Polar (singular-value) decomposition and principal components
2012 September 21

Suppose we have a population of N individuals, each of whom has values for d attributes. We denote the values by $x_{n,j}$, where n, ranging from 1 to N, denotes the individual and j, ranging from 1 to d, denotes the attribute. We will assume that the values have all been centered so that within the (large) population we are considering, they have mean zero, and we will further assume that the values have been rescaled so that they all lie in the same interval, which we choose to be the interval $[-1, 1]$. What we are interested in is the sample correlation matrix,

$$C_{jk} = \frac{1}{N} \sum_{n=1}^{N} x_{n,j} x_{n,k}$$

It is easy to see that if we define an $N \times d$ *population matrix* X, whose matrix elements are $X_{nj} = x_{n,j}$, the correlation matrix is $C = X^T X/N$. The other way of combining X and X^T gives a different sort of sample correlation matrix, $B = XX^T/d$, whose matrix elements are

$$B_{nm} = \frac{1}{d} \sum_{j=1}^{d} x_{n,j} x_{m,j} \ .$$

Whereas B gives the correlation between attributes, averaged over the population, D expresses the correlation between individuals, averaged over attributes.

Let's recast things in terms of bras and kets and operators, partly because I prefer abstractions to representations and partly to make a connection with the notation we use in quantum mechanics. We let $|e_j\rangle$ be the basis vector for the jth attribute, $|f_n\rangle$ the basis vector for the nth individual, and

$$X = \sum_{n,j} x_{n,j} |f_n\rangle\langle e_j|$$

the operator corresponding to the matrix X. The first sample correlation operator is

$$\begin{aligned} C = \frac{1}{N} X^T X &= \frac{1}{N} \sum_{n,j,m,k} |e_j\rangle x_{n,j} \underbrace{\langle f_n|f_m\rangle}_{=\delta_{nm}} x_{m,k} \langle e_k| \\ &= \sum_{j,k} |e_j\rangle \left(\frac{1}{N} \sum_n x_{n,j} x_{n,k} \right) \langle e_k| \\ &= \sum_{j,k} C_{jk} |e_j\rangle\langle e_k| \ , \end{aligned}$$

and the second is

$$\begin{aligned}B = \frac{1}{d}XX^T &= \frac{1}{d}\sum_{n,j,m,k}|f_n\rangle x_{n,j}\underbrace{\langle e_j|e_k\rangle}_{=\delta_{jk}}x_{m,k}\langle f_m| \\ &= \sum_{n,m}|f_n\rangle\left(\frac{1}{d}\sum_j x_{n,j}x_{m,j}\right)\langle f_m| \\ &= \sum_{n,m}B_{nm}|f_n\rangle\langle f_m|\ .\end{aligned}$$

Since we're dealing with real quantities, we are working in a real vector space: inner products are real and symmetric, adjoints become transposes, and unitary operators reduce to orthogonal matrices. In order not to work with nonsquare matrices like X, we tack on additional columns of zeroes to the population matrix X, to make it square; these new columns can be thought of as describing attributes for which all individuals have a zero value. We add additional basis vectors $|e_j\rangle$ for these irrelevant attributes.

Since the correlation matrix C is symmetric and positive, we can diagonalize it in an orthonormal basis,

$$C = \sum_j \lambda_j^2|u_j\rangle\langle u_j|\ ,$$

with nonnegative eigenvalues λ_j^2. The vectors $|u_j\rangle$ can be thought of as new attributes, which are just the right combinations of the original attributes to be uncorrelated. The new attributes are called the *principal components*; finding the principal components is a standard technique for determining the important attributes of a population, as opposed to the original attributes, which were probably chosen because they were easy to observe. Just as $\langle f_n|X|e_j\rangle = x_{n,j}$ is the value of the jth original attribute for the nth individual, $\langle f_n|X|u_j\rangle = y_{n,j}$ can be thought of as the value of the jth new attribute for the nth individual. The sample correlation matrix for the new attributes,

$$\frac{1}{N}\sum_{n=1}^N y_{n,j}y_{n,k} = \frac{1}{N}\sum_{n=1}^N\langle u_j|X^T|f_n\rangle\langle f_n|X|u_k\rangle = \frac{1}{N}\langle u_j|X^TX|u_k\rangle = \langle u_j|C|u_k\rangle = \lambda_j^2\delta_{jk}\ ,$$

is, of course, diagonal, with the eigenvalues λ_j^2 expressing the strength of the self-correlation, or variability, of the new attributes. There is no good reason to use normalized vectors for the new attributes, except that the eigenvalues then express the strength of self-correlation of the new attributes.

The polar-decomposition theorem says that there is an orthogonal matrix O such that

$$X = O\sqrt{X^TX} = \sqrt{N}O\sqrt{C} = \sqrt{XX^T}O = \sqrt{d}\sqrt{B}O\ .$$

The operator O maps the eigenvectors of X^TX to the eigenvectors of XX^T, i.e.,

$$O|u_j\rangle = |n_j\rangle\ , \qquad dB = XX^T = OX^TXO^T = NOCO^T = N\sum_j\lambda_j^2|n_j\rangle\langle n_j|\ ,$$

and the population operator X takes the form

$$X = \sqrt{N} \sum_j \lambda_j |n_j\rangle\langle u_j| \;, \quad X|u_j\rangle = \sqrt{N}\lambda_j |n_j\rangle \;.$$

The quantities $\sqrt{N}\lambda_j$ are the singular values of X, and the orthonormal vectors $|u_j\rangle$ and $|n_j\rangle$ are the right and left singular vectors of X. We can now interpret the singular values and basis vectors $|n_j\rangle$ by noting that

$$y_{n,j} = \langle f_n|X|u_j\rangle = \sqrt{N}\lambda_j \langle f_n|n_j\rangle \;;$$

i.e., the value of the jth new attribute for the nth individual has is governed by the singular value $\sqrt{N}\lambda_j$ and the projection of $|f_n\rangle$ onto $|n_j\rangle$. It is natural to rescale the values of the new attributes by the corresponding singular value,

$$\frac{y_{n,j}}{\sqrt{N}\lambda_j} = \langle f_n|n_j\rangle \;,$$

and then the value of the jth new attribute for nth individual is simply obtained by projecting the individual's vector onto the left singular vector for the new attribute.

It is useful to do a simple example to put some flesh on these abstractions. Suppose we have a population that has two attributes, both of which take on values ± 1 and which are completely correlated: if an individual has value $+1$ for the first attribute, it has $+1$ for the second attribute as well; similarly, if an individual has value -1 for the first attribute, it has -1 for the second. In essence, there is no distinction between the two attributes, and a principal-components analysis should tell us that there is really only one attribute. In this situation, the population matrix has the form

$$X = \begin{pmatrix} x_1 & x_1 \\ x_2 & x_2 \\ \vdots & \vdots \end{pmatrix} \;,$$

where there are N rows and the xs take on values ± 1. The identity of the two columns expresses the complete correlation of the two attributes. We assume that there are an equal number of $+1$s and -1s in each column, realizing that in a real population, there would be fluctuations about this condition that we would have to worry about in our analysis. The correlation matrix,

$$C = \frac{1}{N} X^T X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \;,$$

has eigenvectors

$$|u_1\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \;, \quad \text{eigenvalue } \lambda_1^2 = 2 \;,$$

$$|u_2\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \;, \quad \text{eigenvalue } \lambda_2^2 = 0 \;.$$

Now we have

$$\sqrt{2N}|n_1\rangle = X|u_1\rangle = \sqrt{2N}\frac{1}{\sqrt{N}}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix} ,$$

$$X|u_2\rangle = 0 .$$

The values of the new attributes are $y_{n,1} = \sqrt{2}x_n$ and $y_{n,2} = 0$, confirming that there is really just one attribute.