Contents

A reference for much of this material can be found in Härdle and Simar [2015], chapters 8 and 9.

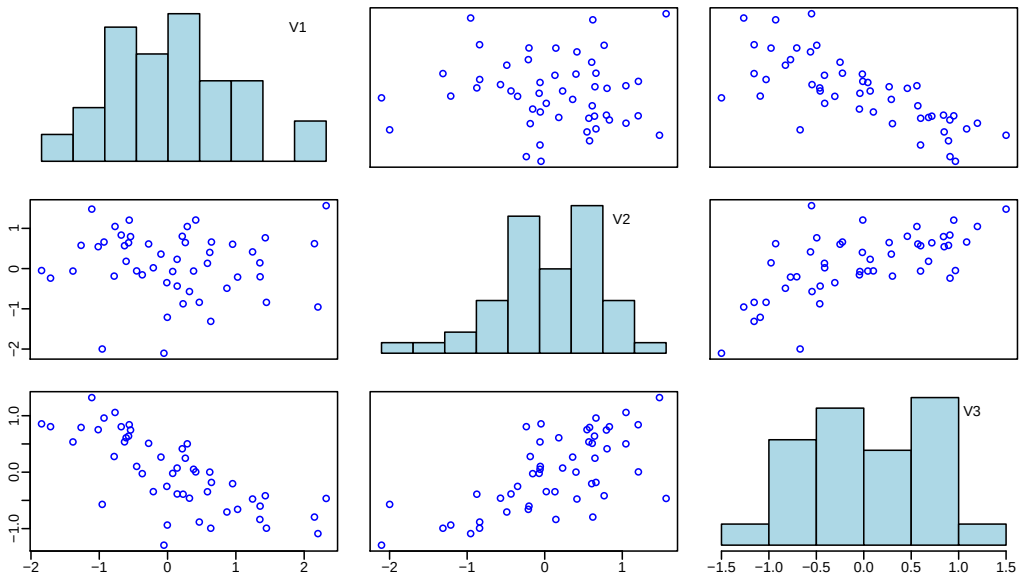


Figure 1: Scatterplots of 3 synthetically generated variables for 50 units.

1 Motivation

Let $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be a data matrix, with rows corresponding to *cases* or *units* and columns corresponding to *features* or *variables*. How can we visualize such data? A standard approach is with pairwise scatterplots. Figure 1 displays pairwise scatterplots for three variables of a synthetic dataset. From such a set of plots, we can get a sense of the univariate distributions of the variables and pairwise joint distributions.

Less apparent from such plots are how the data look in p dimensions. Figure 2 indicates that differences among the n points in $\mathbb{R}^3$ can be well-described by differences in a two-dimensional plane. Given a data matrix $\mathbf{Y}$, how can we find such a plane? How much of the variability among the n rows of $\mathbf{Y}$ occurs in this plane?

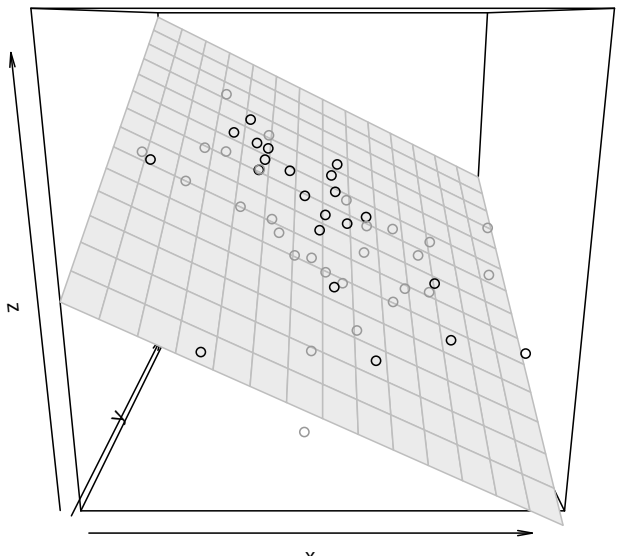


Figure 2: Three dimensional plot of synthetic data.

2 Linear projection of the rows

We'll begin by describing the data variation along a one-dimensional line.

Definition 1. *Let* $\mathbf{v} \in \mathbb{R}^p$ *be a unit vector, so that* $\mathbf{v}^\top\mathbf{v} = 1$. *The projection of* $\mathbf{y} \in \mathbb{R}^p$ *onto* $\mathbf{v}$ *is the vector* $\hat{\mathbf{y}} = f\mathbf{v}$, *where* $f = \mathbf{y}^\top\mathbf{v}$.

The projection $\hat{\mathbf{y}}$ is the best approximation of $\mathbf{y}$ along the linear subspace spanned by $\mathbf{v}$, in a least-squares sense:

Exercise 1. *Show that* $\mathbf{y}^\top\mathbf{v} = \arg\min_f ||\mathbf{y} - f\mathbf{v}||^2$.

We can interpret $\hat{\mathbf{y}}$ in multiple ways:

$$\hat{\mathbf{y}} = \mathbf{v}(\mathbf{v}^\top\mathbf{y}) = (\mathbf{v}\mathbf{v}^\top)\mathbf{y}.$$

The first representation gives $\hat{\mathbf{y}}$ as a vector times a scalar. The second gives $\hat{\mathbf{y}}$ as a rank-one $p \times p$ matrix times a vector.

Now let

- $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be a data matrix with rows equal to $\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top$;
- $\hat{\mathbf{Y}} \in \mathbb{R}^{n \times p}$ have rows equal to $\hat{\mathbf{y}}_1^\top, \ldots, \hat{\mathbf{y}}_n^\top$, with $\hat{\mathbf{y}}_i = (\mathbf{v}\mathbf{v}^\top)\mathbf{y}_i$.

Exercise 2. *Show that* $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{v}\mathbf{v}^\top$.

Comments:

1. The $p \times p$ matrix $\mathbf{v}\mathbf{v}^\top$ is a rank-1 projection matrix, projecting the p-dimensional rows of $\mathbf{Y}$ onto a 1-dimensional linear subspace spanned by $\mathbf{v}$. This space may be represented by the matrix $\mathbf{v}\mathbf{v}^\top$ (a point on the Grassmann manifold).
   - $\dim(\text{rows}(\mathbf{Y})) = p$
   - $\dim(\text{rows}(\hat{\mathbf{Y}})) = 1$
2. The matrix $\mathbf{v}\mathbf{v}^\top$ is idempotent:

$$\begin{aligned}[\mathbf{v}\mathbf{v}^\top][\mathbf{v}\mathbf{v}^\top] &= \mathbf{v}(\mathbf{v}^\top\mathbf{v})\mathbf{v}^\top \\ &= \mathbf{v}\mathbf{v}^\top\end{aligned}$$

   This should be intuitive: $\hat{\mathbf{Y}}$ is the projection of $\mathbf{Y}$ onto the space $\mathbf{v}\mathbf{v}^\top$. Projecting again should have no effect, so we expect $\hat{\mathbf{Y}}\mathbf{v}\mathbf{v}^\top = \hat{\mathbf{Y}}$.

3. Let $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ be the matrix of residuals. Then

$$\mathbf{R} = \mathbf{Y} - \mathbf{Y}\mathbf{v}\mathbf{v}^\top = \mathbf{Y}(\mathbf{I} - \mathbf{v}\mathbf{v}^\top).$$

   $\mathbf{R}$ is the projection of $\mathbf{Y}$ onto the $p-1$ dimensional orthogonal complement of the space $\mathbf{v}\mathbf{v}^\top$. The residuals are orthogonal to the fitted values, in the sense that $\hat{\mathbf{y}}_i^\top \mathbf{r}_i = 0$.

Exercise 3. *Show with a picture and mathematically that* $\hat{\mathbf{Y}}\mathbf{R}^\top = \mathbf{0}_{n\times n}$.

Exercise 4. *Show that the matrix* $(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)$ *is idempotent.*

---

3 Least squares and standardized linear combinations

Question: What 1-d subspace gives the best approximation of $\mathbf{Y}$?

Question: What is $\arg\min_{\mathbf{v}} ||\mathbf{Y} - \mathbf{Y}\mathbf{v}\mathbf{v}^\top||^2$?

First we will prove the following:

Lemma 1. $||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y}||^2 - ||\hat{\mathbf{Y}}||^2$.

This is a consequence of $\mathbf{v}\mathbf{v}^\top$, or $(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)$, being idempotent. Here is the

messy derivation:

$$\begin{aligned}
||\mathbf{Y}-\hat{\mathbf{Y}}||^2 &= \sum_i \sum_j (y_{i,j}-\hat{y}_{i,j})^2 \\
&= \sum_i ||\mathbf{y}_i-\hat{\mathbf{y}}_i||^2 \\
&= \sum_i (\mathbf{y}_i-\hat{\mathbf{y}}_i)^\top(\mathbf{y}_i-\hat{\mathbf{y}}_i) \\
&= \sum_i ||\mathbf{y}_i||^2 - 2\mathbf{y}_i^\top\hat{\mathbf{y}}_i + ||\hat{\mathbf{y}}_i||^2 \\
&= \sum_i \left(||\mathbf{y}_i||^2 - ||\hat{\mathbf{y}}_i||^2\right) = ||\mathbf{Y}||^2 - ||\hat{\mathbf{Y}}||^2,
\end{aligned}$$

where $\mathbf{y}_i^\top\hat{\mathbf{y}}_i = ||\hat{\mathbf{y}}_i||^2$ because $\mathbf{v}\mathbf{v}^\top$ is idempotent.

In this class we will try to avoid summation notation and writing out indices as much as possible. To this end, we now introduce the trace operation, and illustrate its use by providing an alternative derivation of the lemma.

Definition 2. *The trace of a square matrix is the sum of its diagonal elements, so that for* $\mathbf{A} \in \mathbb{R}^{m\times m}$, $tr(\mathbf{A}) = \sum_i a_{ii}$.

Several properties of the trace operator follow immediately from the definition. For square matrices $\mathbf{A}$ and $\mathbf{B}$,

- $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^\top)$;
- $\text{tr}(\mathbf{A}+\mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.

For square or non-square matrices,

- $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$;
- $\text{tr}(\mathbf{XX}^\top) = \text{tr}(\mathbf{X}^\top\mathbf{X}) = \sum_i \sum_j x_{ij}^2 = ||\mathbf{X}||^2$.

The following identity is so useful we will call it a lemma:

Lemma 2. $||\mathbf{x}||^2 = \mathbf{x}^\top\mathbf{x} = tr(\mathbf{x}\mathbf{x}^\top)$.

Now let's return to proving the identity in Lemma 1

$$\begin{aligned} ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 &= ||\mathbf{Y}(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)||^2 \\ &= \text{tr}(\mathbf{Y}^\top(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)^\top(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)\mathbf{Y}) \\ &= \text{tr}(\mathbf{Y}^\top(\mathbf{I} - \mathbf{v}\mathbf{v}^\top)\mathbf{Y}||^2 \\ &= \text{tr}(\mathbf{Y}^\top\mathbf{Y}) - \text{tr}(\mathbf{Y}^\top\mathbf{v}\mathbf{v}^\top\mathbf{Y}) \\ &= ||\mathbf{Y}||^2 - ||\hat{\mathbf{Y}}||^2. \end{aligned}$$

There, that's prettier.

Exercise 5. *Prove Lemma 1 using the fact that* $\mathbf{Y}\mathbf{R}^\top = 0$.

Now recall that we are trying to find the unit vector $\mathbf{v}$, or the projection matrix $\mathbf{v}\mathbf{v}^\top$, so that $||\mathbf{R}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is minimized. The lemma implies

$$\begin{aligned} \arg\min_{\mathbf{v}} ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 &= \arg\min_{\mathbf{v}} \left(||\mathbf{Y}||^2 - ||\hat{\mathbf{Y}}||^2 \right) \\ &= \arg\max_{\mathbf{v}} ||\hat{\mathbf{Y}}||^2. \end{aligned}$$

So an alternative description of what we are trying to do is to find the $\mathbf{v}$ for which the projection of $\mathbf{Y}$ onto $\mathbf{v}\mathbf{v}^\top$ is largest. We will now derive a more "data-oriented" interpretation. Notice that

$$\begin{aligned} ||\hat{\mathbf{Y}}||^2 &= \text{tr}(\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}) \\ &= \text{tr}(\mathbf{v}\mathbf{v}^\top\mathbf{Y}\mathbf{Y}\mathbf{v}\mathbf{v}^\top) \\ &= \text{tr}(\mathbf{v}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{v}) \\ &= \mathbf{v}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{v} = ||\mathbf{Y}\mathbf{v}||^2. \end{aligned}$$

Let $\mathbf{f} = \mathbf{Y}\mathbf{v}$, an n-dimensional vector with ith element given by $f_i = \mathbf{y}_i^\top\mathbf{v}$.

- f_i is the magnitude of the projection of $\mathbf{y}_i$ onto the space $\mathbf{v}\mathbf{v}^\top$.
- f_i is a standardized linear combination of the p variables of $\mathbf{y}_i$. The vector $\mathbf{f} = \mathbf{Y}\mathbf{v} \in \mathbb{R}^n$ is a derived variable for the n units of $\mathbf{Y}$.
- Minimizing $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is equivalent to
  - maximizing $||\hat{\mathbf{Y}}||^2$
  - maximizing $||\mathbf{Y}\mathbf{v}||^2$
  - maximizing $\mathbf{f}^\top\mathbf{f}/n \approx \text{Var}[(f_1, \ldots, f_n)]$

So the problem of approximating the rows of $\mathbf{Y} \in \mathbb{R}^{n\times p}$ by points in a 1-dimensional subspace of $\mathbb{R}^p$ is equivalent to finding the standardized linear combination (SLC) $\mathbf{f} = \mathbf{Y}\mathbf{v}$ with maximum variance.

Exercise 6. *Show that if the sample mean of each column of* $\mathbf{Y}$ *is zero, then the sample mean of a SLC* $\mathbf{f}$ *is also zero. Don't use any summation notation.*

4 Best one-dimensional approximation

Now let's find the optimal linear subspace, or equivalently, the SLC with maximum variance. This is done by finding the $\mathbf{v}$ that maximizes $\mathbf{v}^\top \mathbf{Y}^\top \mathbf{Y}\mathbf{v}$ subject to $\mathbf{v}^\top\mathbf{v} = 1$.

Let's write $\mathbf{S} = \mathbf{Y}^\top\mathbf{Y}$. Why? If the columns of $\mathbf{Y}$ are centered, then

$$\mathbf{S} = \mathbf{Y}^\top\mathbf{Y} = \sum_i \mathbf{y}_i\mathbf{y}_i^\top = (n-1) \times \text{sample covariance matrix.}$$

Task: Maximize $\mathbf{v}^\top\mathbf{S}\mathbf{v}$ subject to $\mathbf{v}^\top\mathbf{v} = 1$.

Method: Lagrange multipliers.

Procedure: A critical point of $\mathbf{v}^\top\mathbf{S}\mathbf{v}$ along the level set $\mathbf{v}^\top\mathbf{v} = 1$ satisfies

$$\frac{d}{d\mathbf{v}}\left(\mathbf{v}^\top\mathbf{S}\mathbf{v} + \lambda(1 - \mathbf{v}^\top\mathbf{v})\right) = 0.$$

Solution: You need to know that for a matrix $\mathbf{A}$, the gradient of $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ is $2\mathbf{A}\mathbf{x}$. It will also soon be helpful to know that the Hessian of $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ is $2\mathbf{A}$. Using this information, it follows that an optimal $\mathbf{v}$ satisfies

$$\begin{aligned}\frac{d}{d\mathbf{v}}\left(\mathbf{v}^\top\mathbf{S}\mathbf{v} + \lambda(1 - \mathbf{v}^\top\mathbf{v})\right) = 2\mathbf{S}\mathbf{v} - 2\lambda\mathbf{v} &= 0 \\ \mathbf{S}\mathbf{v} &= \lambda\mathbf{v}.\end{aligned}$$

In other words, critical points of $\mathbf{v}^\top\mathbf{S}\mathbf{v}$ are unit eigenvectors of $\mathbf{S}$.

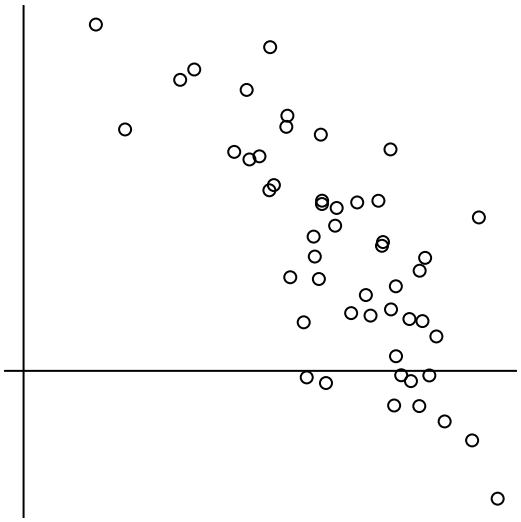
So far we have shown that the critical points of $\mathbf{v}^\top\mathbf{S}\mathbf{v}$ are the eigenvectors of $\mathbf{S}$. Which one maximizes $\mathbf{v}^\top\mathbf{S}\mathbf{v}$? Well, if $\mathbf{v}$ is an eigenvector of $\mathbf{S}$ with eigenvalue λ, then

$$\begin{aligned}\mathbf{v}^\top\mathbf{S}\mathbf{v} &= \mathbf{v}^\top(\lambda\mathbf{v}) \\ &= \lambda\mathbf{v}^\top\mathbf{v} = \lambda.\end{aligned}$$

So the maximizer of $\mathbf{v}^\top\mathbf{S}\mathbf{v}$ is the eigenvector with the largest eigenvalue.

Exercise 7. *Show that all eigenvalues of* $\mathbf{S}$ *are non-negative.*

Let's try out a numerical example. Here is the scatterplot for a 50×2 data matrix $\mathbf{Y}$:



The best approximating one-dimensional linear subspace $\mathbf{v}\mathbf{v}^\top$, and the projection $\mathbf{Y}\mathbf{v}\mathbf{v}^\top$, can be obtained in R as follows:

```
v<-eigen( t(Y)%*%Y )$vec[,1]
v
```

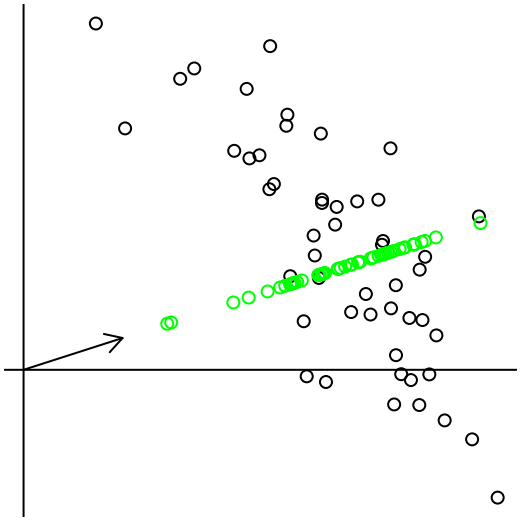
```
## [1] -0.964 -0.266


Yhat<- Y%*%v%*%t(v)
```

Let's check some things we've calculated analytically:

```
eigen( t(Y)%*%Y )$val[1]

## [1] 544


sum( (Y%*%v)^2 )

## [1] 544
```

```
sum( (Y-Yhat)^2 )

## [1] 58.1


sum(Y^2) - sum(Yhat^2)

## [1] 58.1
```

So far so good. Now let's look at our approximation:



This is mathematically correct, but not quite what we had in mind. Most datasets in their raw form have non-zero (sample) means for each variable, and typically the difference between this multivariate mean vector and the origin is much larger than the variation of the data around the mean vector. Since the approximating linear subspace is a subspace and so necessarily contains the origin, the largest eigenvector of $\mathbf{Y}^\top\mathbf{Y}$ will generally be pointing in a direction close to that of the vector of sample means.

Intuitively, what we want to do is first subtract off the sample means from each column, then find the best linear subspace.

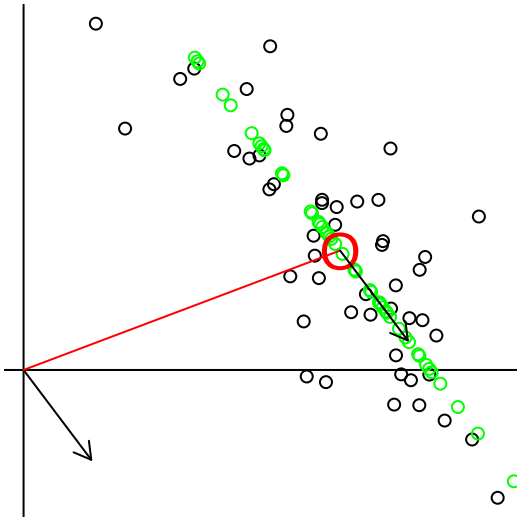
```
ybar <- apply(Y,2,mean) ; S<-crossprod( sweep(Y,2,ybar)  )
v<-eigen( S )$vec[,1]
v

## [1] -0.659  0.752
```

This $\mathbf{v}$ is pointing in a different direction than the first $\mathbf{v}$. The subspace $\mathbf{v}\mathbf{v}^\top$ gives the best one-dimensional description of the variation of the demeaned

data $\mathbf{Y} - \mathbf{1}\bar{\mathbf{y}}^\top$ around the origin, or equivalently of the raw data $\mathbf{Y}$ around the mean $\bar{\mathbf{y}}$.

Our approximation to $\mathbf{Y}$ is that each $\mathbf{y}_i$ is equal to $\bar{\mathbf{y}}$ plus the projection of $\mathbf{y}_i - \bar{\mathbf{y}}$ along the first eigenvector of the centered covariance matrix.

```
Yhat<- rep(1,n)%*%t(ybar)  + sweep(Y,2,ybar)%*%v%*%t(v)
```



This looks much better. Now let's go through everything symbolically.

The centered data are

$$\tilde{\mathbf{Y}} = \begin{pmatrix} y_{11} - \bar{y}_1 & \cdots & y_{1p} - \bar{y}_p \\ \vdots & & \vdots \\ y_{n1} - \bar{y}_1 & \cdots & y_{np} - \bar{y}_p \end{pmatrix} = \mathbf{Y} - \begin{pmatrix} \bar{y}_1 & \cdots & \bar{y}_p \\ \vdots & & \vdots \\ \bar{y}_1 & \cdots & \bar{y}_p \end{pmatrix} = \mathbf{Y} - \mathbf{1}\bar{\mathbf{y}}^\top.$$

You should become very familiar with the following:

$$\begin{aligned}\bar{\mathbf{y}} &= \mathbf{Y}^\top \mathbf{1}/n \\ \bar{\mathbf{y}}^\top &= \mathbf{1}^\top \mathbf{Y}/n \\ \mathbf{Y} - \mathbf{1}\bar{\mathbf{y}}^\top &= \mathbf{Y} - \mathbf{1}\mathbf{1}^\top \mathbf{Y}/n \\ &= (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/n)\mathbf{Y}.\end{aligned}$$

We refer to $\mathbf{C} = (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/n)$ as the n-dimensional centering matrix.

Question: What happens if you center data that is already centered?

Answer: Nothing.

Question: Do $\mathbf{CY}$ and $\mathbf{CCY}$ differ? What is $\mathbf{CC}$?

Exercise 8. *Show that* $\mathbf{C}$ *is idempotent.*

Comments:

- The linear subspace of $\mathbb{R}^n$ spanned by the unit vector $\mathbf{1}/\sqrt{n}$ corresponds to the rank-1 idempotent projection matrix $\mathbf{1}\mathbf{1}^\top/n$.
- The projection of *the columns* of $\mathbf{Y}$ onto this space is $\mathbf{1}\bar{\mathbf{y}}^\top$. Here, we are thinking of the p columns of $\mathbf{Y}$ as p points in n-dimensional space, and approximating these p-points by $c_j\mathbf{1}/\sqrt{n}$ for $j = 1, \ldots p$. Of course, the best approximation in a least squares sense is $c_j = \bar{y}_j$.
- The matrix $\mathbf{C} = \mathbf{I} - \mathbf{1}\mathbf{1}^\top/n$ is a rank $n-1$ idempotent projection matrix.
- The space defined by $\mathbf{C}$ is the orthogonal complement to the linear space spanned by the unit vector $\mathbf{1}/\sqrt{n}$.

It may be useful to write

$$\begin{aligned}\mathbf{Y} &= (\mathbf{1}\mathbf{1}^\top/n + \mathbf{I} - \mathbf{1}\mathbf{1}^\top/n)\mathbf{Y} \\ &= \mathbf{1}\mathbf{1}^\top\mathbf{Y}/n + (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/n)\mathbf{Y} \\ &= \mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{C}\mathbf{Y}\end{aligned}$$

This provides a decomposition of $\mathbf{Y}$ into complementary orthogonal parts: the mean, and variation around the mean.

This decomposition indicates that we can think of the total variation of the data around the origin as being the variation of the mean around the origin plus the variation of the data around the mean:

$$\begin{aligned}||\mathbf{Y}^2|| &= ||\mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{C}\mathbf{Y}||^2 \\ &= \text{tr}((\mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{C}\mathbf{Y})^\top(\mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{C}\mathbf{Y})) \\ &= \text{tr}(\bar{\mathbf{y}}\mathbf{1}^\top\mathbf{1}\bar{\mathbf{y}}^\top) + 2\text{tr}(\mathbf{Y}^\top\mathbf{C}^\top\mathbf{1}\bar{\mathbf{y}}) + \text{tr}(\mathbf{Y}^\top\mathbf{C}^\top\mathbf{C}\mathbf{Y}) \\ &= n||\bar{\mathbf{y}}||^2 + ||\mathbf{C}\mathbf{Y}||^2\end{aligned}$$

Our two-step approximation of $\mathbf{Y}$ is

$$\mathbf{Y} \approx \mathbf{1}\bar{\mathbf{y}}^\top + (\mathbf{C}\mathbf{Y})\mathbf{v}\mathbf{v}^\top$$

where $\mathbf{v}$ is the first eigenvector of $\mathbf{S} = \mathbf{Y}^\top\mathbf{C}^\top\mathbf{C}\mathbf{Y} = \sum_i(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top$. Since the eigenvectors of $\mathbf{S}$ and $c\mathbf{S}$ are the same, we see that $\mathbf{v}$ is the first eigenvector of the sample covariance matrix $\mathbf{S}/(n-p)$.

It is useful to examine how this approximation differs from approximating the rows of $\mathbf{Y}$ by points in a 1-dimensional linear subspace. Here is the linear subspace approximation:

```
v<-eigen( t(Y)%*%Y )$vec[,1]
f<-Y%*%v
mean(f^2)

## [1] 10.9


var(f)

##       [,1]
## [1,] 0.413
```

The magnitude of $\mathbf{f} = \mathbf{Yv}$ is large, but this is just the mean being different from zero. The variability of $\mathbf{f}$ is small compared to its magnitude. Contrast this with our two-step approximation:

```
ybar <- apply(Y,2,mean) ; S<-crossprod( sweep(Y,2,ybar)  )
v<-eigen( S )$vec[,1]
f<-sweep(Y,2,ybar)%*%v
mean(f^2)

## [1] 1.37


var(f)

##      [,1]
## [1,] 1.39


sum(f^2)/(n-1)

## [1] 1.39
```

Now $\mathbf{f} = \mathbf{CYv}$ is representing variation of the datapoints around the mean vector. The magnitude of $\mathbf{f}$ corresponds to its variance: This is because $\mathbf{f}$ is mean-zero.

Our approximation is called an affine approximation: We are approximating the rows of $\mathbf{Y}$ by points in a common 1-dimensional affine space. You can think of an affine space as a point in space plus a linear subspace. Points in a one-dimensional affine subspace are points on a line, and are equal to

- a common point in $\mathbb{R}^p$ plus
- scalar multiples of a common vector in $\mathbb{R}^p$.

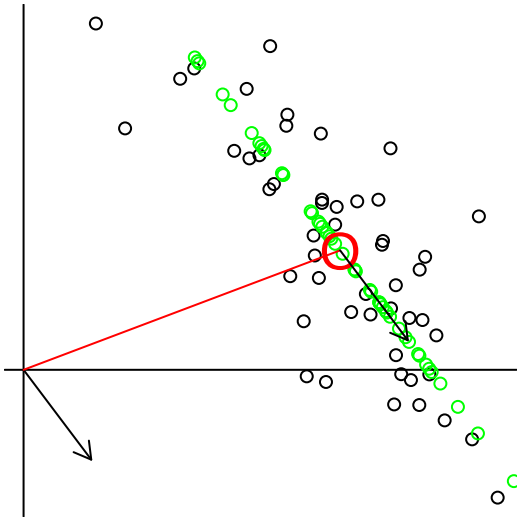
To see that our approximation is of this form, note that for a given row $\mathbf{y}_i$ of $\mathbf{Y}$, our approximation is

$$\mathbf{y}_i = \bar{\mathbf{y}} + ((\mathbf{y}_i - \bar{\mathbf{y}})^\top \mathbf{v})\mathbf{v}.$$

So we are approximating each row $\mathbf{y}_i$ of $\mathbf{Y}$ by

- a common point in $\mathbb{R}^p$ ($\bar{\mathbf{y}}$) plus
- a scalar multiple ($(\mathbf{y}_i - \bar{\mathbf{y}})^\top \mathbf{v}$) of a common vector in $\mathbb{R}^p$ ($\mathbf{v}$).

Thus the approximated rows of $\mathbf{Y}$ lie in a common 1-dimensional affine space (a line). Let's return to the picture:



Our approximation selects a particular affine subspace, obtained from the sample mean $\bar{\mathbf{y}}$ and the first eigenvector of the sample (centered) sum of squares matrix $\mathbf{S} = \mathbf{Y}^\top \mathbf{C}\mathbf{Y}$. Is this the best affine subspace? We have the following result:

Theorem 1. *Let* $\hat{\mathbf{Y}} = \mathbf{1}\boldsymbol{\mu}^\top + \mathbf{f}\mathbf{v}^\top$. *Then* $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ *is minimized in* $(\boldsymbol{\mu}, \mathbf{f}, \mathbf{v})$ *by* $(\bar{\mathbf{y}}, \mathbf{f}_1, \mathbf{v}_1)$, *where*

$$\mathbf{v}_1 = evec_1(\mathbf{Y}^\top \mathbf{C}\mathbf{Y})$$
$$\mathbf{f}_1 = \mathbf{C}\mathbf{Y}\mathbf{v}_1.$$

Exercise 9. *Prove Theorem 1.*

Matrix approximation and variance representation

Now let's recall what we learned earlier: The SLC $\mathbf{f} = \mathbf{Y}\mathbf{v}$ that maximizes $\mathbf{f}^\top\mathbf{f}$ is $\mathbf{f}_1 = \mathbf{Y}\mathbf{v}_1$, where $\mathbf{v}_1$ is the first eigenvector of $\mathbf{Y}^\top\mathbf{Y}$.

Similarly, let $\mathbf{CY}$ be a centered data matrix. Then the SLC $\mathbf{f} = \mathbf{CYv}$ that maximizes $\mathbf{f}^\top\mathbf{f} = (n-1) \times \text{Var}[\mathbf{f}]$ is $\mathbf{f}_1 = \mathbf{CY}\mathbf{v}_1$, where $\mathbf{v}_1$ is the first eigenvector of $\mathbf{Y}^\top\mathbf{C}^\top\mathbf{CY} = \mathbf{Y}^\top\mathbf{CY} = \mathbf{S}$. Thus the $\mathbf{f}$ that gives the best one-dimensional representation of $\mathbf{Y}$ is also the SLC of the centered data that maximizes the variation. This result provides a link between matrix approximation and variance representation.

To summarize: Let $\mathbf{CY}$ be a centered data matrix and $\mathbf{v}$ a unit vector so that $\mathbf{v}^\top\mathbf{v} = 1$. Then

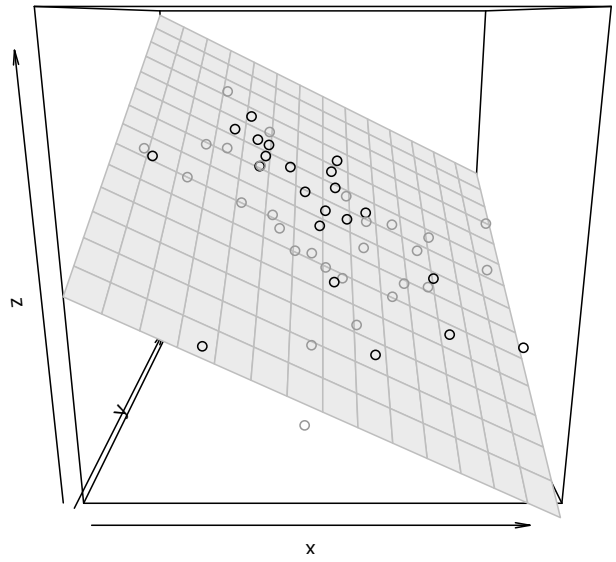
1. $\mathbf{f} = \mathbf{CYv}$ is a mean-zero derived variable.
2. $\mathbf{f}^\top\mathbf{f} = \mathbf{v}^\top\mathbf{Y}^\top\mathbf{CYv} = \mathbf{v}^\top\mathbf{Sv}$, and so $\text{Var}[\mathbf{f}] = \mathbf{v}^\top\mathbf{Sv}/(n-1)$.
3. The $\mathbf{v}$ for which $\text{Var}[\mathbf{f}]$ is maximized is $\mathbf{v}_1 = \text{evec}_1(\mathbf{S})$, giving a variance of $\text{Var}[\mathbf{f}_1] = \lambda_1/(n-1)$.

Much of multivariate analysis is more concerned with variance representation and features of the sample sum of squares $\mathbf{S}$, and less with the problem of approximating a data matrix $\mathbf{Y}$. From this perspective, the interesting results so far are the following:

- the vector $\mathbf{v}_1 = \text{evec}_1(\mathbf{S})$ gives the axis of maximal data variation (around the mean) and is called the *first principal axis*;
- the derived variable $\mathbf{f}_1 = \mathbf{CY}\mathbf{v}_1$ is the SLC with maximum (sample) variance, and is called the *first principal component*.

5 Higher-dimensional approximations

That we called $\mathbf{f}_1$ the first principal component suggests that there is a second, and maybe even a third. These additional principal components are of interest when most of the variation among the rows of $\mathbf{Y}$ occurs not in a 1-dimensional affine subspace, but possibly a subspace of higher dimension. For example, consider the synthetic data we saw earlier:



These 3-dimensional data are not well-represented by point on a line in three dimensions, but may be well represented by points on a plane (specifically, the plane in the figure). We can generalize this idea to higher dimensions:

1-d affine approximation : $\mathbf{y}_i \approx \bar{\mathbf{y}} + \mathbf{v}_1[\mathbf{v}_1^\top(\mathbf{y}_i - \bar{\mathbf{y}})]$

2-d affine approximation : $\mathbf{y}_i \approx \bar{\mathbf{y}} + \mathbf{v}_1[\mathbf{v}_1^\top(\mathbf{y}_i - \bar{\mathbf{y}})] + \mathbf{v}_2[\mathbf{v}_2^\top(\mathbf{y}_i - \bar{\mathbf{y}})]$

q-d affine approximation :

$$\mathbf{y}_i \approx \bar{\mathbf{y}} + \sum_{j=1}^{q} \mathbf{v}_j[\mathbf{v}_j^\top(\mathbf{y}_i - \bar{\mathbf{y}})]$$
$$\mathbf{Y} \approx \mathbf{1}\bar{\mathbf{y}}^\top + (\mathbf{C}\mathbf{Y})\mathbf{V}\mathbf{V}^\top$$

How to find the best q-dimensional affine subspace? Not surprisingly, the best subspace is given by the first q eigenvectors of $\mathbf{S} = \mathbf{Y}^\top\mathbf{C}\mathbf{Y}$. To understand this, we need to know a bit more about eigenvectors.

Exercise 10. *Show that if* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *are eigenvectors of* $\mathbf{S}$ *with unequal eigenvalues* λ_1 *and* λ_2*, then* $\mathbf{v}_1^\top\mathbf{v}_2 = 0$*.*

Exercise 11. *If* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *are eigenvectors of* $\mathbf{S}$*, under what conditions can* $a_1\mathbf{v}_1 + a_2\mathbf{v}_2$ *be an eigenvector of* $\mathbf{S}$*?*

The result from Exercise 10 is important. From this result we can deduce the very useful eigendecomposition theorem:

Theorem 2. *Let* $\mathbf{S} \in \mathbb{R}^{p\times p}$ *be a symmetric matrix with* p *distinct eigenvalues* $\lambda_1 > \cdots > \lambda_p$*. Then*

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top,$$

where $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ *and* $\mathbf{V}$ *is orthogonal, so* $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$*.*

Proof. Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$, the unit eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_p$ in order. Then

$$\mathbf{S}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}$$
$$\mathbf{S}\mathbf{V}\mathbf{V}^\top = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$
$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

The first line holds because the columns of $\mathbf{V}$ are the eigenvectors. The third line holds because of the results from Exercise 10: if $\mathbf{v}_j^\top\mathbf{v}_k = 0$ then $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$. Since $(\mathbf{V}^{-1})^\top = (\mathbf{V}^\top)^{-1}$, we also have $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$. □

The representation also holds if some of the eigenvalues are equal. For example, if $\lambda_k = \lambda_{k+1}$, then we can take $\mathbf{v}_k$ and $\mathbf{v}_{k+1}$ to be an arbitrary orthogonal basis for the two-dimensional subspace spanned by the eigenvectors with this eigenvalue.

Some nice results follow quickly from the eigendecomposition:

- If $\mathbf{S}$ is invertible then $\mathbf{S}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top$
- The trace of $\mathbf{S}$ is $\mathrm{tr}(\mathbf{S}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \sum \lambda_j$

Exercise 12. *Derive the above two results.*

Exercise 13. *Illustrate the above two results numerically.*

Now let's return to the problem of finding the best q-dimensional affine subspace. Analogous to Theorem 1 we have the following:

Theorem 3. *Let* $\hat{\mathbf{Y}} = \mathbf{1}\boldsymbol{\mu}^\top + \mathbf{F}_q\mathbf{V}_q^\top$, *with* $\mathbf{F}_q \in \mathbb{R}^{n\times q}$ *and* $\mathbf{V}_q \in \mathbb{R}^{p\times q}$. *Then* $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ *is minimized by*

$$\begin{aligned}\boldsymbol{\mu} &= \bar{\mathbf{y}} \\ \mathbf{V}_q &= (\mathbf{v}_1, \ldots, \mathbf{v}_q), \text{ where } \mathbf{v}_j = evec_j(\mathbf{Y}^\top\mathbf{C}\mathbf{Y}) \\ \mathbf{F}_q &= \mathbf{C}\mathbf{Y}\mathbf{V}_q.\end{aligned}$$

A proof of this, and some related results, appear in Gabriel [1978]). A slightly different approach to the proof is to minimize the residual sum of squares sequentially, first given $\boldsymbol{\mu}$ and $\mathbf{V}_q$, then given $\boldsymbol{\mu}$, and then optimize over $\boldsymbol{\mu}$.

Accuracy of the approximation: How good is the approximation for each choice of q? Not surprisingly the approximation improves as q increases, and the approximation error is zero when $q = p$. Let's investigate this more precisely. We have

$$\begin{aligned} ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 &= ||(\mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{CY}) - (\mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{CY}\mathbf{V}_q\mathbf{V}_q^\top)||^2 \\ &= ||\mathbf{CY} - \mathbf{CY}\mathbf{V}_q\mathbf{V}_q^\top||^2 \\ &= ||\tilde{\mathbf{Y}}||^2 - \text{tr}(\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}\mathbf{V}_q\mathbf{V}_q^\top) \end{aligned}$$

where $\tilde{\mathbf{Y}} = \mathbf{CY}$. Now let $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ be the eigendecomposition of $\mathbf{S} = \tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}$. Then recalling that the columns of $\mathbf{V}$ are orthogonal, we have

$$\begin{aligned} ||\tilde{\mathbf{Y}}||^2 - \text{tr}(\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}\mathbf{V}_q\mathbf{V}_q^\top) &= \text{tr}(\mathbf{\Lambda}) - \text{tr}(\mathbf{\Lambda}[\mathbf{V}^\top\mathbf{V}_q\mathbf{V}_q^\top\mathbf{V}]) \\ &= \sum_{j=1}^{p}\lambda_j - \sum_{j=1}^{q}\lambda_j = \sum_{j=q+1}^{p}\lambda_j. \end{aligned}$$

Intuitively,

- the total variation of $\mathbf{Y}$ around its mean is $||\mathbf{CY}||^2 = \text{tr}(\mathbf{S}) = \sum_1^p \lambda_j$;
- the maximum variation representable in q-dimensions is $\sum_1^q \lambda_j$;
- the minimum variation unexplained in q-dimensions is $\sum_{q+1}^p \lambda_j$.

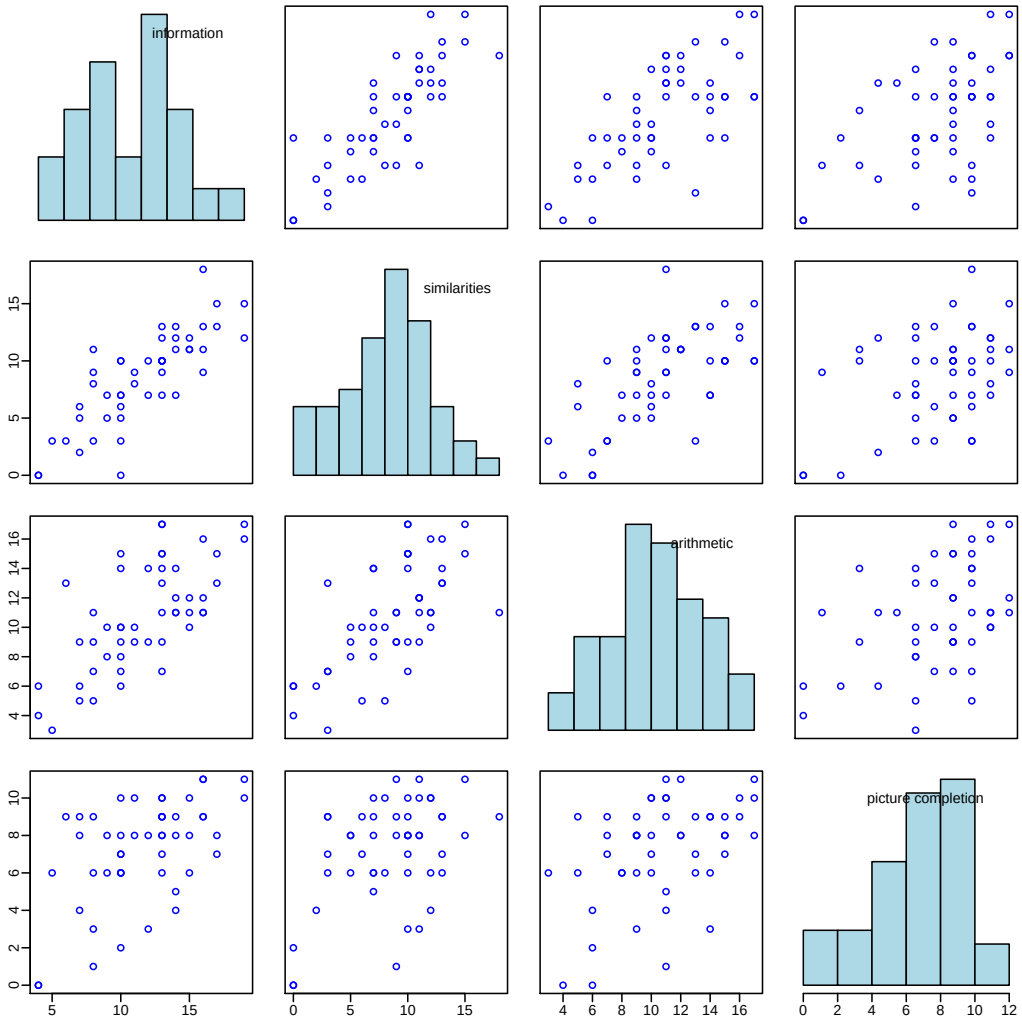
Let's look at a numerical example:

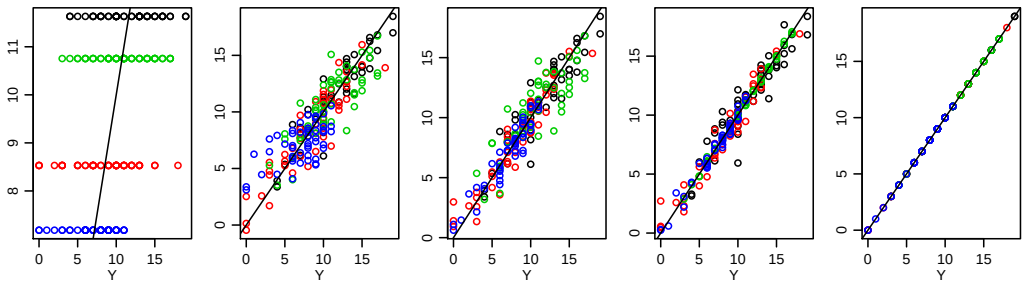
```
dim(wais)
```

```
## [1] 49  5
```

```
head(wais)
```

```
##      senile information similarities arithmetic picture completion
## [1,]      1           9            5         10                  8
## [2,]      1          10            0          6                  2
## [3,]      1           8            9         11                  1
## [4,]      1          13            7         14                  9
## [5,]      1           4            0          4                  0
## [6,]      1           4            0          6                  0
```






```
ERR<-sweep(Yq,c(1,2),Y,"-")

apply(ERR^2,3,sum)

## [1] 2.45e+03 6.32e+02 3.73e+02 1.23e+02 1.67e-28


rev( cumsum( rev( eigen(S)$val )  ) )

## [1] 2452  632  373  123
```

6 Principal components analysis

Let $\mathbf{S} = \mathbf{Y}^\top\mathbf{C}\mathbf{Y}$ be the sample sum of squares matrix. Let $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ be the eigendecomposition of $\mathbf{S}$, so that

$$\begin{aligned}\mathbf{V}^\top\mathbf{V} &= \mathbf{V}^\top\mathbf{V} = \mathbf{I} \\ \mathbf{\Lambda} &= \text{diag}(\lambda_1, \ldots, \lambda_p) \\ \mathbf{S}\mathbf{v}_j &= \lambda_j\mathbf{v}_j.\end{aligned}$$

Let $\mathbf{F} = \mathbf{C}\mathbf{Y}\mathbf{V}$, so the jth column of $\mathbf{F}$ is $\mathbf{f}_j = \mathbf{C}\mathbf{Y}\mathbf{v}_j$.

You should be able to show the following:

Theorem 4.

1. *The SLC* $\mathbf{f}_j$ *is zero mean and has (sample) variance* $\lambda_j/(n-1)$
2. *The SLCs* $\mathbf{f}_j$ *and* $\mathbf{f}_k$ *are uncorrelated.*
3. $\mathbf{f}_1$ *is the SLC with the maximum variance among all SLCs.*
4. $\mathbf{f}_p$ *is the SLC with the minimum variance among all SLCs.*
5. $\mathbf{f}_j$ *is the SLC with the maximum variance among all SLCs uncorrelated with* $\mathbf{f}_1, \ldots, \mathbf{f}_{j-1}$.

In matrix form, items 1 and 2 are

$$\bar{\mathbf{f}} = \mathbf{F}^\top \mathbf{1}/n = 0$$
$$\text{Var}[\mathbf{F}] = \mathbf{F}^\top \mathbf{F}/(n-p) = \mathbf{\Lambda}/(n-p).$$

Terminology:

- The axis determined by $\mathbf{v}_j$ is the *jth principal axis.*
- The jth column of $\mathbf{F}$ is $\mathbf{f}_j = \mathbf{CY}\mathbf{v}_j$, and is the *jth principal component.*

Principal components analysis (PCA) generally consists of computation and examination of $\mathbf{V}$, $\mathbf{\Lambda}$ and $\mathbf{F}$. The computation of $\mathbf{F}$ results from a linear transformation $\mathbf{CY} \rightarrow \mathbf{CYV} = \mathbf{F}$, which transforms the correlated mean-zero column variables of $\mathbf{CY}$ to the uncorrelated mean-zero column variables of $\mathbf{F}$. Why is this transformation of interest?

- Sometimes we imagine that the correlation among the observed variables is a result of them each being a linear combination of some unobserved independent latent variables. PCA hopes to recover these underlying variables.

- Sometimes the variability in a p-dimensional dataset can be well-described in $q \ll p$ dimensions. PCA can be used to obtain and compare the accuracy of low-dimensional representations.

We now illustrate these and other uses of PCA in the context of a few numerical examples.

6.1 WAIS data

The WAIS dataset consists of data from four subtests of the Wechsler Adult Intelligence Scale (WAIS) on 49 elderly individuals.

```
Y<-wais[,-1]
ybar<-apply(Y,2,mean)
ybar

##        information       similarities         arithmetic
##              11.63               8.53              10.76
## picture completion
##               7.18


cov(Y)

##                    information similarities arithmetic picture completion
## information              13.78        12.26       9.16               5.63
## similarities             12.26        16.63       9.61               5.03
## arithmetic                9.16         9.61      13.02               4.38
## picture completion        5.63         5.03       4.38               7.65
```

Let's do a PCA, that is, analyze the eigendecomposition of the sample covariance matrix.

```
CY<- sweep(Y,2,ybar,"-")
S<-crossprod(CY)
eS<-eigen(S)
V<-eS$vec
L<-eS$val
V
```

```
##         [,1]     [,2]   [,3]    [,4]
## [1,] -0.560 -0.00821  0.233  0.7954
## [2,] -0.609 -0.48248  0.337 -0.5319
## [3,] -0.490  0.11998 -0.859 -0.0922
## [4,] -0.276  0.86761  0.308 -0.2755
```

```
L
```

```
## [1] 1820  260  249  123
```

Notice that the entries of $\mathbf{v}_1$, describing the first principal axis, are all of the same sign and roughly equal magnitude. Statistically, this is representing the positive covariance among all of the variables. Scientifically, this is indicating that to a rough (one dimensional) approximation, differences between people are primarily differences in overall ability on all of the subtests.

With this in mind, we might refer to the first PC $\mathbf{f}_1$ as a measure of a person's overall ability. The second PC is dominated by the score on "picture completion", suggesting that this task is somewhat different than the other three.

```
F<-CY%*%V

t(F)%*%F
```

```
##           [,1]      [,2]      [,3]      [,4]
```

```
## [1,]  1.82e+03  4.80e-14 -2.95e-14  2.50e-13
## [2,]  4.80e-14  2.60e+02 -3.72e-14  2.12e-13
## [3,] -2.95e-14 -3.72e-14  2.49e+02 -3.70e-14
## [4,]  2.50e-13  2.12e-13 -3.70e-14  1.23e+02
```

```
L
```

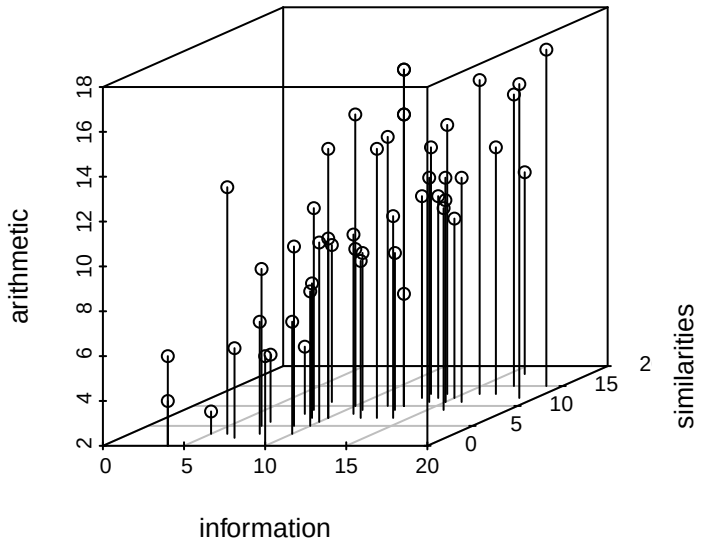
```
## [1] 1820  260  249  123
```

The variance of subjects in terms of their first PC is much larger than the variation in the remaining three dimensions.

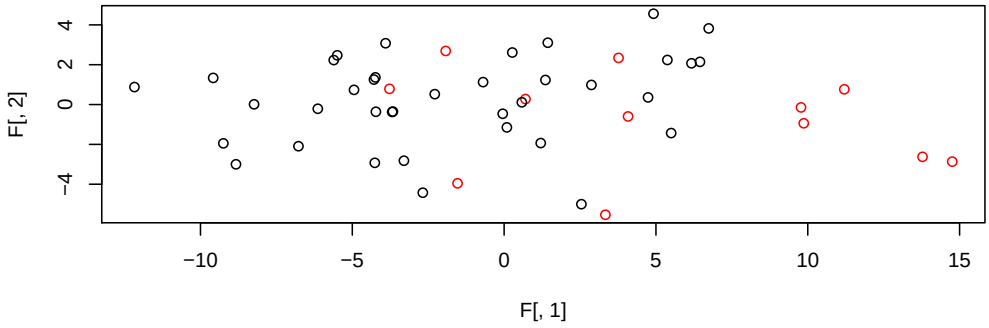
```
cbind( L, L/sum(L), cumsum(L)/sum(L))
```

```
##         L
## [1,] 1820 0.7421 0.742
## [2,]  260 0.1059 0.848
## [3,]  249 0.1017 0.950
## [4,]  123 0.0502 1.000
```

The first PC explains nearly 75% of the variance in the data. If we could visualize things in four dimensions, this would mean that the data would fall roughly along a line in four dimensions. Let's look in three dimensions:



The dataset also comes with a binary variable `senile` which indicates that a subject exhibits senility. Since the first PC represents overall cognitive function, we might expect it to be associated with `senility`.




```
senile<-wais[,1]
tapply(F[,1],senile,mean)

##     0     1
## -1.73  5.34


t.test(F[,1]~senile)

##
##  Welch Two Sample t-test
##
## data:  F[, 1] by senile
## t = -3, df = 20, p-value = 0.003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.35  -2.78
## sample estimates:
## mean in group 0 mean in group 1
##           -1.73            5.34
```

The red dots indicate senile subjects. Although there is evidence of an association, it is a weak one. If we were looking for a way to predict senility from the subtest scores, PCA might not be the best approach. While sometimes

an exogenous variable is well predicted by the first PCA (SLC with maximum variance) this is not true by logical necessity. We will discuss more appropriate classification methods later in the course.

While the PCs are uncorrelated with each other, they do have correlations with the original variables, and these correlations help us interpret the PCs. Not surprisingly, these correlations are associated with the elements of $\mathbf{V}$:

$$\begin{aligned}(n-1)\text{Cov}[\mathbf{Y},\mathbf{F}] &= \mathbf{Y}^\top\mathbf{F} \\ &= \mathbf{Y}^\top\mathbf{Y}\mathbf{V} \\ &= \mathbf{S}\mathbf{V} \\ &= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{\Lambda}\end{aligned}$$

```
V%*%diag(L) / (n-1)
```

```
##       [,1]    [,2]  [,3]   [,4]
## [1,] -21.2 -0.0444  1.21  2.041
## [2,] -23.1 -2.6105  1.75 -1.365
## [3,] -18.6  0.6491 -4.46 -0.237
## [4,] -10.5  4.6944  1.60 -0.707
```

```
cov(Y,F)
```

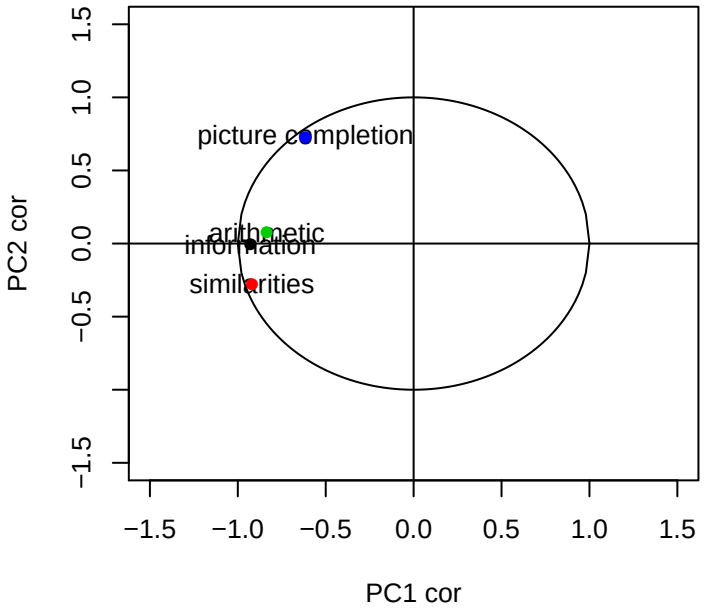
```
##                     [,1]    [,2]  [,3]   [,4]
## information        -21.2 -0.0444  1.21  2.041
## similarities       -23.1 -2.6105  1.75 -1.365
## arithmetic         -18.6  0.6491 -4.46 -0.237
## picture completion -10.5  4.6944  1.60 -0.707
```

Now recall the variance of $\mathbf{F}$ is given by $\mathbf{\Lambda}/(n-1)$. Let $\mathbf{\Psi}$ be the diagonal matrix of sample variances of the columns of $\mathbf{Y}$. Then

$$\begin{aligned}\text{Cor}[\mathbf{Y},\mathbf{F}] &= \sqrt{n-1}\mathbf{\Psi}^{-1/2}\text{Cov}[\mathbf{Y},\mathbf{F}]\mathbf{\Lambda}^{1/2}/(n-1) \\ &= \mathbf{\Psi}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}/\sqrt{n-1}\end{aligned}$$

```
cor(Y,F)

##                      [,1]     [,2]   [,3]    [,4]
## information        -0.928 -0.00514  0.143  0.3432
## similarities       -0.919 -0.27521  0.188 -0.2089
## arithmetic         -0.836  0.07734 -0.542 -0.0409
## picture completion -0.615  0.72951  0.254 -0.1595


diag(1/apply(Y,2,sd)) %*% V %*% diag(sqrt(L)) / sqrt(n-1)

##        [,1]     [,2]   [,3]    [,4]
## [1,] -0.928 -0.00514  0.143  0.3432
## [2,] -0.919 -0.27521  0.188 -0.2089
## [3,] -0.836  0.07734 -0.542 -0.0409
## [4,] -0.615  0.72951  0.254 -0.1595
```

So the first three subtests are strongly correlated with the first PC, while the fourth subtest is most strongly correlated with the second PC. It is sometimes useful to plot this as follows:



It is no coincidence that the point in this correlation plot fall inside a unit circle:

```
CFY<-cor(Y,F)

apply(CFY^2,1,sum)

##        information       similarities         arithmetic
##                  1                  1                  1
## picture completion
##                  1


apply(CFY^2,1,cumsum)

##      information similarities arithmetic picture completion
## [1,]       0.862        0.845      0.698              0.378
```

```
## [2,]       0.862       0.921       0.704              0.910
## [3,]       0.882       0.956       0.998              0.975
## [4,]       1.000       1.000       1.000              1.000
```

Exercise 14. *Show that the rows of* $\mathrm{Cor}[\mathbf{Y}, \mathbf{F}]$ *are unit vectors.*

6.2 NHANES

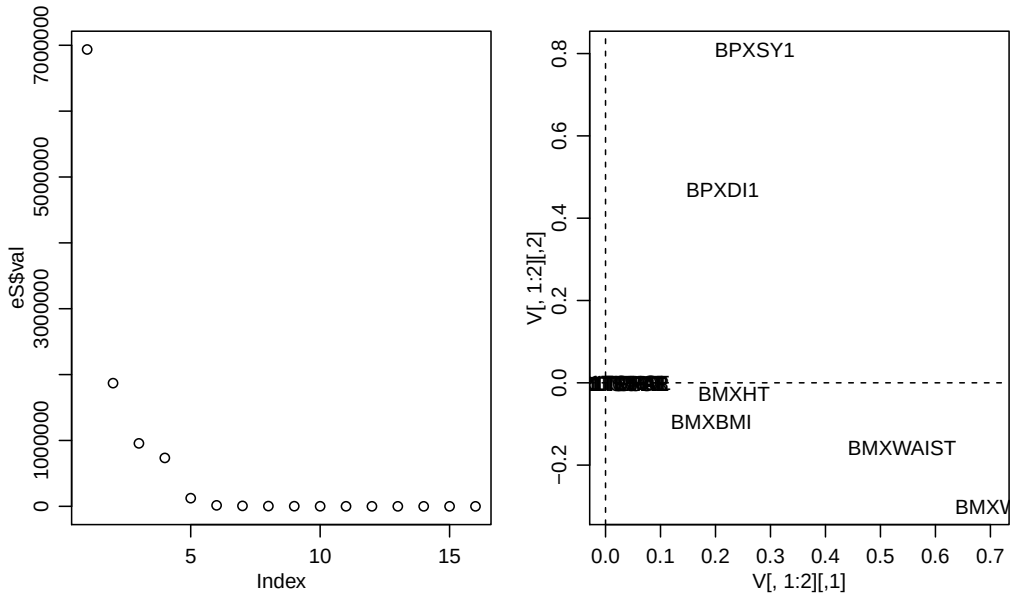
Let's gain more familiarity with PCA in practice with another example. The data for this example come from the National Health and Nutrition Examination Survey (NHANES).

```
dim(nhanes)

## [1] 6352   16


colnames(nhanes)

##  [1] "lnDR1TKCAL" "lnDR1TPROT" "lnDR1TCARB" "lnDR1TSUGR" "lnDR1TFIBE"
##  [6] "lnDR1TTFAT" "lnDR1TSFAT" "lnDR1TMFAT" "lnDR1TPFAT" "lnDR1TCHOL"
## [11] "BMXWT"      "BMXHT"      "BMXBMI"     "BMXWAIST"   "BPXSY1"
## [16] "BPXDI1"
```

First we center the columns and compute eigendecomposition.

```
Y<-nhanes
CY<-sweep(nhanes,2,apply(nhanes,2,mean))
S<-crossprod(CY)
eS<-eigen(S)
```

Since we have a moderate number of variables, we'll plot the eigenspectrum and coefficients of principal axes.



The first plot seems to indicate that the variation can be well-explained in one dimension. The second plot gives a clue as to what is going on.

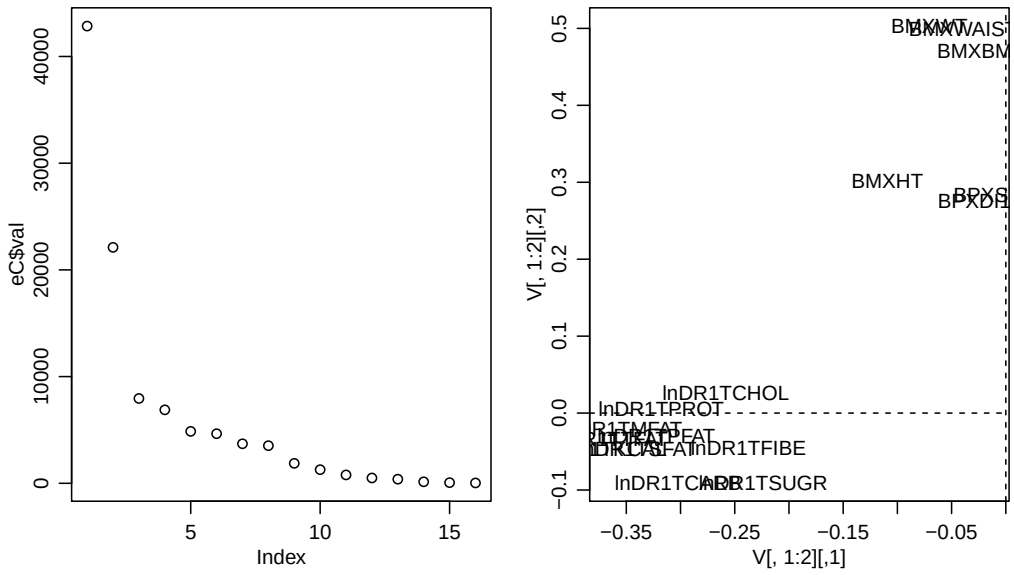
```
round( cbind( V[,1:2] , apply(Y,2,sd) ),3)

##              [,1]   [,2]   [,3]
## lnDR1TKCAL  0.001 -0.001  0.472
## lnDR1TPROT  0.002 -0.001  0.548
## lnDR1TCARB  0.000 -0.001  0.504
## lnDR1TSUGR -0.001 -0.003  0.709
## lnDR1TFIBE  0.001  0.000  0.599
## lnDR1TTFAT  0.002 -0.002  0.601
## lnDR1TSFAT  0.001 -0.002  0.633
## lnDR1TMFAT  0.002 -0.001  0.607
## lnDR1TPFAT  0.001 -0.002  0.627
## lnDR1TCHOL  0.003  0.000  0.926
## BMXWT       0.706 -0.300 24.032
## BMXHT       0.234 -0.025 12.577
## BMXBMI      0.193 -0.093  7.198
## BMXWAIST    0.539 -0.156 18.825
## BPXSY1      0.272  0.810 17.671
```

```
## BPXDI1      0.213  0.469 14.808
```

This calculation highlights that PCA will be highly sensitive to the scale of your variables. If you have several variables that are measured on different scales, PCA will just be telling you about these scales, instead of interesting information about how the variables are associated.

For such cases, PCA is often performed on the scaled variables. Equivalently, we analyze the eigendecomposition of $\text{Cor}[\mathbf{Y}]$ as opposed to $\text{Cov}[\mathbf{Y}]$.

```
CSY<-sweep(CY,2,apply(CY,2,sd),"/")
C<-crossprod(CSY)
eC<-eigen(C)

V<-eC$vec
L<-eC$val
```



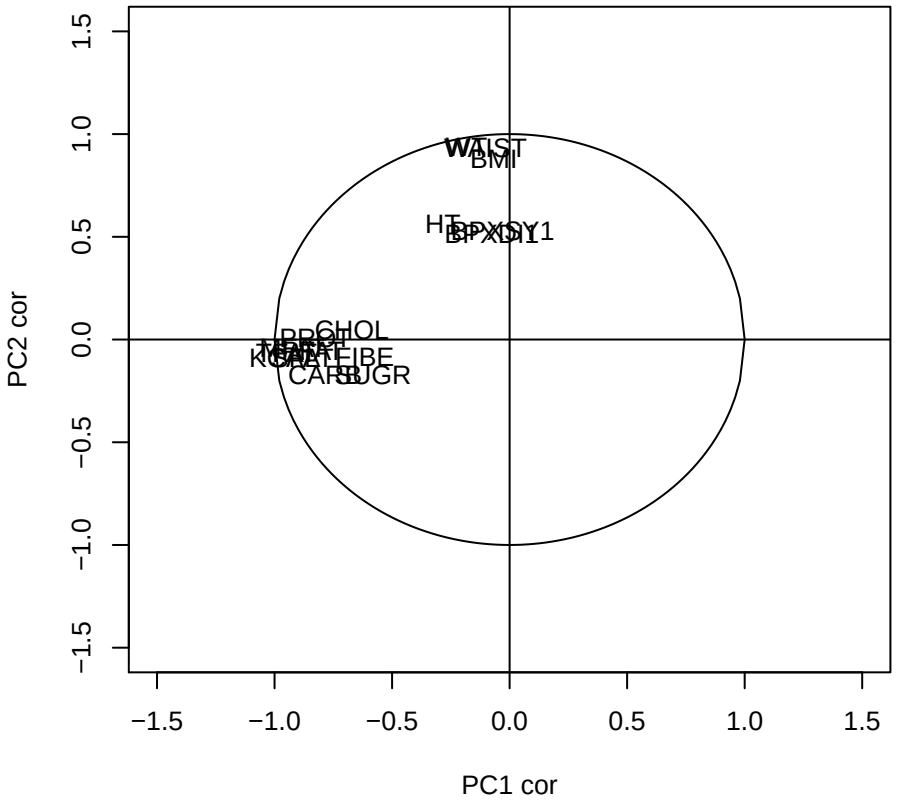
These plots look a lot more informative. It looks like a majority of the variation in the centered and scaled data can be described in two dimensions.

Numerically, we have the following:

```
cbind( L, L/sum(L), cumsum(L)/sum(L))

##             L
##  [1,] 42848.7 0.421673 0.422
##  [2,] 22111.9 0.217603 0.639
##  [3,]  7952.0 0.078255 0.718
##  [4,]  6886.2 0.067767 0.785
##  [5,]  4867.6 0.047902 0.833
##  [6,]  4649.1 0.045751 0.879
##  [7,]  3705.9 0.036469 0.915
##  [8,]  3527.8 0.034717 0.950
##  [9,]  1873.2 0.018434 0.969
## [10,]  1278.7 0.012583 0.981
## [11,]   786.6 0.007741 0.989
## [12,]   501.0 0.004930 0.994
## [13,]   387.2 0.003811 0.998
## [14,]   142.2 0.001400 0.999
## [15,]    64.0 0.000630 1.000
## [16,]    34.1 0.000335 1.000
```

So about 64% of the scaled data can be explained along the first two principal axes.

We can examine the associations between the principle components and the original (scaled and centered) variables by examining the entries of $\mathbf{V}$, or alternatively with a correlation plot:



The plot indicates that

- the dietary variables (KCAL, PROT, CARB, etc.) are well-explained in a one-dimensional subspace;
- the weight variables (WT, BMI, WAIST) are well explained by a separate, orthogonal one-dimensional subspace;
- some other variables (HT, BPXSY1, BPXSY2) are not well-explained by these subspaces.

```
colnames(CSY)[12:14]<-colnames(CSY)[c(13:14,12)]
CFY<-cor(CSY,F)
```

```
##       KCAL PROT CARB SUGR FIBE TFAT SFAT MFAT PFAT CHOL   WT  BMI WAIST
##  [1,] 0.92 0.68 0.62 0.34 0.38 0.90 0.78 0.85 0.70 0.45 0.03 0.00  0.01
##  [2,] 0.93 0.68 0.64 0.37 0.39 0.90 0.78 0.85 0.70 0.45 0.91 0.78  0.88
##  [3,] 0.95 0.74 0.93 0.76 0.49 0.94 0.81 0.89 0.72 0.70 0.92 0.79  0.89
##  [4,] 0.95 0.74 0.93 0.76 0.50 0.94 0.82 0.89 0.72 0.70 0.98 0.91  0.94
##  [5,] 0.95 0.74 0.93 0.78 0.58 0.94 0.82 0.90 0.73 0.71 0.98 0.97  0.96
##  [6,] 0.95 0.75 0.94 0.94 0.90 0.94 0.83 0.90 0.76 0.78 0.99 0.98  0.96
##  [7,] 0.95 0.90 0.94 0.94 0.94 0.99 0.85 0.94 0.84 0.93 0.99 0.98  0.96
##  [8,] 0.95 0.91 0.94 0.95 0.96 0.99 0.85 0.95 0.84 0.93 0.99 0.98  0.96
##  [9,] 0.95 0.91 0.95 0.95 0.96 0.99 0.98 0.95 0.99 0.94 0.99 0.98  0.96
## [10,] 0.98 0.97 0.95 0.97 1.00 1.00 0.98 0.95 0.99 0.99 0.99 0.98  0.96
## [11,] 0.99 1.00 0.99 1.00 1.00 1.00 0.98 0.95 0.99 1.00 0.99 0.98  0.96
## [12,] 0.99 1.00 0.99 1.00 1.00 1.00 1.00 1.00 1.00 1.00 0.99 0.99  0.96
## [13,] 0.99 1.00 0.99 1.00 1.00 1.00 1.00 1.00 1.00 1.00 0.99 1.00  1.00
## [14,] 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 0.99 1.00  1.00
## [15,] 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00  1.00
## [16,] 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00  1.00
##         HT BPXSY1 BPXDI1
##  [1,] 0.08   0.00   0.01
##  [2,] 0.40   0.28   0.27
##  [3,] 0.40   0.28   0.27
##  [4,] 0.42   0.66   0.70
##  [5,] 0.89   0.75   0.70
##  [6,] 0.97   0.76   0.73
##  [7,] 0.97   0.76   0.75
##  [8,] 1.00   1.00   1.00
##  [9,] 1.00   1.00   1.00
## [10,] 1.00   1.00   1.00
## [11,] 1.00   1.00   1.00
## [12,] 1.00   1.00   1.00
## [13,] 1.00   1.00   1.00
## [14,] 1.00   1.00   1.00
## [15,] 1.00   1.00   1.00
```

```
## [16,] 1.00   1.00   1.00
```

7 Summary:

1. For a typical dataset $\mathbf{Y} \in \mathbb{R}^{n \times p}$, the differences between rows are p-dimensional.

2. For some datasets $\mathbf{Y} \in \mathbb{R}^{n \times p}$, much of this variation can be described in less than p dimensions.

3. If this is the case, we can well-approximate a row $\mathbf{y}_i$ as

$$\mathbf{y}_i \approx \bar{\mathbf{y}} + f_{i,1}\mathbf{v}_1 + \cdots + f_{i,q}\mathbf{v}_q$$

   for some $q < p$.

4. The best q-dimensional approximation is obtained when $\mathbf{v}_j$ is the jth eigenvector of $\mathbf{S}$, and when $f_{i,j}$ is the projection of $(\mathbf{y}_i - \bar{\mathbf{y}})$ onto $\mathbf{v}_j$.

5. The axis determined by $\mathbf{v}_j$ is the jth principal axis, and the variable $\mathbf{f}_j = \mathbf{C}\mathbf{Y}\mathbf{v}_j$ is the jth principal component. The elements of $\mathbf{f}_j$ are called the jth principal component scores.

6. The 1st principal axis gives the direction of maximum variation in the data. The jth principal axis gives the direction of maximal variation that is orthogonal to the first $j-1$ principal axes.

7. The 1st principal component is a mean-zero SLC with maximum variation among all SLCs. The jth principal component is mean-zero SLC with maximum variation among all SLCs that are orthogonal to/uncorrelated with the first $j-1$ principal components.

8. The variation among the rows of an $n \times p$ dataset is in terms of p correlated variables. With principal components, we approximate this variation in terms of $q \leq p$ uncorrelated variables, $\mathbf{f}_1, \ldots, \mathbf{f}_q$.

References

K. R. Gabriel. Least squares approximation of matrices by additive and multiplicative models. *J. Roy. Statist. Soc. Ser. B*, 40(2):186–196, 1978. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1978)40:2<186:LSAOMB>2.0.CO;2-P&origin=MSN.

Wolfgang Karl Härdle and Léopold Simar. *Applied multivariate statistical analysis*. Springer, Heidelberg, fourth edition, 2015. ISBN 978-3-662-45170-0; 978-3-662-45171-7. doi: 10.1007/978-3-662-45171-7. URL http://dx.doi.org/10.1007/978-3-662-45171-7.