El método de Principal Components

Table of Contents

El método

El punto inicial del método de PCA es la matriz de covarianza. Partimos entonces de

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

Tenemos los elementos i-ésimos de las varianzas y covarianzas. Si la covarianza no es igual a cero, esto indica que tenemos una relación entre las dos variables. La fuerza de esta relación esta representada por el coeficiente de correlación $r_{ij} = s_{ij}/s_i s_j$.

La dimensión de la matriz de covarianza, la dimensión de la matriz de eigenvalores, son las mismas, y son matrices cuyos indices corren desde $l_1, l_2, ..., l_p$.

En la clase 3, (sección pasada del libro) se describió un método para transformar p variables correlacionadas $x_1, ..., x_p$, en p nuevas variables no correlacionadas $z_1, ..., z_p$. Los ejes coordenados de estas nuevas variables son descritos por los vectores característico $\mathbf{u}_i$ los cuales hacen la matriz $\mathbf{U}$ de los cosenos directores que se usan en la transformación:

$$\mathbf{z} = \mathbf{U}'[\mathbf{x} - \bar{\mathbf{x}}]$$

Las variables transformadas son conocidas como *componentes principales* de x o *pc*. El i-ésimo componente principal es

$$\mathbf{z}_i = \mathbf{u}_i'[\mathbf{x} - \overline{\mathbf{x}}]$$

La cual tendrá media cero y varianza l_i como raices características. Las variables transformadas son los *componentes principales* y las variables sin transformar son los *z-scores*.

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; ];
```

Calculamos U

```
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 ) ];
```

La primera observación de los datos es x =

```
samples(1,:)
```

```
ans =

   10.0000   10.7000
```

Sustituyendo en la ecuación de z

```
(samples(1,:)-mean(samples)) * U
```

```
ans =

    0.4831    0.5065
```

Tenemos entonces a z_1 y z_2 sus varianzas son l_1 y l_2 respectivamente. Después veremos que $l_1 + l_2$ es igual a la suma de las varianzas de las variables originales.

Podemos calcular todos los valores de z

```
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
```

La varianza de z_1 es l_1 y la varianza de z_2 es l_2.

Algunas propiedades del principal components

Si se quiere transformar un conjunto de variables x mediante una transformación lineal $z = U'[x - \bar{x}]$ ya sea U ortonormal o no, la matriz de covarianza de las nuevas variables $\mathbf{S}_z$, puede ser determinada directamente de la matriz de covarianza de las observaciones originales $\mathbf{S}$. En la segunda forma no es necesario calcular las z y despues calcular la matriz de covarianza y nos estamos ahorrando un paso.

$$\mathbf{S}_z = \mathbf{U}'\mathbf{S}\mathbf{U}$$

```
U' * S * U
```

```
ans =

    1.4465    0.0000
    0.0000    0.0864
```

Sin embargo el hecho de que $\mathbf{U}$ sea ortonormal no es condición suficiente para asegurar que las variables no estan correlacionadas. Solo esta solución de los vectores caracterísiticos, va a producir una matriz $\mathbf{S}_z$ tal que es una matriz diagonal como $\mathbf{L}$ que produce nuevas variables que no tienen correlación.

Lo que estamos haciendo es un promedio ponderado, es posible interpretar los componentes con los signos y esto se relaciona con la variabilidad. Por ejemplo los elementos del primer vector

```
U(:,1)
```

```
ans =

    0.7236
    0.6902
```

son muy similares y ambos positivos, indicando el primer pc z_1 es un promedio ponderado de ambas variables. Esto esta relacionado con la variabilidad que x_1 y x_2 tengan en común. Ya hemos visto que u_1 define la línea de regresión ortogonal a la que Pearson se refirió como la "línea de mejor ajuste".

Los coeficientes del segundo vector también son casi iguales excepto por el signo y, por lo tanto, el segundo pc, z_2, representa las diferencias en las mediciones de los dos métodos que probablemente representarían la variabilidad de las pruebas y las mediciones. La linea definida por u_2 fue referida por Pearson como la "lineal de peor ajuste". Sin embargo, este término es apropiado para el vector característico correspondiente a la raíz característica más pequeña, no para el segundo, a menos que solo haya dos, como es el caso aquí.

De acuerdo con el objetivo del análisis multivariado de resumir los resultados con la menor cantidad de números posible, existen dos cantidades de un solo número para medir la variabilidad general de un conjunto de datos multivariados. Estos son

El determinante de la matriz de covarianza

```
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.

La suma de las variables

$$s_1^2 + s_2^2 + ... + s_p^2 = Tr(\mathbf{S})$$

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).

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

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

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

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

```
ans =

    0.1250
```

```
ans =

    0.1250
```

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

$s_1^2 + s_2^2 = 0.7986 + 0.7343 = 1.539 = l_1 + l_2$

```
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 pc, pueden tratarse como componentes de la varianza. La razón de cada raíz característica al total indicará la proporción de la variabilidad total explicada por cada componente principal. $z_1, 1.4465/1.5329 = .944$ y para $z_2, .0864/1.5329 = .056$. Esto dice que aproximadamente 94% de la variabilidad total de estos datos químicos (representados por $Tr(\mathbf{S})$ está asociado con, explicado o "explicado por" la variabilidad del proceso y $6 \% $ se debe a la variabilidad relacionada con la prueba y la medición. Dado que las raíces características son estimaciones de muestra, estas proporciones también son estimaciones de muestra.

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