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 $\mu_X(1, 1, \ldots, 1) = \mu_X E$:

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

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

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i \tag{2}$$

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 has the following geometric interpretation:

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

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

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_{XX}\sigma_{YY}} \tag{3}$$

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. 3 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.

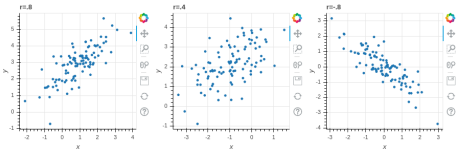

Figure 1: Correlation

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):

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

Exercises.

1. Prove that the two expressions for σ_X^2 given in section 1.2.1 are the same.