1 Principal Component Analysis

1.1 Introduction

Suppose that, as usual, we begin with a collection of measurements of different features for a group of samples. Some of these measurements will tell us quite a bit about the difference among our samples, while others may contain relatively little information. For example, if we are analyzing the effect of a certain weight loss regimen on a group of people, the age and weight of the subjects may have a great deal of influence on how successful the regimen is, while their blood pressure might not. One way to help identify which features are more significant is to ask whether or not the feature varies a lot among the different samples. If nearly all the measurements of a feature are the same, it can't have much power in distinguishing the samples, while if the measurements vary a great deal then that feature has a chance to contain useful information.

In this section we will discuss a way to measure the variability of measurements and then introduce principal component analysis (PCA). PCA is a method for finding which linear combinations of measurements have the greatest variability and therefore might contain the most information. It also allows us to identify combinations of measurements that don't vary much at all. Combining this information, we can sometimes replace our original system of features with a smaller set that still captures most of the interesting information in our data, and thereby find hidden characteristics of the data and simplify our analysis a great deal.

1.2 Variance and Covariance

1.2.1 Variance

Suppose that we have a collection of measurements $(x_1, \ldots, x_n)$ of a particular feature X. For example, x_i might be the initial weight of the *ith* participant in our weight loss study. The mean of the values $(x_1, \ldots, x_n)$ is

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The simplest measure of the variability of the data is called its *variance.*

Definition: The (sample) variance of the data $x_1, \ldots, x_n$ is

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu_X^2 \tag{1}$$

The square root of the variance is called the *standard deviation.*

As we see from the formula, the variance is a measure of how 'spread out' the data is from the mean.

Recall that in our discussion of linear regression we thought of our set of measurements $x_1, \ldots, x_n$ as a vector – it's one of the columns of our data matrix. From that point of view, the variance has a geometric interpretation – it is $\frac{1}{N}$ times the square of the distance from the point $X = (x_1, \ldots, x_n)$ to the point $\mu_X(1, 1, \ldots, 1) = \mu_X E$:

$$\sigma_X^2 = \frac{1}{n}(X - \mu_X E) \cdot (X - \mu_X E) = \frac{1}{n}\|X - \mu_X E\|^2. \tag{2}$$

1.2.2 Covariance

The variance measures the dispersion of measures of a single feature. Often, we have measurements of multiple features and we might want to know something about how two features are related. The *covariance* is a measure of whether two features tend to be related, in the sense that when one increases, the other one increases; or when one increases, the other one decreases.

Definition: Given measurements $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ of two features X and Y, the covariance of X and Y is

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i \tag{3}$$

There is a nice geometric interpretation of this, as well, in terms of the dot product. If $X = (x_1, \ldots, x_n)$ and $Y = (y_1 \ldots, y_n)$ then

$$\sigma_{XY} = \frac{1}{N}((X - \mu_X) \cdot (Y - \mu_Y)).$$

From this point of view, we can see that σ_{XY} is positive if the $X - \mu_X$ and $Y - \mu_Y$ vectors "point roughly in the same direction" and its negative if they "point roughly in the opposite direction."

1.2.3 Correlation

One problem with interpreting the variance and covariance is that we don't have a scale – for example, if σ_{XY} is large and positive, then we'd like to say that X and Y are closely related, but it could be just that the entries of $X - \mu_X$ and $Y - \mu_Y$ are large. Here, though, we can really take advantage of the geometric interpretation. Recall that the dot product of two vectors satisfies the formula

$$a \cdot b = \|a\|\|b\|\cos(\theta)$$

where θ is the angle between a and b. So

$$\cos(\theta) = \frac{a \cdot b}{\|a\|\|b\|}.$$

Let's apply this to the variance and covariance, by noticing that

$$\frac{(X-\mu_X)\cdot(Y-\mu_Y)}{\|(X-\mu_X)\|\|(Y-\mu_Y)\|} = \frac{\sigma_{XY}}{\sigma_{XX}\sigma_{YY}}$$

so the quantity

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{4}$$

measures the cosine of the angle between the vectors $X-\mu_X$A and $Y-\mu_Y$.

Definition: The quantity r_{XY} defined in eq. 4 is called the (sample) *correlation coefficient* between X and Y. We have $0 \leq |r_{XY}| \leq 1$ with $r_{XY} = \pm 1$ if and only if the two vectors $X-\mu_X$ and $Y-\mu_Y$ are collinear in $\mathbf{R}^n$.

Figure 1 illustrates data with different values of the correlation coefficient.

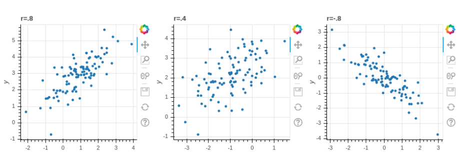

Figure 1: Correlation

1.2.4 The covariance matrix

In a typical situation we have many features for each of our (many) samples, that we organize into a data matrix X. To recall, each column of X corresponds to a feature that we measure, and each row corresponds to a sample. For example, each row of our matrix might correspond to a person enrolled in a study, and the columns correspond to height (cm), weight (kg), systolic blood pressure, and age (in years):

Table 1: A sample data matrix X

sample	Ht	Wgt	Bp	Age
A	180	75	110	35
B	193	80	130	40
...	...	...	...	...
U	150	92	105	55

If we have multiple features, as in this example, we might be interested in the variance of each feature and all of their mutual covariances. This "package" of information can be obtained "all at once" by taking advantage of some matrix algebra.

Definition: Let X be a $k \times N$ data matrix, where the N columns of X correspond to different features and the k rows to different samples. Let X_0 be the centered version of this data matrix, obtained by subtracting the mean μ_i of column i from all the entries x_{si} in that column. Then the $N \times N$ symmetric matrix

$$D_0 = \frac{1}{N} X_0^{\intercal} X_0$$

is called the (sample) covariance matrix for the data.

Proposition: The diagonal entries d_{ii} of D_0 are the variances of the columns of X:

$$d_{ii} = \sigma_i^2 = \frac{1}{N} \sum_{s=1}^{k} (x_{si} - \mu_i)^2$$

and the off-diagonal entries $d_{ij} = d_{ji}$ are the covariances of the i^{th} and j^{th} columns of X:

$$d_{ij} = \sigma_{ij} = \frac{1}{N} \sum_{s=1}^{k} (x_{si} - \mu_i)(x_{sj} - \mu_j)$$

Proof: This follows from the definitions, but it's worth checking the details, which we leave as an exercise.

1.2.5 Linear Combinations of Features (Scores)

Sometimes useful information about our data can be revealed if we combine different measurements together to obtain a "hybrid" measure that captures something interesting. For example, in the Auto MPG dataset that we studied in the section on Linear Regression, we looked at the influence of both vehicle weight w and engine displacement e on gas mileage; perhaps their is some value in considering a hybrid "score" defined as

$$S = a * w + b * e$$

for some constants a and b – maybe by choosing a good combination we could find a better predictor of gas mileage than using one or the other of the features individually.

As another example, suppose we are interested in the impact of the nutritional content of food on weight gain in a study. We know that both calorie content and the level dietary fiber contribute to the weight gain of participants eating this particular food; maybe there is some kind of combined "calorie/fiber" score we could introduce that captures the impact of that food better.

Definition: Let X_0 be a (centered) $k \times N$ data matrix giving information about N features for each of k samples. A linear synthetic feature, or a linear score, is a linear combination of the N features. The linear score is defined by constants $a_1, \ldots, a_n$ so that If $y_1, \ldots, y_N$ are the values of the features for a particular sample, then the linear score for that sample is

$$S = a_1 y_1 + a_2 y_2 + \cdots + a_N y_N$$

Lemma: The values of the linear score for each of the k samples can be calculated as

$$\begin{bmatrix} S_1 \\ \vdots \\ S_k \end{bmatrix} = X_0 \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}. \tag{5}$$

Proof: Multiplying a matrix by a column vector computes a linear combination of the columns – that's what this lemma says. Exercise 3 asks you to write out the indices and make sure you believe this.

1.2.6 Mean and variance of scores

When we combine features to make a hybrid score, we assume that the features were centered to begin with, so that each features has mean zero. As a result, the mean of the hybrid features is again zero.

Lemma: A linear combination of features with mean zero again has mean zero.

Proof: Let S_i be the score for the i^{th} sample, so

$$S_i = \sum_{j=1}^{N} x_{ij} a_j.$$

where X_0 has entries x_{ij}. Then the mean value of the score is

$$\mu_S = \frac{1}{k}\sum_{i=1}^{k} S_i = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{N} x_{ij}a_j.$$

Reversing the order of the sum yields

$$\mu_S = \frac{1}{k}\sum_{j=1}^{N}\sum_{i=1}^{k} x_{ij}a_j = \sum_{j=1}^{N} a_j \frac{1}{k}(\sum_{i=1}^{k} x_{ij}) = \sum_{j=1}^{N} a_j\mu_j = 0$$

where $\mu_j = 0$ is the mean of the j^{th} feature (column) of X_0.

The variance is more interesting, and gives us an opportunity to put the covariance matrix to work. Remember from 2 that, since a score S has mean zero, it's variance is $\sigma_S^2 = S \cdot S$ – where here the score S is represented by the column vector with entries $S_1, \ldots S_k$ as in eq. 5.

Lemma: The variance of the score S with weights $a_1, \ldots a_N$ is

$$\sigma_S^2 = a^\intercal D_0 a = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix} D_0 \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \tag{6}$$

More generally, if S_1 and S_2 are scores with weights $a_1, \ldots, a_N$ and $b_1, \ldots, b_N$ respectively, then the covariance $\sigma_{S_1 S_2}$ is

$$\sigma_{S_1 S_2} = a^\intercal D_0 b.$$

Proof: From eq. 2 and 5 we know that

$$\sigma_S^2 = S \cdot S$$

and

$$S = X_0 a.$$

Since $S \cdot S = \frac{1}{N} S^\intercal S$, this gives us

$$\sigma_S^2 = \frac{1}{N}(X_0 a)^\intercal (X_0 a) = \frac{1}{N} a^\intercal X_0^\intercal X_0 a = a^\intercal D_0 a$$

as claimed.

For the covariance, use a similar argument with eq. 3 and eq. 5. writing $\sigma_{S_1 S_2} = \frac{1}{N} S_1 \cdot S_2$ and the fact that S_1 and S_2 can be written as $X_0 a$ and $X_0 b$.

The point of this lemma is that the covariance matrix contains not just the variances and covariances of the original features, but also enough information to construct the variances and covariances for *any linear combination of features.*

In the next section we will see how to exploit this idea to reveal hidden structure in our data.

1.2.7 Geometry of Scores

Let's begin by looking at fig. 2, which shows a scatter plot of some simulated data having 50 samples and two features. This data has been centered, so it can be represented in a 50×2 data matrix X_0 each row of which is the coordinates (x_0, x_1) of one of the points in the picture.

The scatter plot shows that the data points are arranged in a more or less elliptical cloud oriented at an angle to the xy-axes which represent the two given features. The two individual histograms show the distribution of the two features – each has mean zero, with the x-features distributed between -2 and 2 and the y feature between -4 and 4. Looking just at the two features individually, meaning only at the two histograms, we can't see the overall elliptical structure.

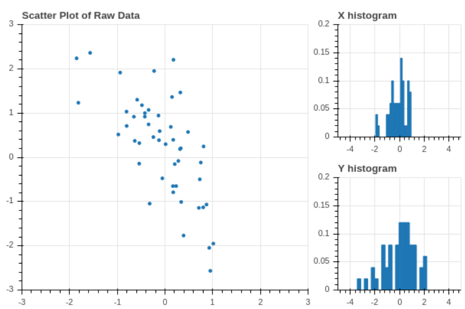

Figure 2: Simulated Data with Two Features

How can we get a better grip on our data in this situation? We can try to find a "direction" in our data that better illuminates the variation of the data. For example, suppose that we pick a unit vector at the origin pointing in a particular direction in our data. See fig. 3.

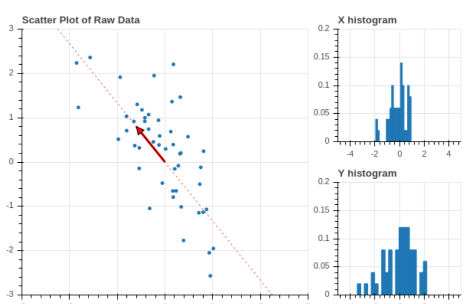

Figure 3: A direction in the data

Now we can orthogonally project the datapoints onto the line defined by this vector, as shown in fig. 4.

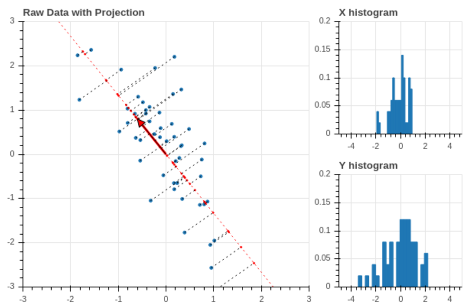

Figure 4: Projecting the datapoints

Recall that if the unit vector is defined by coordinates $u = [u_0, u_1]$, then the orthogonal projection of the point x with coordinates (x_0, x_1) is $(x \cdot u)u$. Now

$$x \cdot u = u_0 x_0 + u_1 x_1$$

so the coordinates of the points along the line defined by u are the values of the score Z defined by $u = [u_0, u_1]$. Using our work in the previous section, we see that we can find all of these coordinates by matrix multiplication:

$$Z = X_0 u$$

where X_0 is our data matrix. Now let's add a histogram of the values of Z to our picture:

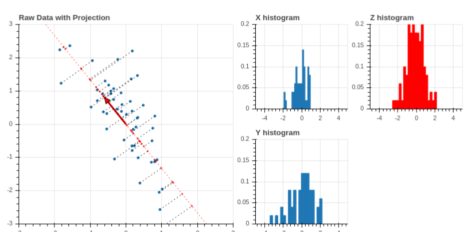

Figure 5: Distribution of Z

This histogram shows the distribution of the values of Z along the tilted line defined by the unit vector u.

Finally, using our work on the covariance matrix, we see that the variance of Z is given by

$$\sigma_Z^2 = \frac{1}{50} u^\intercal X_0^\intercal X_0 u = u^\intercal D_0 u$$

where D_0 is the covariance matrix of the data X_0.

Lemma: Let X_0 be a $k \times N$ centered data matrix, and let $D_0 = \frac{1}{N} X_0^\intercal X_0$ be the associated covariance matrix. Let u be a unit vector in "feature space" $\mathbf{R}^N$.

Then the score $S = X_0 u$ can be interpreted as the coordinates of the points of X_0 projected onto the line generated by u. The variance

$$\sigma_S^2 = u^\intercal D_0 u = \sum_{i=1}^{k} s_i^2$$

where $s_i = X_0[i,:]u$ is the dot product of the i^{th} row $X_0[i,:]$ with u. It measures the variability in the data "in the direction of the unit vector u".

1.2.8 Principal Components

1.2.8.1 Change of variance with direction As we've seen in the previous section, if we choose a unit vector u in the feature space and find the projection $X_0 u$ of our data onto the line through u, we get a "score" that we can use to measure the variance of the data in the direction of u. What happens as we vary u?

To study this question, let's continue with our simulated data from the previous section, and introduce a unit vector

$$u(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \end{bmatrix}.$$

This is in fact a unit vector, since $\sin^2(\theta) + \cos^2(\theta) = 1$, and it is oriented at an angle θ from the x-axis.

The variance of the data in the direction of $u(\theta)$ is given by

$$\sigma_\theta^2 = u(\theta)^\intercal D_0 u(\theta).$$

A plot of this function for the data we have been considering is in fig. 6. As you can see, the variance goes through two full periods with the angle, and it reaches a maximum and minimum value at intervals of $\pi/2$ – so the two angles where the variance are maximum and minimum are orthogonal to one another.

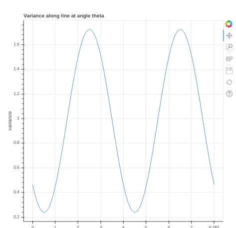

Figure 6: Change of variance with angle theta

The two directions where the variance is maximum and minimum are drawn on the original data scatter plot in fig. 7 .

Let's try to understand why this is happening.

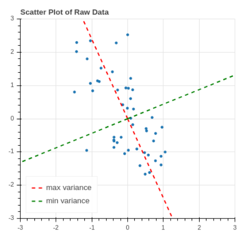

Figure 7: Data with principal directions

1.2.8.2 Directions of extremal variance Given our centered, $k \times N$ data matrix X_0, with its associated covariance matrix $D_0 = \frac{1}{N} X_0^\intercal X_0$, we would like to find unit vectors u in $\mathbf{R}^N$ so that

$$\sigma_u^2 = u^\intercal D_0 u$$

reaches its maximum and its minimum. Here σ_u^2 is the variance of the "linear score" $X_0 u$ and it represents how dispersed the data is in the "u direction" in $\mathbf{R}^N$.

In this problem, remember that the coordinates of $u = (u_1, \ldots, u_N)$ are the variables and the symmetric matrix D_0 is given. As usual, we to find the maximum and minimum values of σ_u^2, we should look at the partial derivatives of σ_u^2 with respect to the variables u_i and set them to zero. Here, however, there is a catch – we want to restrict u to being a unit vector, with $u \cdot u = \sum u_i^2 = 1$.

So this is a *constrained optimization problem*:

- Find extreme values of the function

$$\sigma_u^2 = u^\intercal D_0 u$$

- Subject to the constraint $\|u\|^2 = u \cdot u = 1$ (or $u \cdot u - 1 = 0$)

As we learned in multivariate calculus, we can use the technique of *Lagrange Multipliers* to solve such a problem.

To apply this method, we introduce the function

$$S(u, \lambda) = u^\intercal D_0 u - \lambda(u \cdot u - 1) \tag{7}$$

Then we compute the gradient

$$\nabla S = \begin{bmatrix} \frac{\partial S}{\partial u_1} \\ \vdots \\ \frac{\partial S}{\partial u_N} \\ \frac{\partial S}{\partial \lambda} \end{bmatrix} \tag{8}$$

and solve the system of equations $\nabla S = 0$. Here we have written the gradient as a column vector for reasons that will become clearer shortly.

Computing all of these partial derivatives looks messy, but actually if we take advantage of matrix algebra it's not too bad. The following two lemmas explain how to do this.

Lemma: Let M be a $k \times N$ matrix with constant coefficients and let u be a $N \times 1$ column vector whose entries are $u_1, \ldots u_N$. The function $F(u) = Mu$ is a linear map from $\mathbf{R}^N \to \mathbf{R}^k$. Its (total) derivative is a linear map between the same vector spaces, and satisfies

$$D(F)(v) = Mv$$

for any $N \times 1$ vector v. If u is a $1 \times k$ matrix, and $G(u) = uM$, then

$$D(G)(v) = vM$$

for any $1 \times k$ vector v. (This is the matrix version of the derivative rule that $\frac{d}{dx}(ax) = a$ for a constant a.)

Proof: Since $F : \mathbf{R}^N \to \mathbf{R}^k$, we can write out F in more traditional function notation as

$$F(u) = (F_1(u_1, \ldots, u_N), \ldots, F_k(u_1, \ldots, u_N)$$

where

$$F_i(u_1, \ldots u_N) = \sum_{j=1}^{N} m_{ij} u_j.$$

Thus $\frac{\partial F_i}{\partial u_j} = m_{ij}$. The total derivative $D(F)$ is the linear map with matrix

$$D(F)_{ij} = \frac{\partial F_i}{\partial u_j} = m_{ij} = M.$$

The other result is proved the same way.

Lemma: Let D be a symmetric $N \times N$ matrix with constant entries and let u be an $N \times 1$ column vector of variables $u_1, \ldots, u_N$. Let $F : \mathbf{R}^N \to R$ be the function $F(u) = u^\intercal D u$. Then the derivative gradient $\nabla_u F$ is a vector field – that is, a vector-valued function of u, and is given by the formula

$$\nabla_u F = 2Du$$

Proof: Let d_{ij} be the i, j entry of D. We can write out the function F to obtain

$$F(u_1, \ldots, u_N) = \sum_{i=1}^{N} \sum_{j=1}^{N} u_i d_{ij} u_j.$$

Now $\frac{\partial F}{\partial u_i}$ is going to pick out only terms where u_i appears, yielding:

$$\frac{\partial F}{\partial u_i} = \sum_{j=1}^{N} d_{ij} u_j + \sum_{j=1}^{N} u_j d_{ji}$$

Here the first sum catches all of the terms where the first "u" is u_i; and the second sum catches all the terms where the second "u" is u_i. The diagonal terms $u_i^2 d_{ii}$ contribute once to each sum, which is consistent with the rule that the derivative of $u_i^2 d_{ii} = 2u_i d_{ii}$. To finish the proof, notice that

$$\sum_{j=1}^{N} u_j d_{ji} = \sum_{j=1}^{N} d_{ij} u_j$$

since D is symmetric, so in fact the two terms are the same Thus

$$\frac{\partial}{\partial u_i} F = 2 \sum_{j=1}^{N} d_{ij} u_j$$

But the right hand side of this equation is twice the i^{th} of Du, so putting the results together we get

$$\nabla_u F = \begin{bmatrix} \frac{\partial F}{\partial u_1} \\ \vdots \\ \frac{\partial F}{\partial u_N} \end{bmatrix} = 2Du.$$

The following theorem puts all of this work together to reduce our questions about how variance changes with direction.

1.2.8.3 Critical values of the variance **Theorem:** The critical values of the variance σ_u^2, as u varies over unit vectors in $\mathbf{R}^N$, are the eigenvalues $\lambda_1, \ldots, \lambda_N$ of the covariance matrix D, and if e_i is a unit eigenvector corresponding to λ_i, then $\sigma_{e_i}^2 = \lambda_i$.

Proof: Recall that we introduced the Lagrange function $S(u, \lambda)$, whose critical points give us the solutions to our constrained optimization problem. As we said in eq. 7:

$$S(u, \lambda) = u^\intercal D_0 u - \lambda(u \cdot u - 1) = u^\intercal D_0 u - \lambda(u \cdot u) + \lambda$$

Now apply our Matrix calculus lemmas. First, let's treat λ as a constant and focus on the u variables. We can write $u \cdot u = u^\intercal I_N u$ where I_N is the identity matrix to compute:

$$\nabla_u S = 2D_0 u - 2\lambda u$$

For λ we have

$$\frac{\partial}{\partial \lambda} S = -u \cdot u + 1.$$

The critical points occur when

$$\nabla_u S = 2(D_0 - \lambda)u = 0$$

and

$$\frac{\partial}{\partial\lambda}S = 1 - u \cdot u = 0$$

The first equation says that λ must be an eigenvalue, and u an eigenvector:

$$D_0 u = \lambda u$$

while the second says u must be a unit vector $u \cdot u = \|u\|^2 = 1$. The second part of the result follows from the fact that if e_i is a unit eigenvector with eigenvalue λ_i then

$$\sigma_{e_i}^2 = e_i^\intercal D e_i = \lambda_i \|e_i\|^2 = \lambda_i.$$

To really make this result pay off, we need to recall some key facts about the eigenvalues and eigenvectors of symmetric matrices. Because these facts are so central to this result, and to other applications throughout machine learning and mathematics generally, we will prove them in the next section.

1.2.8.4 Eigenvalues and eigenvectors of symmetric matrices

Let D be an $N \times N$ symmetric matrix with real entries. Then:

- All of the eigenvalues $\lambda_1, \ldots, \lambda_N$ of D are real.
- If $u^\intercal D u \geq 0$ for all $u \in \mathbf{R}^N$, then all eigenvalues λ_i are non-negative.
- If v is an eigenvector for D with eigenvalue λ, and w is an eigenvector with a different eigenvalue λ', then v and w are orthogonal: $v \cdot w = 0$.

- There is an orthonormal basis $u_1, \ldots, u_N$ of $\mathbf{R}^N$ made up of eigenvectors of D corresponding to the eigenvalues λ_i.
- Let Λ be the diagonal matrix with entries $\lambda_1, \ldots, \lambda_N$ and let P be the matrix whose columns are made up of the vectors u_i. Then:

$$D = P \Lambda P^\intercal.$$

If we combine this theorem with our result from **??** 1.2.8.4 then we get a complete picture. Let D_0 be the covariance matrix of our data. Since

$$\sigma_u^2 = u^\intercal D_0 u \geq 0(\mathrm{it'sasumofsquares})$$

we know that the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ are all nonnegative. Choose a corresponding sequence $u_1, \ldots u_N$ of orthogonal eigenvectors where all $\|u_i\|^2 = 1$. Since the u_i form a basis of $\mathbf{R}^N$, any score is a linear combination of the u_i:

$$S = \sum a_i u_i.$$

Since $u_i^\intercal D u_j = \lambda_j u_i^\intercal u_j = 0$ unless $i = j$, in which case it is λ_i, we can compute

$$\sigma_S^2 = \sum_{i=1}^{N} \lambda_i a_i^2,$$

and $\|S\|^2 = \sum a_i^2$ since the u_i are an orthonormal set. So in these coordinates, are optimization problem is:

- maximize $\sum \lambda_i a_i^2$
- subject to the constraint $\sum a_i^2 = 1$.

We don't need any fancy math to see that the maximum happens when $a_1 = 1$ and the other $a_j = 0$, and in that case, the maximum is λ_1. (If λ_1 occurs more than once, there may be a whole subspace of directions where the variance is maximal). Similarly, the minimum value is λ_N and occurs when $a_N = 1$ and the others are zero.

Definition: The orthonormal unit eigenvectors u_i for D_0 are the *principal directions* or *principal components* for the data X_0.

Theorem: The maximum variance occurs in the principal direction(s) associated to the largest eigenvalue, and the minimum variance in the principal direction(s) associated with the smallest one. The covariance between scores in principal directions associated with different eigenvalues is zero.

At this point, the picture in fig. 7 makes sense – the red and green dashed lines are the principal directions, they are orthogonal to one another, and the point in the directions where the data is most (and least) "spread out."

Sometimes the results above are presented in a slightly different form, and may be referred to, in part, as Rayleigh's theorem.

Corollary: (Rayleigh's Theorem) Let D be a real symmetric matrix and let

$$H(v) = \max_{v \neq 0} \frac{v^\intercal D v}{v^\intercal v}.$$

Then $H(v)$ is the largest eigenvalue of D.

Proof: The maximum of the function $H(v)$ is the solution to the same optimization problem that we considered above.

Exercises.

1. Prove that the two expressions for σ_X^2 given in section 1.2.1 are the same.
2. Prove that the covariance matrix is as described in the proposition in 1.2.4.
3. Let X_0 be a $k \times N$ matrix with entries x_{ij} for $1 \leq i \leq k$ and $1 \leq j \leq N$. If a linear score is defined by the constants $a_1, \ldots a_N$, check that equation eq. 5 holds as claimed.

4. Why is it important to use a unit vector when computing the variance of X_0 in the direction of u? Suppose $v = \lambda u$ where u is a unit vector and $\lambda > 0$ is a constant. Let S' be the score $X_0 v$. How is the variance of S' related to that of $S = X_0 u$?