UNIVERSITY OF WISCONSIN
DATA SCIENCE

Important note: Transcripts are **not** substitutes for textbook assignments.

Learning Objectives

By the end of this lesson, you will be able to:

- Explain why principal components analysis (PCA) is an unsupervised method.
- Visualize the components of PCA using appropriate tools.
- Define and interpret principal components and loadings.
- Select an appropriate number of components.
- Apply PCA with the `prcomp` function.

Iris Example

Using iris data set, already loaded in *R*, with variables.

We know: Four size measurements (all in cm) of iris plant: *Sepal.Length*, *Sepal.Width*, *Petal.Length*, *Petal.Width*

We don't know: Species; ~~"setosa"~~, ~~"versicolor"~~, and ~~"virginica"~~

1. Identify how appropriate for PCA.
2. Visualize with plots.
3. Describe possible conclusions

During this lecture presentation, we will again be examining the iris data set, assuming that we do know the size measurements of the iris plants, but we do not know the associated classifications in the species. So for this purpose, we are trying to clarify the measurements of the plants to organize that information without an end response in terms of a predictive purpose. During this lecture, we'll work to explain why this data set can be appropriate for principal components analysis and how we may visualize some of that process with plots. In addition, we'll describe what the principal components, as well as their weights, are able to tell us about the data set.

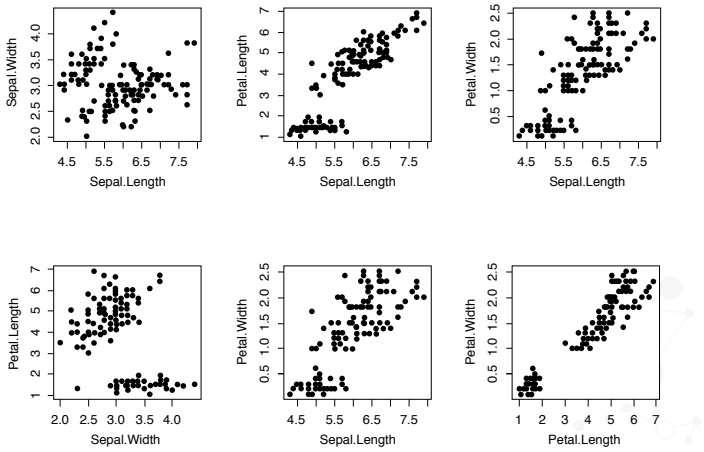
Notes:
Link: iris data set

Iris: Measurements-Only

Species = ?? ●

Which features most important in describing values?

- Visually, looking for direction of highest variability.
- Avoid using multiple variables to relate same information.



Previously when applying clustering methods to grouping the observations in this data set, we had selected pedal length and pedal with as jointly useful towards this purpose since both also appeared individually to be able to distinguish groupings. The consequence, though, is that the two variables provide very similar information about the data in terms of direction or overall behavior, since these two variables are highly correlated. In this case, because both variables are size measurements for the pedals, they're both providing some sort of size, assessment of the pedals of the iris plants. Thus, we'll next be considering a method for retaining near-full information about the data while reducing the number of variables we consider to summarize the data.

Dimension reduction

Why not retain all features?

Problem: Iris correlations; repetitive information.

```
> round(cor(iris[,1:4]),3)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length        1.000      -0.118        0.872       0.818
Sepal.Width        -0.118       1.000       -0.428      -0.366
Petal.Length        0.872      -0.428        1.000       0.963
Petal.Width         0.818      -0.366        0.963       1.000
```

Consideration: Visualizations only in lower dimensions

Solution: Dimension reduction (not variable reduction) → "projection"

A valid question may be, why go through the process of dimension reduction? Why not simply retain all features to describe the data in the full set? One of the reasons has to do with repetitive information. As we can see in the iris data set, the correlations between the predictors are generally quite high, and in the case of pedal length and pedal width, very, very high with the correlation of 0.96, which says that they provide information that is similar in nature regarding the iris plants. So with the idea of being able to better understand the data in a more simplistic way, and also to be better able to visualize the data, which we really can only do well in lower dimensions, we proceed with dimension reduction. Note that this is not variable reduction, we're not getting rid of any of the variables, but we're really orienting it, or projecting it, into a new space.

Principal Components Analysis: Purpose

Goal: Project the data into a lower-dimensional space (determined by the principal components) while retaining the maximum information as measured by variability of the data.

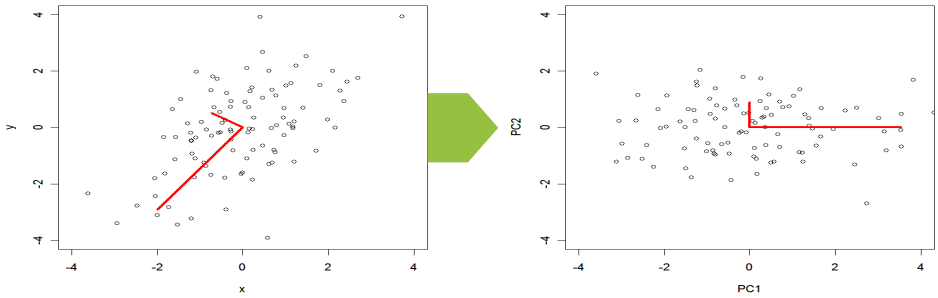
- Feature space (referring to the p-dimensional space of the p variables) may be summarized more concisely.
- Identify directions in feature space that explain the greatest amount of variation.
- Some guidance available for best number of components.

So our goal is to reorient the space in which we are viewing the data along lines that correspond to the most variability. That is, we want to identify directions that explain movement in the data. Principal components analysis, abbreviated as PCA, achieves that purpose for us through the magic of matrix algebra.

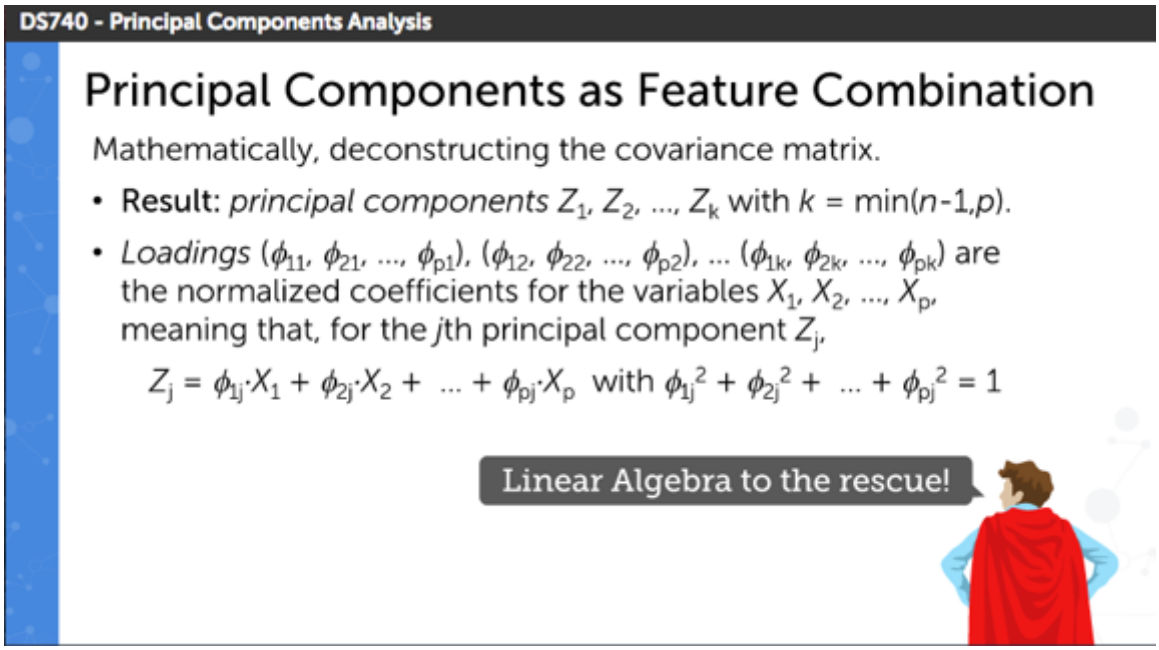
PCA Conceptually

Use these “important” directions (principal components) to better understand the data.

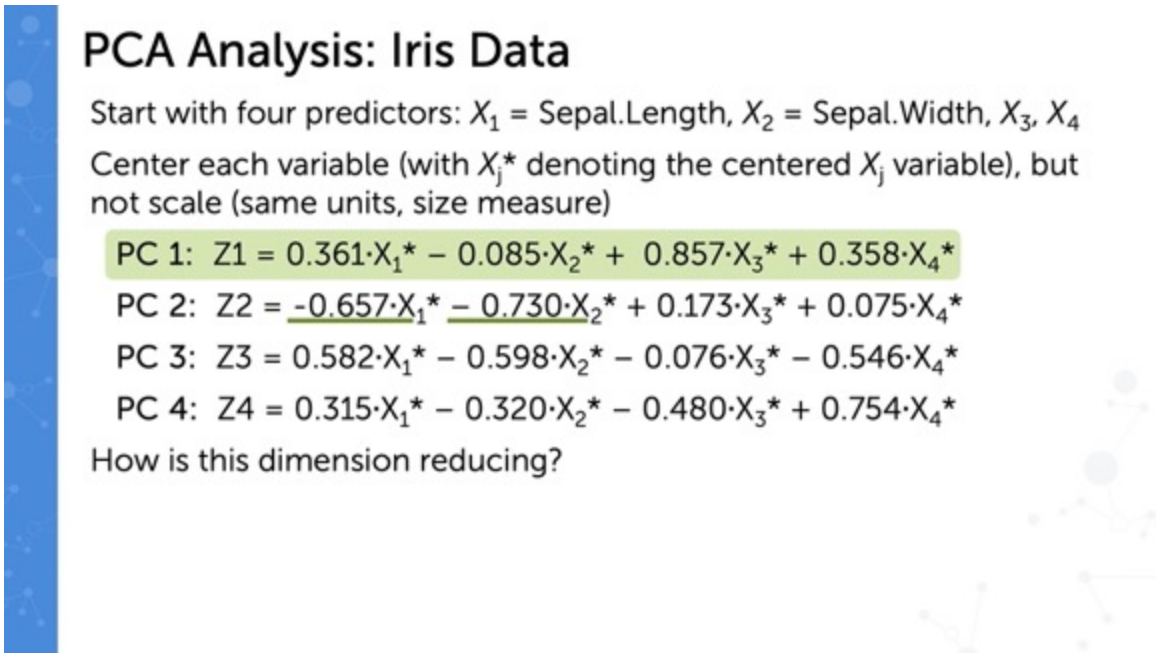
Two variable PCA:



In this slide, we see a visualization of that re-orientation. On the original axes we have a direction in which the data are more spread out across an angle from the lower left quadrant into the upper right quadrant. This is identified by the longer, red direction identified by a segment. That segment, along with the directionality of the data, is then oriented along, what is known as, principal component 1, while the shorter and less important direction in terms of the variability of the data, is oriented in the direction of principal component 2.



The theory behind principal components is based on linear algebra. It involves deconstructing the covariance matrix of the centered variables. This sounds complex, but actually results in a fairly simple result. The principal components turn out to be nothing but linear combinations of the variables denoted as x1, x2, up to xp. And the principal components themselves, those linear combinations of the x's, are denoted as z sub 1, z sub 2, up to z sub k, where k can't get any bigger than n minus 1 or p. The coefficients that result for the linear combination of those x's into the principal components are known as loadings. So for example, to obtain principal component z1, we use loading phi sub 1, 1 times x1, plus phi sub 2, 1 times x2, all the way up to phi sub p1, times xp where those coefficients are normalized to have the sum of their squares be one.



This computation of principal components is illustrated using the Iris data set. For this data set, we center each variable, but choose not to scale since all the variables are recorded in the same units and all act as some general measure of size. Denoting the four variables as x1, x2, x3, and x4, as labeled at the top, and denoting the centered values of those with a star, we pull out the principal components derived from a matrix decomposition.

For example, PC1, the first principal component, which is the most important principal component, is equal to a linear combination of x1 star through x4 star with the heaviest weight on x3 star, the petal length. And sepal length and petal width have the next highest weights.

Overall this might be considered mostly the dimensions of the petal. PC2 has the next highest importance, and the loadings are highest on the sepal length and sepal width with the largest coefficients in front of x1 star and x2 star. The size of the loadings is what matters in terms of the importance of the variables to the principal component.

Note that we started off with four variables, and we wound up with four principal components. So how is this dimension reducing? And we'll look at that on the next couple of slides.

PCA: Relative Component Importance

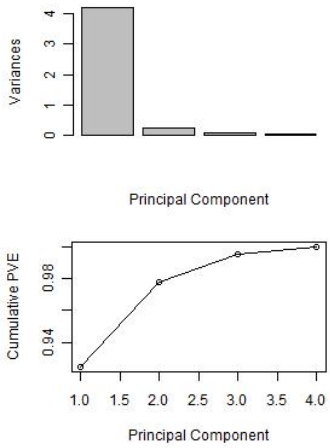
PCA based on explaining variability.

Define $v_j = \text{Var}(Z_j)$; then the **proportion of variability explained** (PVE) by the jth principal component Z_j is $v_j/(v_1 + v_2 + \ldots + v_k)$.

Iris data, with $v_1 = 0.9246$, $v_2 = 0.0531$, $v_3 = 0.0171$, $v_4 = 0.0052$

Cumulative PVE by

- Component 1: 92.46%
- Components 1 and 2: 97.77%



Describing dimension reduction entails a discussion of the variability explained, because that is the basis for principal components analysis. A visual of the variances across the principal components is shown in the bar chart at the right of the slide and shows a dramatic shift from the variance explained at principal component 1 to the remaining principal components. A more useful measure describes the proportion of variability explained by each of the principal components, and is simply found by taking the variance of a principal component divided by the sum of the variances of all k principal components. The variances for the principal components are given for the iris data and we see the dramatically higher variance for principal component 1.

A more useful plots and more useful summary relies on cumulative proportion of variability explained, and this can be plotted across the principal component by combining the proportion of variability explained accumulated across the components. So we see that after components 1 and 2 are included, there is a bit of a shift, what's called an elbow, in the plot of the cumulative PVE, which suggests that we're probably not gaining much by adding any components beyond two. Thus, we would limit ourselves to the first two principal components.

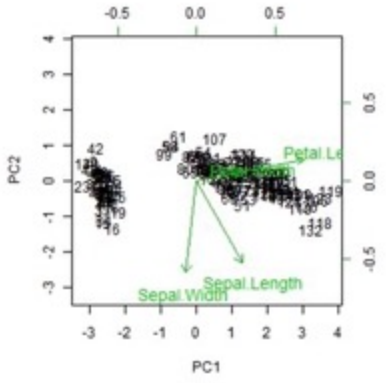
PCA Analysis: Visual Concept via Biplot

First principal component (on x-axis) is direction of greatest variation.

PC 1: $Z_1 = 0.361 \cdot X_1^* - 0.085 \cdot X_2^* + 0.857 \cdot X_3^* + 0.358 \cdot X_4^*$

PC 2: $Z_2 = -0.657 \cdot X_1^* - 0.730 \cdot X_2^* + 0.173 \cdot X_3^* + 0.075 \cdot X_4^*$

With centered values X_j^*



Visualizing the behavior of principal components can be seen in what is known as a biplot. This plots any two of our principal components, where the first of the two listed is plotted along the horizontal axis and the second is plotted on the vertical axis. By default, we will use PC1 and PC2 in this command.

We see the data points reoriented along these axes. That is the ones as noted at the bottom and left-hand side of this plot. And the observations are denoted by their numeric index. In addition, the lines in green designate the direction of the loadings that are used to produce these principal components. And the axis labels on the right and top show the loading values.

So for example, sepal length has loading value 0.361 on PC1. And so the sepal length text is at about 0.36 along that top axis and has a loading of negative 0.657 in PC2. And you can see it has a value of about negative 0.657 on that axis on the right-hand side. Similarly, if we take a look at each of the other variables, we can correspond their coordinates in the principal component's space to the loadings and the linear combinations that make up the principal components.

PCA Analysis: Process

1. Always center variables.
2. Scale if desire variables to count equally in measure (generally, do so unless units are same for all variables).
3. Calculate loadings and principal components.
4. Describe the proportion of variability explained by consecutively adding in principal components.
5. Select appropriate number of principal components.
6. Use principal components to interpret and visualize behavior in data.

When running through a principal components analysis, we always center variables in terms of the matrix decomposition. And generally we do scale, unless units are the same for all variables or have the same meaning for all variables, since we do want the variables to count equally in the measure. We then use an R function to calculate the loadings and the principal components, and proceed to describe the proportion of variability explained by adding in each of those principal components, which allows us to then select a reasonable number for how many principal components to retain. We also use appropriate graphical methods to interpret and to visualize the principal components, and their relationship to the original variables.

PCA: Computation

`prcomp`: input is matrix of variables.

- Options: `center` (= *T* default) and `scale` (= *F* default).
- `Scores` refer to the values of PCs for the data points.

`summary`: of fitted PCA, provides standard deviation, PVE, and cumulative PVE.

`plot`: of fitted PCA, plots the variances.

`biplot`: plots first two PCs, with the data points' scores plotted in the principal component space.

We will be using the prcomp function that is in the base package to compute our principal components analysis. The primary input is a matrix of variables. And while the default is to center set equal to true, if we want a scale, we do have to reset that to be true, because the default is false. One of the outputs that will examine are scores, which refer to the values of the PC for the data points, so that will be part of prcomp analysis.

There's also a summary function available for prcomp, which gives us some of the important numeric features-- the standard deviation, the proportion of variability explained, and the cumulative PVE. The plot gives us a very simple plot of the variances, the biplot that we saw on a previous slide. And in order to produce the biplot, we need a separate function for plotting the two PCs appropriately. The data point scores will be plotted on a principal components space, along with the vectors corresponding to the loadings for each of the variables used to build the principal components.

Some Possible Uses

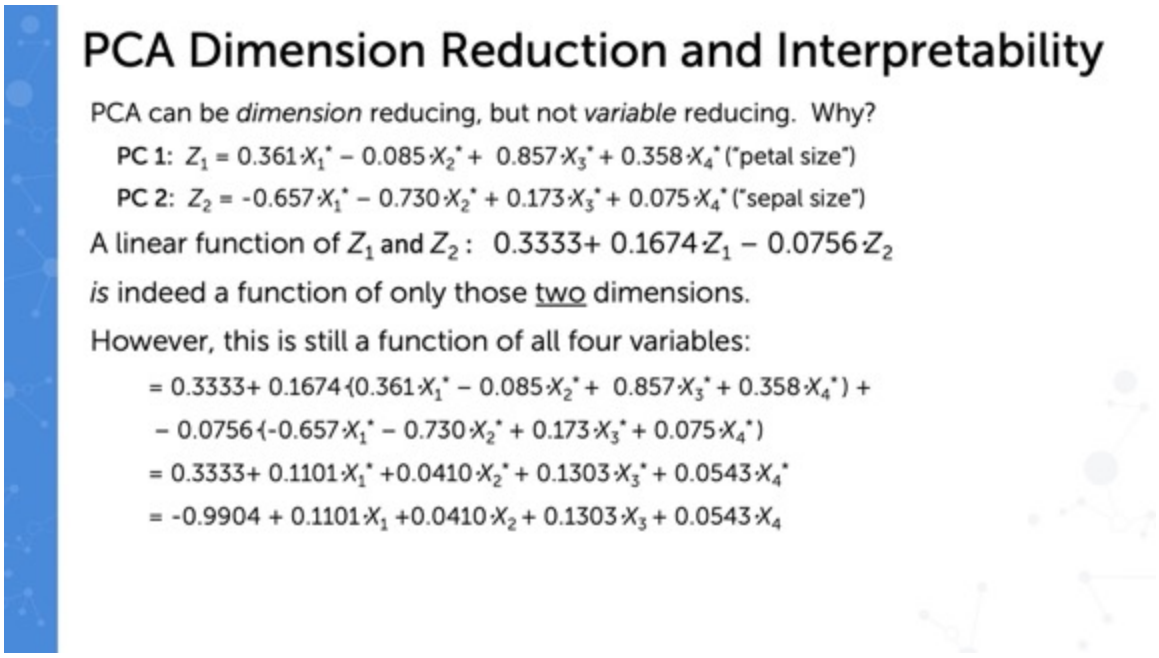
Honing in on important features allows researcher to focus.

Customers / purchasing audiences: identify features to which marketing can appeal.

Web use / preference: identify important characteristics of individuals who might be interested in certain pages.

Gene expression data: identify genes most correlated to a particular disease, or characterize genes that relate to genetic heritage.

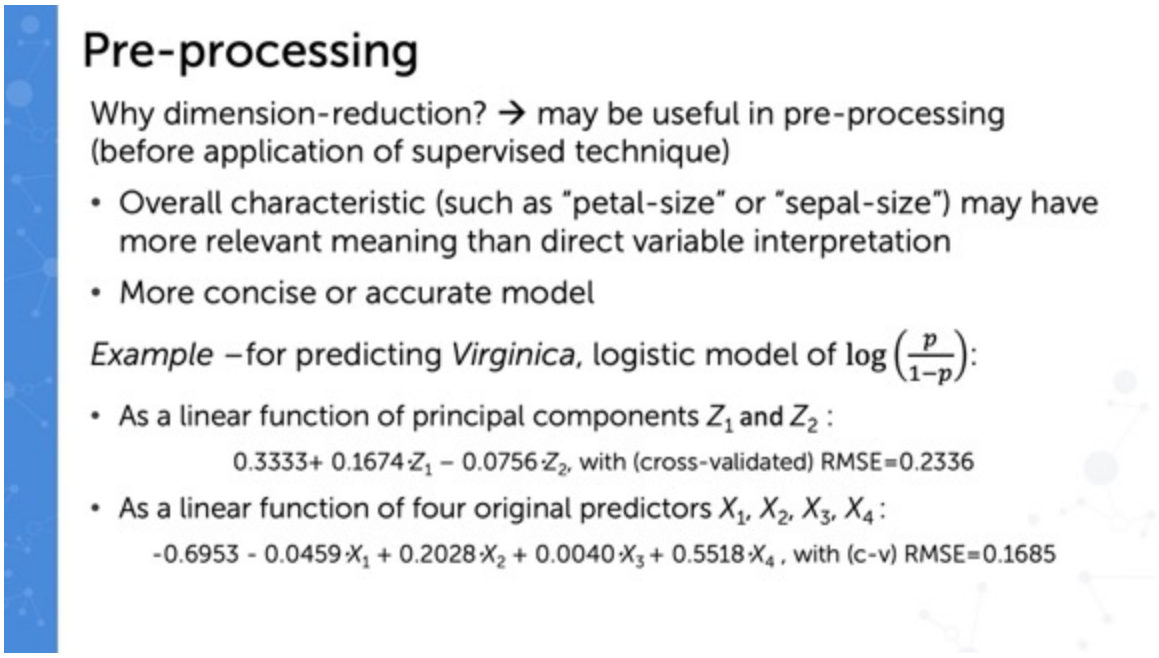
As with clustering, our primary purpose with pure principal components analysis is explorative, and that is to help identify important features, which a research study can further focus on after they're identified. For instance, we might wish to appeal to a certain type of customer and we may be able to identify features of such customers that would follow along with a particular ad or marketing strategy. With gene expression data, this might come into play with a characterization of genes related to a particular genetic heritage. Often times, that also corresponds to a spatial or geographic location.



We'll use the next couple slides to try and understand some of the uses towards which PCA can be applied. We note at the top that PCA can be dimension reducing, but not variable reducing. So going back to our Iris example, the first PC could be considered to be petal size or mostly petal size. And the second PC could be considered to be a measure of the sepal size based on the loadings in those PCs and the corresponding variable descriptions.

Then suppose we had some linear function of z1 and z2. So a constant plus another constant times z1 plus another constant times z2. This is a function of only those two dimensions and could be interpreted as the z2 value balances out to some extent with the z1 value. That is z2 goes in an opposite direction of z1 in terms of whatever it is that we are trying to record or identify with this linear function.

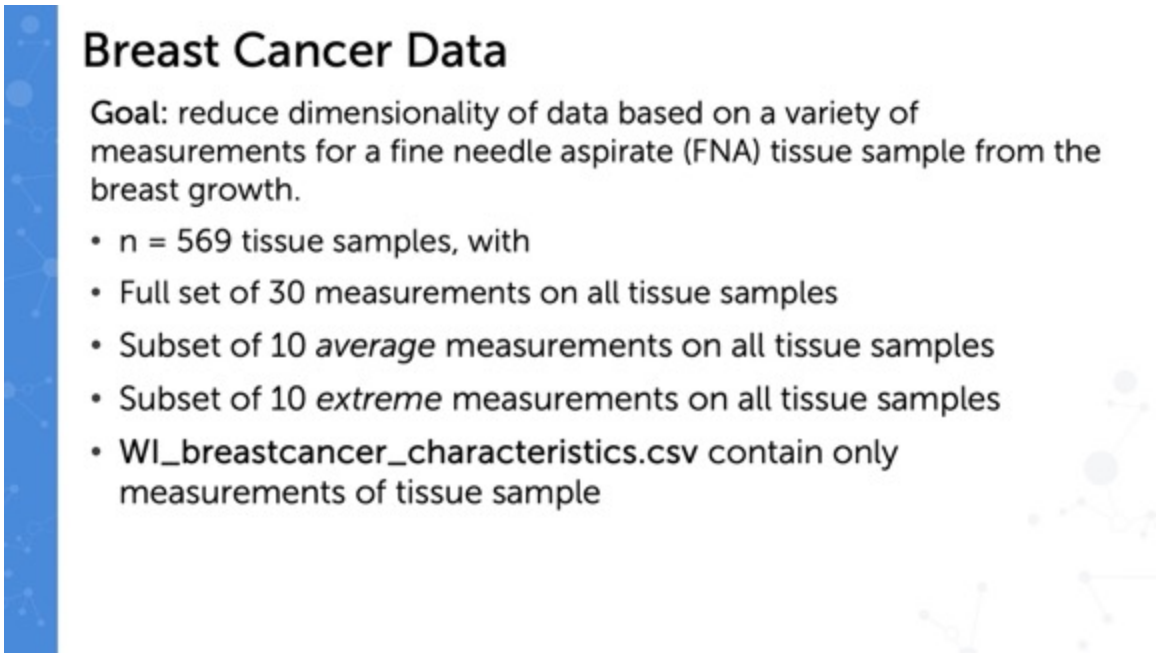
If, however, we put in the definitions of z1 and z2 into this equation, we get back a function in terms of the centered x's-- that is the x1 star through x4 star-- which in turn can be back computed into a function on the original x's. So it doesn't take away any of the predictors. It doesn't trim them out. But it does focus in on the variable directions that are most meaningful. And we might be able to derive some sort of interpretation from those variables.



One potential use of dimension reduction can be in preprocessing of our data before application to some sort of supervised technique. That is we may wish to move in the directions of highest variability that captures some sort of overall characteristic. In our example with the iris plants, that was petal size or sepal size. This may have a more relevant meaning than if we were to look at the variables individually. We may also be able to get a more concise or more accurate model.

And in the situation, again, in our iris example, if we were to use a logistic regression model of the logit, modeling that as a linear function of our principal components-- this is the equation you saw in the previous slide-- we get a function of z1 and z2 that puts more emphasis on z1, the petal size, and then counterbalances that a little bit with sepal size. Whereas the linear function of the original predictors seems to put more weight on x2 and x4, which are simply the widths of the petal and sepal respectively. So they have different interpretations, and make use of the data a little bit differently.

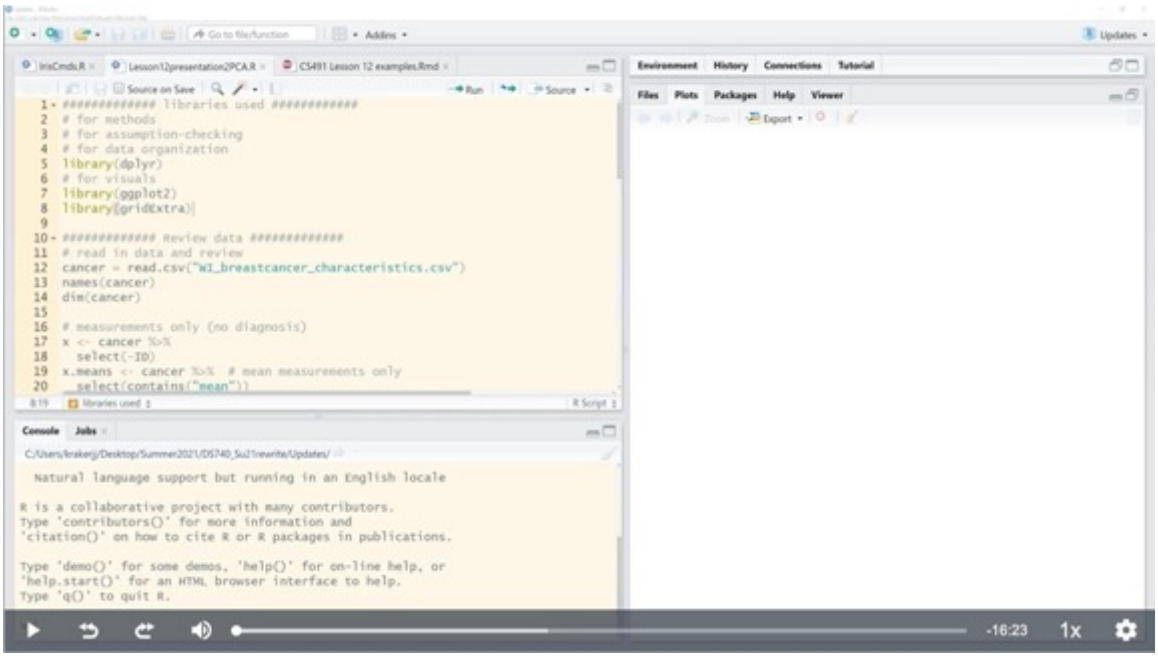
What we do want to check is which of these is the better fitted model if we're looking at a supervised model. And it turns out that in this case, using the original four predictors actually has a lower cross-validated RMSC. So we would go with that. But the model built on the principal components does have a little bit simpler of an interpretation.



We revisit the Breast Cancer data set that we saw in clustering to assess which of the 30 different variables have the most importance in terms of the variance of the data observations, the data measurements. Due to the high difference in the magnitudes of the variables, we will need to scale this data. In this situation, we're not looking at it as a preprocessing, but rather as an identification of important features corresponding to unusual observations, which may in turn help us identify what are some distinctions among these characteristics as measured on these tissue samples.

Notes:
Public-source data: wdbc.data
Public-source data description: breast-cancer-wisconsin.names
This commonly-analyzed data set appears in a number of publications, including:

- Fraley & Raftery, 2002.
- W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, 1905, p. 861-870, San Jose, CA, 1993.



In this recording, we revisit the Wisconsin breast cancer data as was previously introduced in the presentation on clustering. Again, we'll be looking at the measurements of the tissue samples treated as pieces of information in an unsupervised setting. Our intent here, thus, is to explore the different variable values and, in particular, the directions in which we might be able to explain the variability within the data set.

We open our libraries for organization and visuals; set our working directory to be able to access the data frame; and revise this to include the 30 different characteristics, excluding ID. Additionally, we'll take a look at a subset of those characteristics that are the mean of the measurements of that particular characteristic only and the extreme measurements.

Each of those should contain 10 such measurements. So if we take a look at p, that is all 30 measurements. p.means are the 10 mean measurements. And p.extremes are the 10 extreme measurements.

As we've talked about before-- that is, in the clustering example-- if we take a look at the values of these different variables, these different measurements, they are on very different scales. And thus, when we fit principal components analysis, we are going to consider both centering and scaling of these measurements.

We do this by inside the prcomp function of our matrix, x. We use the center=T and the scale=T arguments, and then we'll store this in pc.info. We're going to take a look at some of the values, or components, that are stored in pc.info. First thing that we're going to take a look at is the summary of pc.info.

And what this shows us is a description of the importance of these different components. In particular, we can peel off different pieces, including the standard deviation, the proportion of variance explained, and the cumulative proportion. So if I took a look just at the importance taken from this summary-- so that's the $importance. If I wanted the cumulative proportion, I could just take the third row of this matrix, which shows the cumulative proportion of variance explained.

A plot of pc.info will display the variances. So that is, when I take a look at the values of the standard deviations, it will plot the standard deviation squared, which gives me the variances. This can help us visually see the cumulative effect of the additional variance explained by the different components.

When taking a look at the proportion of variance explained and the amount added on, it's important to recognize that that is capped at-- the meaningful number is capped at the min of n minus 1 and p. Since our n is quite large, this is going to be capped at p. And so that is the number of principal components that will, additionally, be able to explain more variance.

We can also compute these directly, if we wanted to, based off the standard deviation. So if I took the standard deviation from off my output, I computed the variances and then the proportional variance and then use the cumsum some function to add those up to accumulate them, I would get the CumulativePVE.

So we can see that, indeed, the 30th component does add on some additional information about the variance.

We could also, as noted before, just get this directly from the output of the summary function by taking a look at the importance information and the third row of that. Now, these values are rounded to five decimal places, so they'll give a little bit less accuracy. But for the purposes of what they'll need them, that is sufficiently accurate.

And so we can take a look at a plot of the CumulativePVE across the principal component number. Plotting in this way is something known as a scree plot. Scree refers to the connectivity between the points in a way of visualizing breaks, or elbows, in the pattern.

After the first 1, 2, we see a little bit of an elbow. And certainly, after 3, 4, 5, 6, we definitely see an elbow. And since at 7 principal components, we're explaining over 90% of the variability, that seems to be a pretty good place. 6 or 7-- 6 being the elbow point, and 7 getting us above 90%.

The last part of the output and visualizations that we're going to be using help us take a look at the loadings for the components. And here, when I take a look at the rotation part of the output, these are the loadings for the principal components. This is a very large matrix. It's going to be a 30 by 30 matrix, and so there's quite a lot of information in there.

So let's just take a look at the loadings. That is, the coefficients for the first pc. And so this is the way the different original variables' measurements are combined to give us the first pc.

And this can be a way to take a look at if this appears to be more of a particular component. It's not obvious here that there's any one component or grouping of components that stands out, but we may be able to identify some more significant components in there. In the pc1scores, these are the values for our various observations in that pc direction 1 and pc2scores in pc direction 2.

So let's take a look at both these put together in one plot called a biplot. In this biplot-- and this one looks rather messy because we have all 500-some observations denoted by its number of observations. And those observations are plotted as pc values 1 and pc value 2.

So, for example, if I was looking at observation 79-- looks like it has-- it's the ordered pair approximately negative 12 and almost negative 5. So if I take a look at the pc1 and pc2scores-- 79-- it's about a negative 12 and-- almost negative 12, negative 5.

The other part of this is taking a look at the loading. So the loadings we can think of as these directional arrows. And again, it's a little bit difficult to see, and that's why we're going to take a look at this revisited for a subset of the variables. But Dimension.mean has an arrow that ends right about here. And it looks like this other set of coordinate axes is what we have to look at for this arrow.

And it looks like about a negative not quite 0.1 and a negative 0.3. So if I take a look at my pc.loadings and make sure I've defined all my pc.loadings, I get about-- well, about a negative 0.06 and a negative 0.37.

So let's take a look at this for a smaller example. And by a smaller example, I mean instead of taking a look at all 30 variables, let's start with 10 variables, the 10 mean measurements. So I'm going to run principal components on this matrix with 10 variables in it. And since I've stored this in pc.info, I can use all my previous summary commands and take a look.

So now this one displ-- it looks like I really only need about two of these, perhaps. And when I take a look at my scree plot, there definitely is an elbow there. Taking a look at proportion-- over 90%-- maybe three or four. But certainly no more than four are really needed.

So visualizing this, again, in our biplot-- Now, this is a bit visually easier to see in terms of the red lines are not quite so overlapped. So when I take a look at Dimension.mean, my direction here looks to be about negative 0.07 and negative 0.48. So if I took a look at the pc.loadings-- negative 0.07 and negative 0.58-- does not appear to visually arrow all the way out to the Dimension.mean.

So Dimension.mean is the location. That is, about negative 0.07 and negative 0.58. The location of that name is where we should be looking for those coordinates.

If we wanted to look at the score for one of our observations, for 79, again, we see a similar sort of coordinate as we recognize before at 79-- negative 8 and about negative 3.7.

If we take a look at the loadings for our first principal component, there aren't any dramatic differences among the loadings other than to say dimension mean-- or the dimension does not appear to be very significant to the loadings. But the other measurements do appear to be reasonably equally included.

For pc.loadings2, taking a look at the loadings here, dimension seems to be very important to this one and a little bit less so Symmetry, Smoothness, Radius-- but Radius in the opposite direction of how it was included up here.

So there aren't clear contextual patterns that would have us describe what is the meaning of these principal components in this case. But we do want to explore these loadings to try and identify which of the various original variables contribute most or, perhaps, at least meaningfully, to the different principal components.

As always, thank you for your attention, and enjoy working with principal components.

Notes:
See the online course for a downloadable R file containing the set of commands used in this demonstration.

Summary

We use PCA to project the data into a lower-dimensional space (determined by the principal components). The goal is to better understand the behavior of data (in terms of variability) and/or to identify the most relevant variables in a data set.

- Identify directions in feature space that explain the greatest amount of variation.
- Summarize the information provided via PVE, and use it to select appropriate number of components.
- Use principal components to interpret and visualize behavior in data.