El método de Principal Components

Table of Contents

.... 1
Algunas propiedades del principal components 2
Correlación de Principal Components y las variables originales 4

Si ahora tenemos el caso de una matriz de covarianza p

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{1p} & s_{2p} & \cdots & s_p^2 \end{bmatrix}$$

Los datos orginales son columnas para el autor Primero introducimos los datos

```
samples = [
  10.0, 10.7;
  10.4,  9.8;
   9.7, 10.0;
   9.7, 10.1;
  11.7, 11.5;
  11.0, 10.8;
   8.7,  8.8;
   9.5,  9.3;
  10.1,  9.4;
   9.6,  9.6;
  10.5, 10.4;
   9.2,  9.0;
  11.3, 11.6;
  10.1,  9.8;
   8.5,  9.2; ];
```

Función dedicada

```
S = cov( samples );
l = [1.4465, 0.0864];

t = ( S - eye(2) * l(1) );
a = -t(1,1) / t(1,2);
U = [1; a] / sqrt( 1 + a^2 );

t = ( S - eye(2) * l(2) );
a = -t(2,2) / t(2,1);
U = [ U, [a; 1] / sqrt( 1 + a^2 ) ];
```

El equivalente en nuestro caso es

```
z = (samples - repmat(mean(samples), size(samples, 1), 1 ) ) * U;
```

```
z


z =

    0.4831    0.5065
    0.1514   -0.4208
   -0.2171    0.2071
   -0.1481    0.2794
    2.2655   -0.0879
    1.2758   -0.1113
   -1.7689    0.0289
   -0.8450   -0.1615
   -0.3418   -0.5032
   -0.5655   -0.0134
    0.6379   -0.0556
   -1.2691   -0.1715
    2.0450    0.2606
   -0.0657   -0.2137
   -1.6376    0.4564
```

La covarianza de estos datos es igual a

```
cov(z)


ans =

    1.4465    0.0000
    0.0000    0.0864
```

Algunas propiedades del principal components

A partir de U es posible obtener la misma matriz. Las varianzas de las z son posibles de obtener de la misma matriz original. En la segunda forma no es necesario calcular las z y despues calcular la matriz de covarianza y nos estamos ahorrando un paso.

```
U' * S * U


ans =

    1.4465    0.0000
    0.0000    0.0864
```

Lo que estamos haciendo es un promedio ponderado, es posible interpretar los componentes con los signos y esto se relaciona con la variabilidad.

Cuando tenemos ++ es variabilidad en el proceso Cuando tenemos +- es variabilidad en las mediciones

Analisis multivariado de resumen de los resultados como Calculamos el determinante de S

```
det(S)


ans =

    0.1250
```

El determinante de S es la varianza generalizada. La raiz cuadrada de esta cantidad es proporcional al área o volumen generado por un conjunto de datos. Nos interesa que este producto sea pequeño porque eso quiere decir que en nuestro proceso no tenemos tanta variabilidad.

Una propiedad util de PCA es que la variabilidad como medida específica es preservada, ya sea que se mida por las S (varianzas originales) o las L (eigenvalores).

$$|\mathbf{S}| = |\mathbf{L}| = l_1 l_2 \ldots l_p$$

esto es, el determinante de la matriz de covarianza original es igual al producto de las raices características. Por ejemplo

$$|\mathbf{S}| = .1250 = (1.4465)(.0864) = l_1 l_2$$

$$Tr(\mathbf{S}) = Tr(\mathbf{L})$$

```
det(S)
det( [l(1), 0; 0, l(2)] )


ans =

    0.1250


ans =

    0.1250


trace(S)
trace( [l(1),0; 0,l(2)] )


ans =

    1.5329


ans =

    1.5329
```

La segunda identidad es particularmente útil porque muestra que las raíces, que son las varianzas de los componentes principales, pueden tratarsecomo componentes de la varianza. La razón de cada raíz característica al total indicará la proporción de la variabilidad total explicada por cada componente principal.

El componente principal que explica mas varianza es el mas grande Es la variabilidad de cada componente principal

```
diag(U' * S * U)/sum(l)
```

```
ans =

    0.9436
    0.0564
```

Correlación de Principal Components y las variables originales

Es posible determinar la relación de cada pc con cada una de las variables originales. La correlación es igual a

$$r_{zx} = \frac{u_{ji}\sqrt{l_i}}{s_j}$$

```
U.* sqrt( repmat(l,2,1) ./ repmat( diag(S),1,2  ) )
```

```
ans =

    0.9739   -0.2270
    0.9687    0.2482
```

Published with MATLAB® R2021b