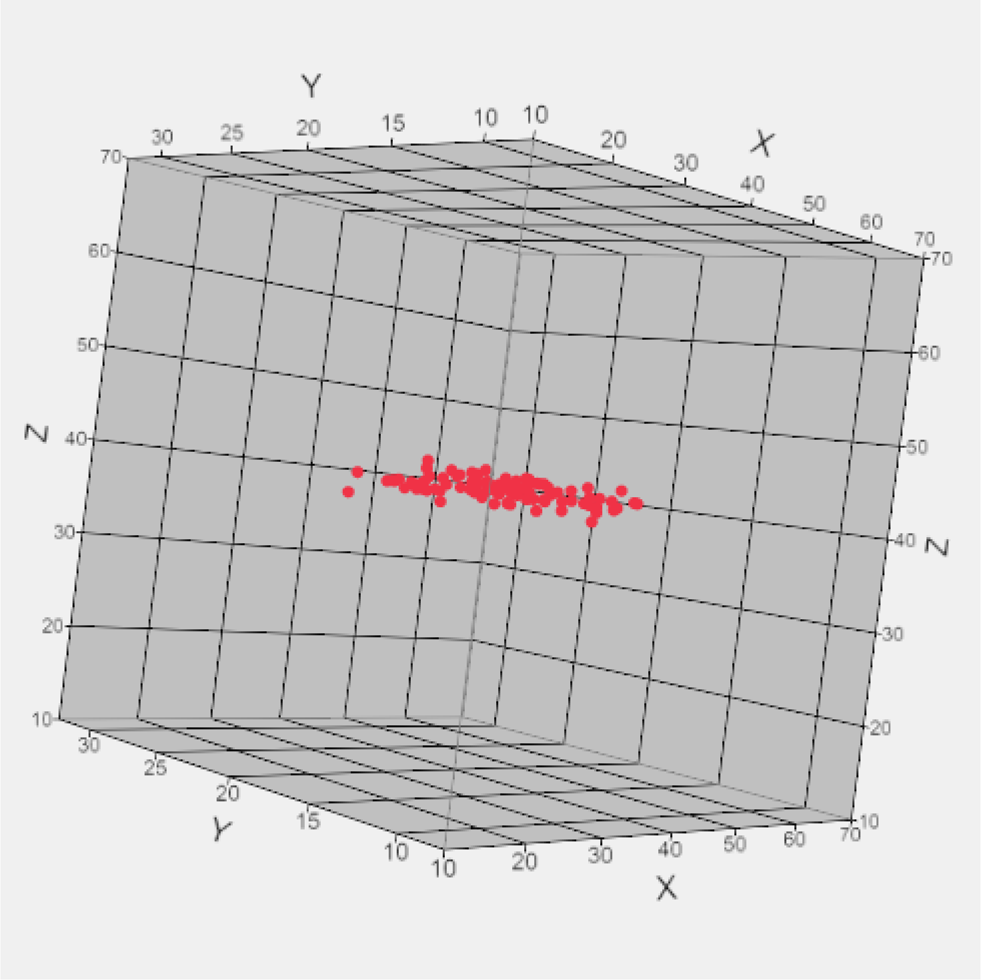
Figure 2. 3D scatterplot of X,Y,Z (*n*=100 data points)



[Plot produced by Scatterplot 3D feature of JMP.]

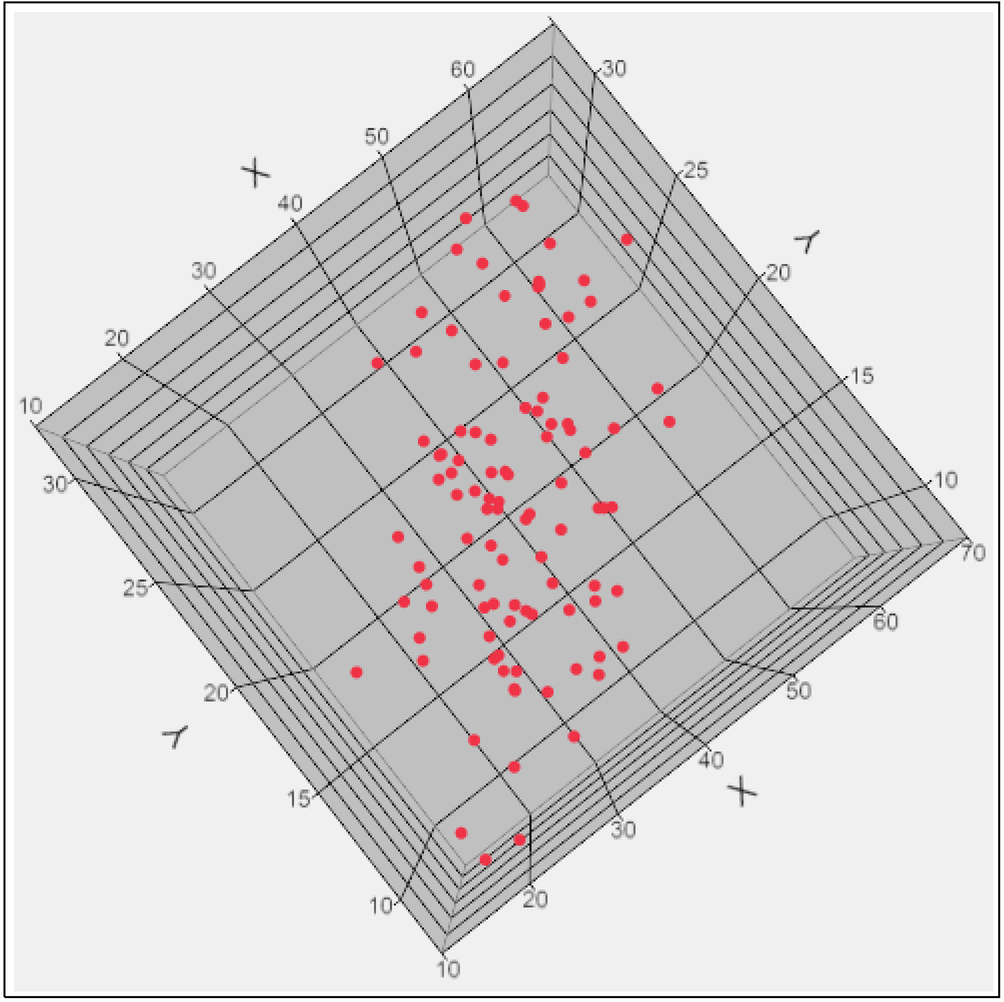
But now, let us move our vantage point so that we view the 3D plot from a different perspective. Let us move down the edge closest to us until we are looking straight into the data box, with the Z walls at the back and the two X-by-Y planes at top and bottom. The following graph results. Some very interesting structure is revealed:

Figure 3. 3D scatterplot of X,Y,Z from a different perspective.



From this perspective, we see easily that Z is essentially a constant, with all of its values concentrated around 40. We are looking edgewise at a disk. All of the interesting action in these data is going on between X and Y within that disk. We can see that by moving our vantage point yet again, until our vantage point is located at the top of the preceding graph and looking down onto the flattened disk:

Figure 4. 3D scatterplot of X,Y,Z from yet another perspective.

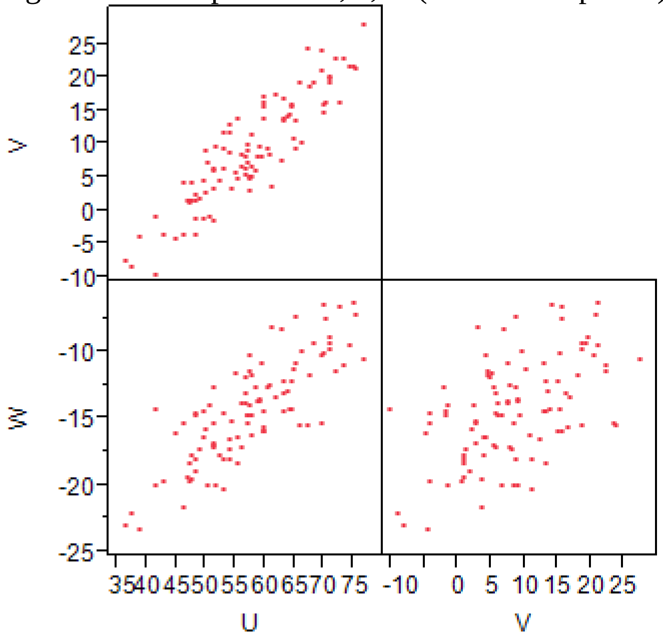


This perspective is like the top left of the initial scatterplot matrix in Figure 1. It shows that Y is positively related to X, with values of Y tending to increase with increasing values of X. But Z does not participate in the relationship because Z is essentially constant. In other words, these three-dimensional data are essentially two-dimensional because Z is effectively constant. What appears initially to be a three-dimensional relationship can be reduced to a two-dimensional relationship – Z is basically unnecessary. In arriving at this conclusion, we did not change the data at all. We merely moved our vantage point to a different perspective. We also see that some perspectives are more valuable for gaining insight than other perspectives. We should try to find interesting perspectives on our data.

Keep the above lessons in mind, and consider a different dataset, consisting of three variables (U,V,W) [n = 100].[2] Here is a scatterplot matrix:

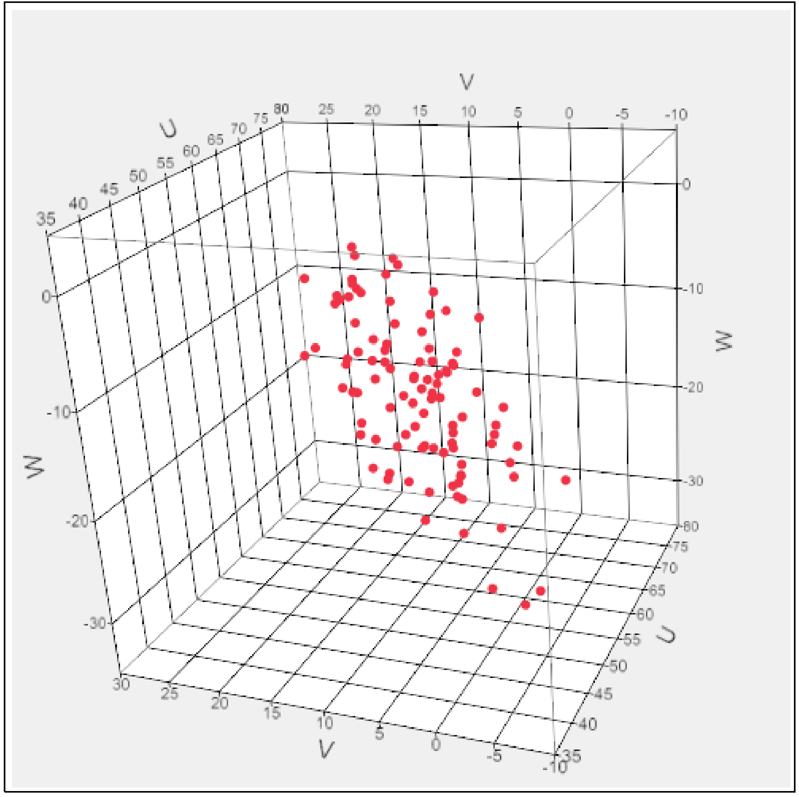
[2] The actual data may be found in the Excel file *PCA XYZ UVW ABC data and prin comps.xlsx*.

Figure 5. Scatterplots of U,V,W (*n*=100 data points)



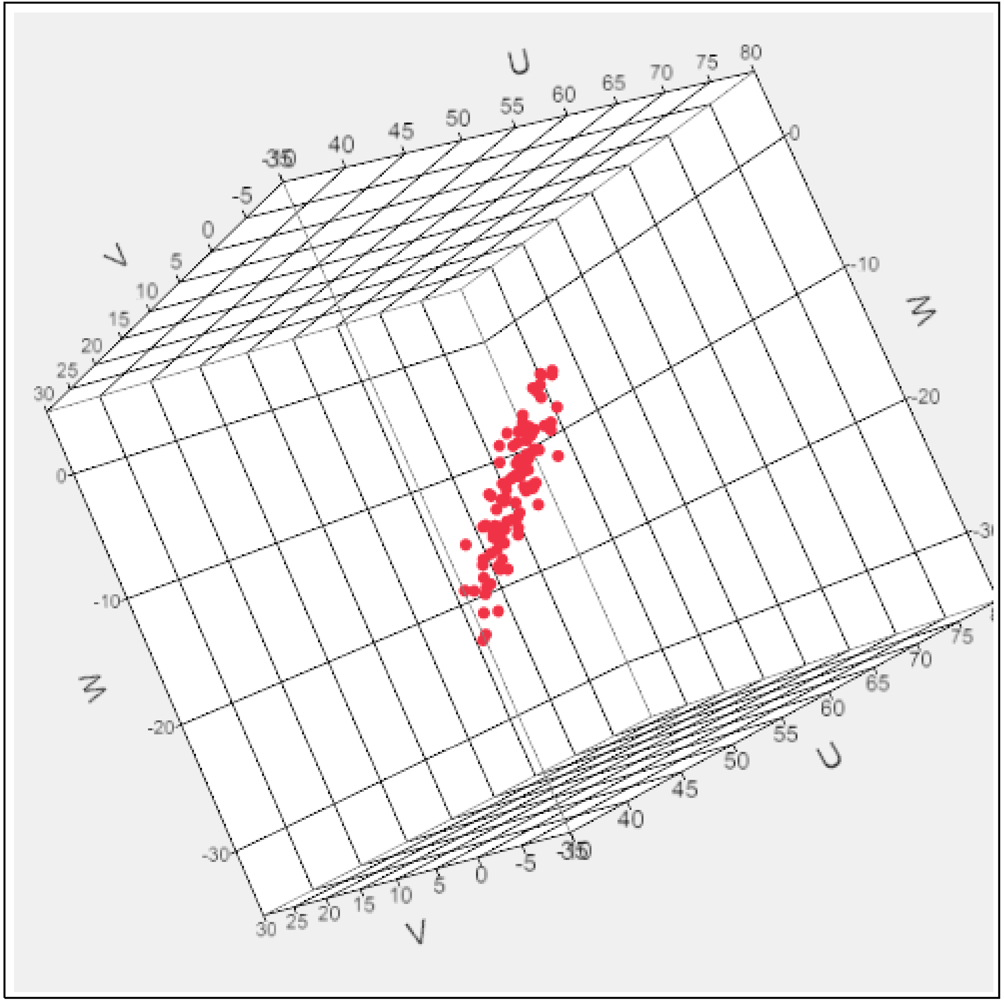
It looks like there may be some interesting things going on. Positive relationships of different strengths appear among all three of the variables. And here is a 3D view:

Figure 6. 3D scatterplot of U,V,W (*n*=100 data points)



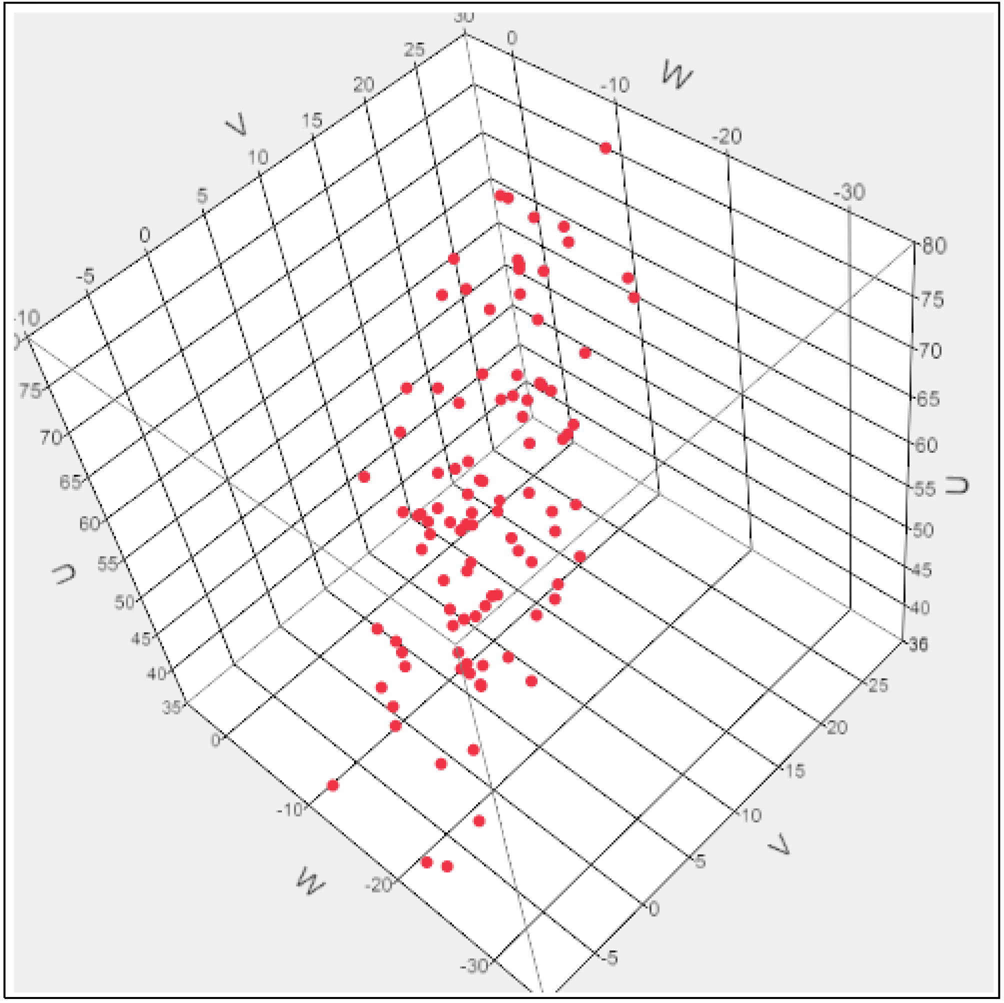
Now let us once again move our vantage point in the preceding plot – to the right and up a bit until we obtain the following perspective:

Figure 7. 3D scatterplot of U,V,W from a different perspective.



Once again we see a flattened disk of points. But the disk does not lie along a plane parallel to one of the walls. Still, the only real action is going on inside the disk, as we can see by shifting our vantage point from edge-on to directly above the disk and looking down onto the disk: (This view is not directly above the cube, but above the disk, since the disk is angled toward – not parallel to – the walls of the cube.)

Figure 8. 3D scatterplot of U,V,W from yet another perspective.



After these manipulations, we have learned that the new dataset of U,V,W variables shares similar features with the first dataset of X,Y,Z variables, in that both are flattened disks of points, with all of the important action occurring in the two-dimensional subspace of the disk. However, unlike the X,Y,Z dataset, in which it is easy to see that there is an "unnecessary" variable, Z, there is no unnecessary variable among U,V,W. A reduction of dimension is achievable with the X,Y,Z dataset – Z is essentially constant, so just drop Z. However, there is no readily apparent way to reduce the dimension of the U,V,W dataset – none of U,V,W is constant. Although the two data sets share similar structure in being confined to two-dimensional disks, we cannot simplify the U,V,W data by dropping one of the variables as we can simplify the X,Y,Z data by dropping Z.

Now I will reveal to you a startling fact that will heighten the injustice of the preceding sentence: The U,V,W dataset is the *same* data as the X,Y,Z dataset! It has just had a change of coordinates. Please look again now at Figure 2. Keep the red data points fixed where they are and maintain your viewing vantage point. But imagine uncoupling the background coordinate grid from the data points so that the grid can swivel in the background without moving the points. (However, keep the grid attached at the origin (0,0,0).) Rotate the grid in any direction, so that the axes maintain 90-degree angles with each other. Have the points changed their relationships with each other or with you? Not at all! Each point remains exactly the same distance from every other point and from you. The geometric relationships among the points are unaffected; their relationships to you, the viewer, are the same because you and the points both remain stationary. But the distances of the points from the swiveling grid walls change. That is what happened to X,Y,Z to get U,V,W. The geometry of the points remained the same, but the background coordinate grid swiveled to a new location. Then the distances of the points from the walls of the grid were re-measured, resulting in the new U,V,W coordinates. Since the geometry of the points is the same, you should be able to see the same point patterns if you compare the 3D graphs of U,V,W with the 3D graphs of X,Y,Z – especially in Figures 4 and 8, which have about the same perpendicular perspectives from above the planes of the disks.[3]

Now look at Figure 2 again and compare it with Figure 3. To get from Figure 2 to Figure 3, we kept the coordinate grid and the points where they were but moved our viewing vantage point down the edge closest to us. As an alternative, what if we had remained where we were, but had rotated the coordinate system together with the points up toward us? You should be able to see that if we are careful, we can get the same 3D view of the data by either method. Our eyes will see the red dots in the same final positions. We can leave the coordinate system and points alone, but move our vantage point; or we can shift the coordinate system and points, but remain where we are. The same view results.

These ruminations imply that there really *should be* a superfluous dimension in the U,V,W data. But what is it? The flattened disk in Figure 7 is the same flattened disk in Figure 3. The equation of the plane of the disk in Figure 3 is Z = 40. But the plane of the disk in Figure 7 does not have such a simple equation in U,V,W coordinates. If we can figure out how to express Z in terms of U,V,W, then we can identify the superfluous dimension. It turns out that $Z = \frac{1}{\sqrt{3}}U - \frac{1}{\sqrt{6}}V - \frac{1}{\sqrt{2}}W$. Since the disks of Figures 3 and 7 are the same and since Z = 40 is the equation of the plane of the disk in Figure 3, therefore $\frac{1}{\sqrt{3}}U - \frac{1}{\sqrt{6}}V - \frac{1}{\sqrt{2}}W = 40$ is the equation of the plane of the disk in Figure 7. For emphasis, let me repeat this: $\frac{1}{\sqrt{3}}U - \frac{1}{\sqrt{6}}V - \frac{1}{\sqrt{2}}W = 40$ is the same *locus* of points as Z = 40. This is because $Z = \frac{1}{\sqrt{3}}U - \frac{1}{\sqrt{6}}V - \frac{1}{\sqrt{2}}W$ for *every* point in 3D space.

So the superfluous dimension is $\frac{1}{\sqrt{3}}U - \frac{1}{\sqrt{6}}V - \frac{1}{\sqrt{2}}W$, expressed in U,V,W coordinates. This is a strange looking dimension! It is not a single variable. Moreover, it is hidden in the U,V,W data – it is not obvious. If you were given the X,Y,Z data, you could find that Z is superfluous by rotating the data as I did until they align in a flattened disk at Z=40, as in Figure 3. Moreover, most of the values in column Z are approximately 40. But if you were given the U,V,W data

[3] The point pattern in Figure 4 does not appear exactly the same as the point pattern in Figure 8 because of my inability to adjust the perspectives manually so that they coincide exactly.

only, you could rotate them and see a flattened disk, but how could you find the equation of the plane of that disk?

Moreover, if $\frac{1}{\sqrt{3}}U-\frac{1}{\sqrt{6}}V-\frac{1}{\sqrt{2}}W$ is a superfluous dimension, what are the *non*-superfluous dimensions? In Figure 3, the non-superfluous dimensions are clearly X and Y. It should be so likewise in Figure 7. But X and Y have different expressions in U,V,W coordinates. It turns out that their expressions are $X=\frac{1}{\sqrt{3}}U+\frac{2}{\sqrt{6}}V$ and $Y=\frac{1}{\sqrt{3}}U-\frac{1}{\sqrt{6}}V+\frac{1}{\sqrt{2}}W$ in the new coordinates.

How do I know these things? Because I *defined* the equations to convert X,Y,Z into U,V,W! I set up the following equations to calculate the U,V,W data from the original X,Y,Z data:

$$\left\{\begin{array}{l} U=\frac{1}{\sqrt{3}}X+\frac{1}{\sqrt{3}}Y+\frac{1}{\sqrt{3}}Z \\ V=\frac{2}{\sqrt{6}}X-\frac{1}{\sqrt{6}}Y-\frac{1}{\sqrt{6}}Z \\ W=\qquad\quad\;\; \frac{1}{\sqrt{2}}Y-\frac{1}{\sqrt{2}}Z \end{array}\right\}\text{[Eqn 1]}$$

To get any U,V,W point, plug the X,Y,Z coordinates into Eqn 1.[4] If you solve these equations simultaneously for X,Y,Z, you obtain

$$\left\{\begin{array}{l} X=\frac{1}{\sqrt{3}}U+\frac{2}{\sqrt{6}}V \\ Y=\frac{1}{\sqrt{3}}U-\frac{1}{\sqrt{6}}V+\frac{1}{\sqrt{2}}W \\ Z=\frac{1}{\sqrt{3}}U-\frac{1}{\sqrt{6}}V-\frac{1}{\sqrt{2}}W \end{array}\right\}\text{[Eqn 2]}$$

The third line of Eqn 2 yields the expression for Z = 40. The superfluous dimension is $\frac{1}{\sqrt{3}}U-\frac{1}{\sqrt{6}}V-\frac{1}{\sqrt{2}}W$ (i.e., Z). The non-superfluous dimensions are $\frac{1}{\sqrt{3}}U+\frac{2}{\sqrt{6}}V$ (i.e., X) and $\frac{1}{\sqrt{3}}U-\frac{1}{\sqrt{6}}V+\frac{1}{\sqrt{2}}W$ (i.e., Y).

You can use these two sets of equations to change X,Y,Z coordinates into U,V,W coordinates [Eqn 1] and vice-versa [Eqn 2]. Application of Eqn 1 to the X,Y,Z data changes our perspective on the data points from that of Figure 2 to that of Figure 6. Therefore, Eqn 1 amounts to a rotation of coordinate systems. If you then apply Eqn 2 to the U,V,W data, they are transformed back into the X,Y,Z data – which amounts to changing our perspective on the data from that of Figure 6 back to that of Figure 2. That is, Eqn 2 also amounts to a rotation of coordinate systems. But Eqn 2 is the inverse rotation to that of Eqn 1. A big take-away from this is that systems of linear equations are equivalent to a rotation of coordinate systems, which are equivalent to a

[4] You may try this with the data in the Excel file *PCA (X,Y,Z) (U,V,W) and manipulations.xlsx*.

change of perspective. By choosing interesting rotations, we choose interesting perspectives on the data. And those interesting perspectives can be approached mathematically through systems of equations.

But not just any systems of equations! We want to preserve the geometry of the data points as we rotate the coordinate systems. We want the distances between data points and the angles between data points to remain the same. If we are successful in preserving distances and angles, then any statistical inferences that depend upon distances and angles would be equally good in either coordinate system. So to do our statistical analysis, we could choose a convenient rotation in which the inferences are easier than in the original coordinates and be confident that our analysis still applies to the original data.

So how do we know if the rotation of Eqn 1 preserves the geometry of the data? It turns out that the geometry is preserved if the equations of rotation (Eqn 1) define an *orthonormal transformation*. This means that the unit basis vectors of the X,Y,Z system remain of unit length and maintain right angles with each other when transformed to the U,V,W system. Let us see if that works out. The unit basis vectors of the X,Y,Z system are (1,0,0), (0,1,0), (0,0,1). Run each of them through Eqn 1 to get U,V,W coordinates of $(\frac{1}{\sqrt{3}},\frac{2}{\sqrt{6}},0)$, $(\frac{1}{\sqrt{3}},\frac{-1}{\sqrt{6}},\frac{1}{\sqrt{2}})$, $(\frac{1}{\sqrt{3}},\frac{-1}{\sqrt{6}},\frac{-1}{\sqrt{2}})$, respectively. The lengths of these new U,V,W vectors are $\sqrt{\left(\frac{1}{\sqrt{3}}-0\right)^2+\left(\frac{2}{\sqrt{6}}-0\right)^2+\left(0-0\right)^2} = 1$, $\sqrt{\left(\frac{1}{\sqrt{3}}-0\right)^2+\left(\frac{-1}{\sqrt{6}}-0\right)^2+\left(\frac{1}{\sqrt{2}}-0\right)^2} = 1$, $\sqrt{\left(\frac{1}{\sqrt{3}}-0\right)^2+\left(\frac{-1}{\sqrt{6}}-0\right)^2+\left(\frac{-1}{\sqrt{2}}-0\right)^2} = 1$. So the X,Y,Z unit basis vectors remain of unit length in U,V,W coordinates. How can we tell if they maintain right angles with each other? Answer: If the cosine of the angle between each pair is zero. A trigonometric property is that the cosine of the angle between two vectors is proportional to the inner product of the coordinates. Now the inner products are

$$(\frac{1}{\sqrt{3}},\frac{2}{\sqrt{6}},0)\bullet(\frac{1}{\sqrt{3}},\frac{-1}{\sqrt{6}},\frac{1}{\sqrt{2}}) = \frac{1}{\sqrt{3}}\frac{1}{\sqrt{3}}+\frac{2}{\sqrt{6}}\frac{-1}{\sqrt{6}}+0\frac{1}{\sqrt{2}} = 0$$

$$(\frac{1}{\sqrt{3}},\frac{2}{\sqrt{6}},0)\bullet(\frac{1}{\sqrt{3}},\frac{-1}{\sqrt{6}},\frac{-1}{\sqrt{2}}) = \frac{1}{\sqrt{3}}\frac{1}{\sqrt{3}}+\frac{2}{\sqrt{6}}\frac{-1}{\sqrt{6}}+0\frac{-1}{\sqrt{2}} = 0$$

$$(\frac{1}{\sqrt{3}},\frac{-1}{\sqrt{6}},\frac{1}{\sqrt{2}})\bullet(\frac{1}{\sqrt{3}},\frac{-1}{\sqrt{6}},\frac{-1}{\sqrt{2}}) = \frac{1}{\sqrt{3}}\frac{1}{\sqrt{3}}+\frac{-1}{\sqrt{6}}\frac{-1}{\sqrt{6}}+\frac{1}{\sqrt{2}}\frac{-1}{\sqrt{2}} = 0$$

So the X,Y,Z unit basis vectors remain perpendicular to each other in U,V,W coordinates. Therefore, Eqn 1 is an orthonormal transformation of X,Y,Z coordinates into U,V,W coordinates. And therefore the Eqn 1 transformation preserves the original geometry of the data points.[5]

[5] You can try this with the X,Y,Z and U,V,W data in the Excel file *PCA XYZ UVW ABC data and prin comps.xlsx*. The length of each X,Y,Z vector is the same as the length of the corresponding U,V,W vector. The inner product (proportional to cosine) of every pair of X,Y,Z vectors is the same as the inner product of the corresponding pair of

There is an easy way to check for an orthonormal transformation. Start by forming the matrix of coefficients of the transformation. For example, here is the matrix of coefficients of Eqn 1:

$$\mathbf{M} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

Note that the columns of **M** are the transformed unit basis vectors. So the sum of squares of each column of **M** should equal 1. (They do.) And the inner products of each pair of different columns should be zero. (They are.) So **M** defines an orthonormal transformation. Any transformation that satisfies these two conditions is an orthonormal transformation (unit lengths of vectors and zero inner products), and all orthonormal transformations satisfy these two conditions.

But wait! What about Eqn 2? Eqn 2 is a different rotation that changes U,V,W coordinates back into original X,Y,Z coordinates. So Eqn 2 should also preserve the geometry going backward. Eqn 2 should also define an orthonormal transformation. Does it? Here is the matrix of coefficients for Eqn 2:

$$\mathbf{M}^{\mathrm{T}} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

It is easy to see that the sum of squares of each column of $\mathbf{M}^{\mathrm{T}}$ is 1 and that the inner product of each pair of different columns is zero.

Moreover, if you compare **M** with $\mathbf{M}^{\mathrm{T}}$, you will notice something interesting: $\mathbf{M}^{\mathrm{T}}$ is the transpose of **M**.[6] That is, if you interchange the rows of **M** with the columns of **M**, you will get $\mathbf{M}^{\mathrm{T}}$. Furthermore, logically, if you follow Eqn 1 immediately with Eqn 2, you should get back where you started. So if you compute …

$$\mathbf{M}\,\mathbf{M}^{\mathrm{T}} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

you do get back the original unit basis vectors. This also means that **M** and $\mathbf{M}^{\mathrm{T}}$ are inverses of each other, in addition to being transposes of each other. This makes geometric sense, as well,

U,V,W vectors. Moreover, the distance between every pair of X,Y,Z vectors is the same as the distance between the corresponding pair of U,V,W vectors.

[6] For a refresher short course on matrix mathematics, see my *Matrix Mathematics Notes.pdf* on Canvas.

because Eqn 2 is the rotation that you would apply to reverse and “cancel out” the rotation of Eqn 1.

Now, Eqn 1 (equivalently, **M**) is not the only orthonormal rotation. There are infinitely many. But only some of them are statistically and practically interesting for a given data set. Principal components analysis provides one particularly interesting orthonormal rotation of data. Although our **M** is orthonormal, it is not a principal component transformation. All PCs are orthonormal, but not all orthonormals are PCs.

Principal Components Analysis (PCA) chooses interesting new coordinates for the data. PCA chooses the new axes sequentially one after another. The first new axis is the one on which the data are maximally spread out. This is useful because it provides the dimension on which one can most differentiate one data point from another. So the first new axis has the most information for distinguishing the points from each other. PCA looks at the data from all possible angles and figures out which of the infinite number of views spreads out the data the most. This statement has a precise mathematical meaning, but can be explained easily by a graphic. Consider the U,V,W dataset again:

Figure 9. U,V,W data with ellipsoid and dimension of maximal variance.

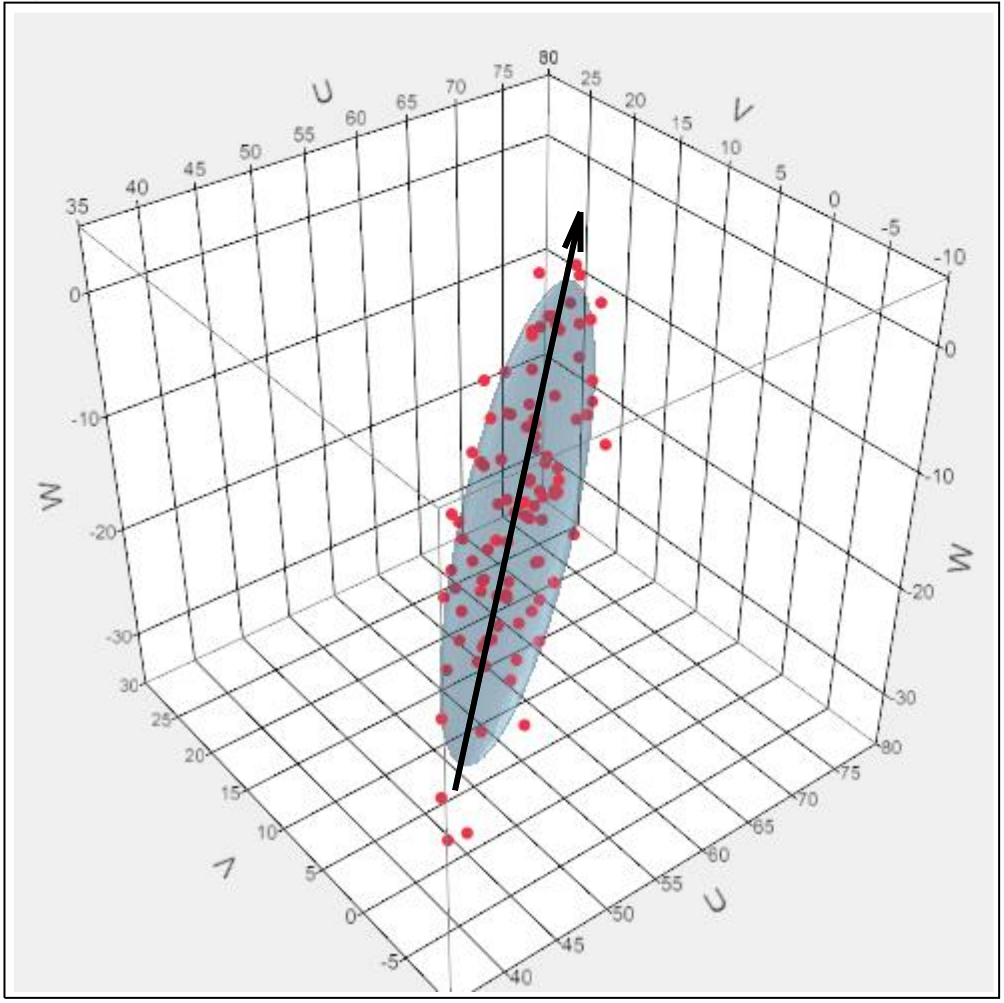
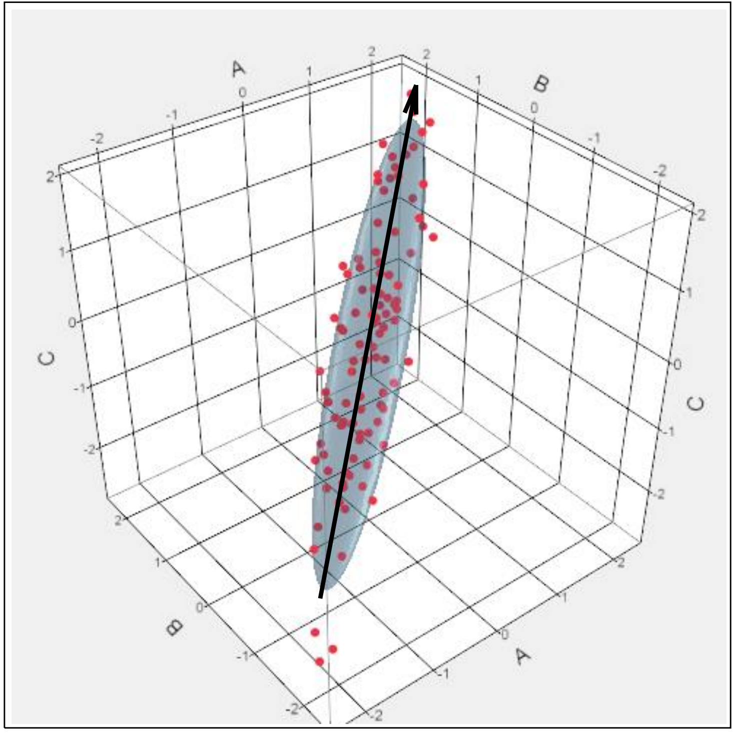


Figure 9 includes an ellipsoid that covers most of the data. The dark arrow through the long axis of the ellipsoid is the dimension on which the data are most spread out. This means that if you project every data point perpendicularly onto the dark arrow and get the coordinates of the projections onto the dark arrow, then the variance of those coordinates is larger than the variance of the data projections onto any other possible arrow. The dark arrow represents the first principal component.

… except that PCA usually standardizes the data before finding the variance-maximizing dimension. Standardization means that the mean of each variable is subtracted from each coordinate and the result is then divided by the standard deviation of the variable: Data point (u_i, v_i, w_i) becomes $\left(\frac{u_i - \bar{u}}{s_u}, \frac{v_i - \bar{v}}{s_v}, \frac{w_i - \bar{w}}{s_w}\right)$. Each of the standardized variables now has mean 0 and variance 1. Standardization is employed in order to treat all variables equally. If one variable has a measurement scale like inches, its variance would tend to be much larger than other variables measured in feet. Since PCA seeks to maximize variance, standardization prevents the PCA procedure from catering to variables that are measured in small units. Standardization forces all variables to be treated the same.[7]

The following Figure 10 shows the standardized data (now called A,B,C) with ellipsoid and first principal component arrow. Not much has happened to the geometry, compared with Figure 9.

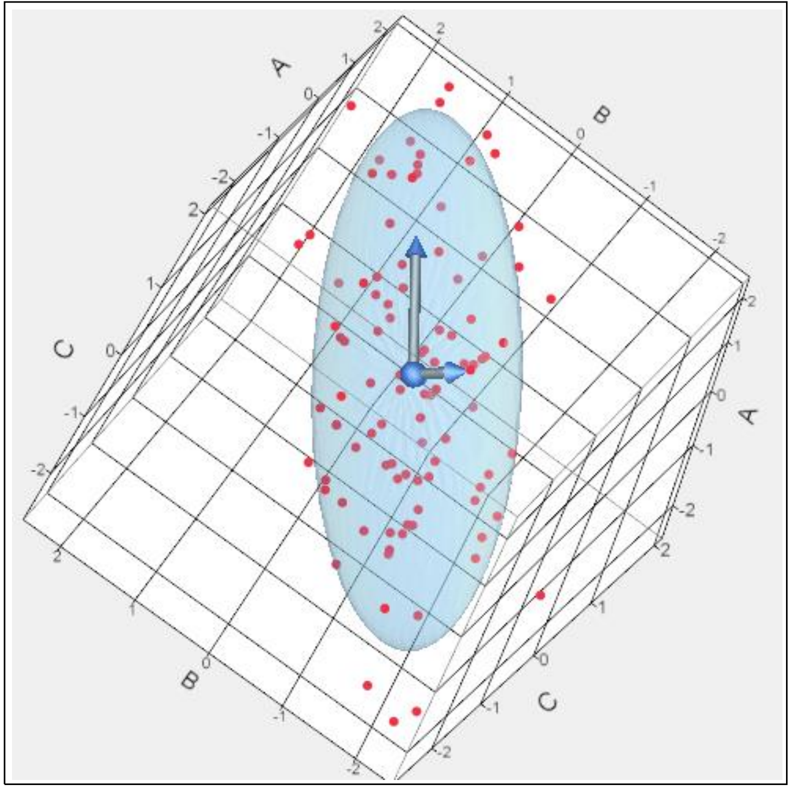
Figure 10. Standardized U,V,W (A,B,C) data with ellipsoid and dimension of maximal variance.



[7] Treating all variables the same may be appropriate, or may not be. There are versions of PCA that do not standardize, but only *center* (subtract the mean) – or even do nothing. The decision to standardize or not is non-trivial! Different results can be obtained.

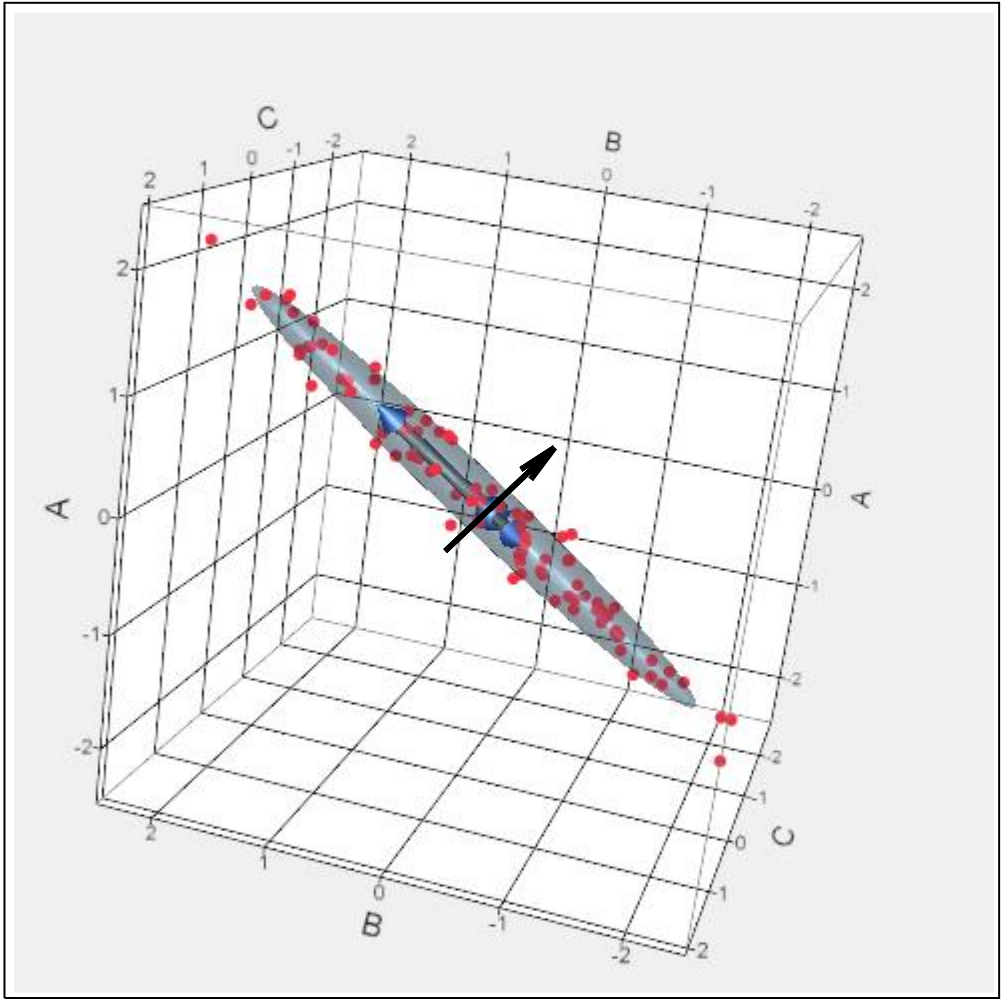
How is the second dimension determined? If the transformation is to be orthonormal, then the second dimension must be perpendicular to the first dimension – perpendicular to the dark arrow in Figure 10. There are infinitely many candidates. Attach a second arrow to the middle of the first and pointing perpendicularly away from the first. Hold the first arrow fixed and rotate the second arrow in a circle all the way around the first. All of the directions in that circle are candidates. Which one to choose? Continuing the idea of spreading out the data as much as possible, PCA chooses the perpendicular direction that maximally spreads out the data, subject to being perpendicular to the first dimension. This is the direction in which the disk of data is the thickest. It is shown from a somewhat different perspective in Figure 11. In Figure 11, the lengths of the arrows are proportional to the lengths of the ellipsoidal axes (more on this momentarily.)

Figure 11. The first two principal components of now standardized U,V,W (A,B,C) data.



There is one dimension left. The third dimension should be chosen perpendicular to the first two dimensions, and in such a way that the data are maximally spread out on it. But there is only one possible dimension left – the dimension that is perpendicular to the plane of the two arrows in Figure 11. The third and last dimension is shown as the dark arrow in Figure 12. I have drawn the arrow exaggeratedly long. The arrow runs through the thin part of the disk of the data, which is shown edge-on in Figure 12. If you look closely at the edge of the data disk, you can see the first two principal components inside the disk.

Figure 12. The third principal component of standardized U,V,W (A,B,C) data.



The geometry suggests (and it is true) that the three principal components coincide with the three principal axes of the ellipsoid that envelops the data. The data are maximally spread out on the longest axis. Subject to being perpendicular to the first axis, the data are maximally spread out on the middle axis. Subject to being perpendicular to the first two axes, the data are maximally spread out on the remaining axis. This is how principal components work. Each axis is perpendicular to all of those that came before and maximally spreads out the data subject to perpendicularity. If there were more than three original variables, there would be more than three principal components. But the pattern would continue. Each component would be perpendicular to all of those that came before and would maximally spread out the data subject to perpendicularity.

After all principal components have been calculated (the usual terminology refers to *extracting* the components), the new coordinates of the data can be calculated in the principal component coordinate system. This amounts to a rotation of the (standardized) original data into the principal component coordinate system. The rotation is orthonormal. So the geometry of the (standardized) data is preserved.

What are the equations of the rotation? Answer:[8]

$$\left\{\begin{array}{l}\text{Prin1} = \quad 0.62889A + 0.56336B + 0.53584C \\ \text{Prin2} = -0.06860A - 0.64629B + 0.76000C \\ \text{Prin3} = -0.77446A + 0.51471B + 0.36780C\end{array}\right\} \text{(Eqn 3)}$$

You can easily check that the matrix of coefficients has columns of unit length and zero inner products between different columns. So the transformation is orthonormal. The principal components of (U,V,W) or (A,B,C) are given by Eqn 3.

A very useful property of the principal components transformation is that the principal components are ordered in terms of importance. This statement requires a little bit of explaining. First, observe that Var(Prin1) ≥ Var(Prin2) ≥ Var(Prin3). This is true because the data are maximally spread out on Prin1. That is, if you compute the Prin1 coordinates of all 100 A,B,C data points and calculate the variance of those 100 coordinates, the result must exceed the variance of the Prin2 coordinates of all 100 data points. (Otherwise, Prin2 would have been a better choice for the first coordinate than Prin1!) Similarly, the variance of the Prin2 coordinates of all 100 data points must exceed the variance of all 100 Prin3 coordinates.

Second (continuing the explanation), it is a platitude that data vary for reasons. Collectively, this dataset varies a lot because of Prin1, less because of Prin2, and not much at all because of Prin3. So most of the explanation for *why* these data vary can be laid at the doorstep of Prin1; less of the explanation is due to Prin2; and very little to Prin3. In fact, so little of the explanation for why the data vary can be attributed to Prin3 that a good argument can be made for dropping Prin3 from a list of factors that account for variation in these data. *This discussion associates variation with explanation.* Big variation, big explanation. Small variation, small explanation. *This discussion also associates variation with information*. Prin1 has most of the information that is in the dataset. Prin2 has less; and Prin3 has very little.

But can this intuition be quantified? Yes! It is clear from the preceding discussion that the importance of a component and the magnitude of its variation are related to the length of its corresponding ellipsoid axis. And that can be measured precisely. Figures 11 and 12 show the lengths of axes 1 and 2 (for components 1 and 2) drawn to scale. Axis 3 is exaggerated in Figure 12, as previously noted. But the software calculates the squared lengths of all three axes and they are 2.4991, 0.4854, and 0.0155, respectively. Notice that the total is exactly 3.0. Thus, the first component is responsible for about 83% of the total variance (2.4992÷3), the second for about 16% (0.4854÷3), and the third for only about 0.5% of the total. The total variance of the three principal components is 3.0 because the principal component rotation is an orthonormal transformation that preserves the geometry, therefore preserves distances, and hence preserves variance. The original variance is the total variance of the (standardized) U,V,W (A,B,C) data that were rotated, namely:

$$Var\left(\frac{u_i - \overline{u}}{s_u}\right) + Var\left(\frac{v_i - \overline{v}}{s_v}\right) + Var\left(\frac{w_i - \overline{w}}{s_w}\right) = 1 + 1 + 1 = 3.$$

[8] Calculated by SAS or by JMP.

This total variance is preserved by the orthonormal rotation to (Prin1, Prin2, Prin3) coordinates. If principal components are extracted from standardized data, then the total variance always equals the number of variables, since a standardized variable has unit variance.

So if variance is the "essence" of a dataset, then Prin1 captures most of the "essence" of our U,V,W data (actually, the standardized A,B,C data). Therefore, it may be argued that Prin1could be used as a reasonable surrogate for the original data for some applications. Perhaps Prin2 should be used along with Prin1. A simplification results – a reduction of the dimension of the data – if Prin1 and possibly Prin2 substitute for the original data. A major decision in the use of principal components is how many components to retain in order to represent the original data. Various rules have been proposed. I will discuss such rules later. As we shall also see later, the orthogonality of Prin1 and Prin2 has beneficial consequences through making the components uncorrelated.

Now we can make another interpretation to link up with information. A dataset contains information. How much information? Intuitively, the "essence" of a dataset must be the information that it contains. But we have just identified the "essence" of a dataset as variance. So, intuitively, it must be true that information and variance are fundamentally unified concepts. This idea should not be so strange. You are familiar with R-square in regression. R-square is the proportion of *Y*-variance that is explained by the predictor variables collectively through the linear combination of predictors that is maximally correlated with *Y*. The more that *Y* varies when an *X* varies, the more of *Y* that *X* explains – the more information about *Y* that *X* contains. In parametric statistical models, the sufficient statistics contain all of the information that the original data contain. However, in nonparametric models and in parametric models for which the sufficient statistics offer little data reduction, there is a need for ways to assess the capture of information that is less than 100%. The proportion of variance captured offers one popular way. In PCA, after standardizing the original variables, they all have the same variance, namely 1. So the total variance of the dataset = number of variables. In the PCA of Eqn 3, the total variance and total information is 3. PCA preserves all of the original information, but offers a useful redistribution of that information. In the PCA of Eqn 3, Prin1 contains about 83% of all of the information in the original dataset; Prin2 contains about 16%; and Prin3 only about 0.5%. There is variation in the dataset primarily because of Prin1, much less because of Prin2, and almost not at all because of Prin3.

A major issue with the use of principal components is how to interpret them. If the original X,Y,Z variables were apartment area, number of bathrooms, and number of bedrooms, respectively, then we readily understand what they mean. However, a principal component is a linear combination of the original variables. So what would $\frac{1}{\sqrt{3}}X+\frac{1}{\sqrt{3}}Y+\frac{1}{\sqrt{3}}Z$ mean? The interpretation of principal components is an art. To interpret a principal component, we need to take into account the magnitudes and signs of the coefficients as well as the inherent meanings of the combined variables. Then we need to be creative in discovering a plausible "story" that could account for the component. For example, in the case of the hypothetical component $\frac{1}{\sqrt{3}}X+\frac{1}{\sqrt{3}}Y+\frac{1}{\sqrt{3}}Z$ with X,Y,Z as area, bathrooms, and bedrooms, the component might represent the "size" of an apartment as a composite. Large apartments tend to have more area, more bathrooms, and more bedrooms. With these three variables equally weighted in the component formula, the component would tend to have larger values for larger apartments and smaller values for smaller apartments. So the interpretation of this component as “size” makes sense in terms of the meanings of its constituent variables and their weights. Frequently in real-world applications, the story is the most important part of the analysis.

So shall we interpret the meaning of a principal component by looking to the values of the weights of the original variables? The interpretation of a component would be aided by knowing how much each of the original variables correlates with the component. Then in order to give the component its meaning, we would look to the meanings of those original variables that correlate highly with the component. Original variables that correlate at a low level with the component would not contribute to the meaning of the component. The correlation between an original variable and a principal component is called a **loading**. Analysts like to look at the **loadings matrix** to interpret component meanings. Here is the loadings matrix for the A,B,C example:

	Prin1	Prin2	Prin3
A	0.99418	−0.04779	−0.09657
B	0.89058	−0.45027	0.06418
C	0.84708	0.52949	0.04586

Each matrix entry is the correlation between the row variable and the column component. Since standardization leaves correlations unchanged, the matrix also shows the correlations between the U,V,W variables and the components. Since all three of the original variables are highly correlated with Prin1 (the terminology is that all three variables “load high” on Prin1), then Prin1 represents a composite of all three original variables. To divine an interpretable meaning from this observation, we would look to features that A,B,C (equivalently, U,V,W) have in common. Prin2 has moderate correlations with B and C, of opposite sign, and therefore Prin2 is a **contrast** component: Data points that score high on Prin2 have high C values and low B values; data points that score low on Prin2 have low C values and high B values. A plays little role in Prin2. To divine an interpretable meaning from this, we would look to features in common to high C/low B and low C/high B. No variable loads high on Prin3. Prin3 would likely be dropped.

There is another point that you should note: If you compare the columns of the matrix loadings with the principal components coefficients – i.e., the row coefficients of the equations of rotation (Eqn 3) – you will see that they are proportional. That is,
0.99418÷0.62889 = 0.89058÷0.56336 = 0.84708÷0.53584 = 1.5808; and
-0.04779÷-0.06860 = -0.45027÷-0.64629 = 0.52949÷0.76000 = 0.6967; and
-0.09657÷-0.77446 = 0.06418÷0.51471 = 0.04586÷0.36780 = 0.12469.
It is the squares of these constants of proportionality that are particularly interesting:
1.5808^2 = 2.4991, and
0.6967^2 = 0.4854, and
0.12469^2 = 0.01555.

These numbers are precisely the squared lengths of the Prin1, Prin2, and Prin3 axes, respectively.[9] That is, they are the variances of the A,B,C projections into the three principal component axes of rotation.

Mathematical note: The principal components are actually the eigenvectors of the correlation matrix of the original (U,V,W) data.[10] The squared lengths of the principal component axes are actually the eigenvalues of the correlation matrix of the original (U,V,W) data. Eigenvectors and eigenvalues are easy for computers to calculate. That makes principal components an eminently feasible technique in working with data.

SUMMARY

You should now understand that PCA is a special kind of orthonormal transformation of (standardized) data. Namely, PCA amounts to a rigid rotation of (standardized) data into a new coordinate system so that the variance of the first coordinate dimension of the rotated data is maximized. Subject to the first coordinate maximization, the variance of the second coordinate dimension of the rotated data is maximized. Subject to the first and second coordinate maximizations, the variance of the third coordinate dimension of the rotated data is maximized; and so on.

There are two key aspects to the PCA rotation: orthonormality and variance maximization. Orthonormality preserves the geometry of the original data in the sense that the distances and angles between all points are the same after rotation as before. All orthonormal transformations preserve geometry. But PCA is a special orthonormal transformation that frontloads the data variance onto the low-order principal component dimensions in the manner described in the preceding paragraph.

Because of these two properties, PCA offers unique opportunities for gaining insight into data. Any statistical analysis that depends only on the distances and angles between data will yield the same results after application of any orthonormal transformation to the original (standardized) data as before. Example: Most aspects of regression. So if we do not like the original coordinate system, we may choose a more convenient orthonormal rotation. If PCA is selected, the new dimensions will be uncorrelated (because orthogonal) and will highlight the important new dimensions that carry the most variance and also will highlight the unimportant new dimensions that carry minimal variances. The unimportant new dimensions may be candidates for elimination, thus simplifying the data by reducing the number of dimensions. The important new dimensions may provide new insight. The meanings of the new rotated dimensions may be inferred by interpretation of the *loadings matrix*, which shows the correlations between the original data and the rotated data. The meaning of a rotated dimension is most strongly related to the meanings of the original dimensions that correlate most strongly with the rotated dimension. These loadings (correlations) that drive the meaning of the components are proportional to the

[9] See discussion in the third paragraph following Eqn 3.

[10] The eigenvectors of (U,V,W) and (A,B,C) are the same: Since correlations are not affected by standardization, then (U,V,W) and (A,B,C) have the same pairwise correlation matrices and therefore the same eigenvectors and eigenvalues. (X,Y,Z) and (U,V,W) have different eigenvectors and eigenvalues since correlation is not preserved under orthonormal transformation.

component coefficients. So the coefficients may also be used to interpret the meanings of the components as long as it is recognized that the coefficients are not the actual correlations.

In addition to the preceding general facts about PCA, you also learned a bit of the mathematics: Suppose **X** is a *n*x*p* data matrix of *n* rows of data and *p* columns of (standardized) variables. Suppose **M** is any *p*x*p* orthonormal matrix.

- Being orthonormal means $\mathbf{M}\mathbf{M}^{T}= \mathbf{I}_{p}$.
- The orthonormal transformation of **X** is **XM**.
- The matrix of the reverse orthonormal transformation is the transpose of **M**, namely, $\mathbf{M}^{T}$.
- Reverse transforming the data back yields $(\mathbf{XM})\mathbf{M}^{T}= \mathbf{X}(\mathbf{M}\mathbf{M}^{T}) = \mathbf{X}\,\mathbf{I}_{p} = \mathbf{X}$.
- $\mathbf{M}^{T}$ is the inverse of **M**: $\mathbf{M}^{-1} = \mathbf{M}^{T}$.

Now that you understand these general features of PCA and the foundations of its mathematics, you are prepared for a more detailed examination of PCA in part 2 of these notes.