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.

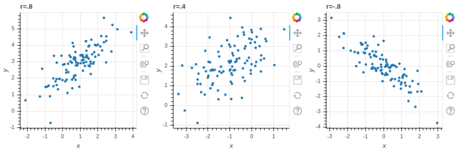

Figure 1: Correlation

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 $\sigma_{S_1S_2}$ is

$$\sigma_{S_1S_2} = a^\intercal D_0 b.$$

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

$$\sigma_S^2 = \frac{1}{N}(X_0a)^\intercal(X_0a) = \frac{1}{N}a^\intercal X_0^\intercal X_0 a = a^\intercal D_0 a$$

as claimed.

For the covariance, use a similar argument with eq. 3 and eq. 5. writing $\sigma_{S_1S_2} = \frac{1}{N}S_1 \cdot S_2$ and the fact that S_1 and S_2 can be written as X_0a and X_0b.

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.*

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.