6.2 La SVD y Componentes Principales

El objetivo de componentes principales es hacer una reducción de la información disponible. Es decir, la información contenida en p variables predictoras $\mathbf{X}=(X_1,\ldots,X_p)$ puede ser reducida a $\mathbf{Z}=(Z_1,\ldots Z_{p'})$, con p'<=p y donde las Z_i's no son correlacionadas. Los componentes principales poblacionales (Hotelling, 1933) de un vector aleatorio **X** son los elementos de una transformación ortogonal de **X** que tiene correlación 0 entre si.
Más especificamente, sea **X** un vector aleatorio de dimensión p y con matriz de covarianza Σ entonces sus componentes principales son definidos por

$$\mathbf{Z=XV'}$$

donde V' es una matriz ortogonal p x p', tal que Σ=VDV' con D matriz diagonal que contiene los valores propios de Σ.
Los elementos z_1, z_2,.. z_p de Z son llamados el primer principal componente de X, el segundo principal componente y así sucesivamente.
Usando el hecho que Var(Z)=Var(XV')=VΣV'=D se desprende que la varianza del i-ésimo componente principal será igual al i-ésimo valor propio de Σ.
En la práctica el análisis de componentes principales involucra el uso de una muestra aleatoria de tamaño n: $X_1,X_2,\ldots X_n$ de **X** y la matriz de covarianza de **X** es estimada por

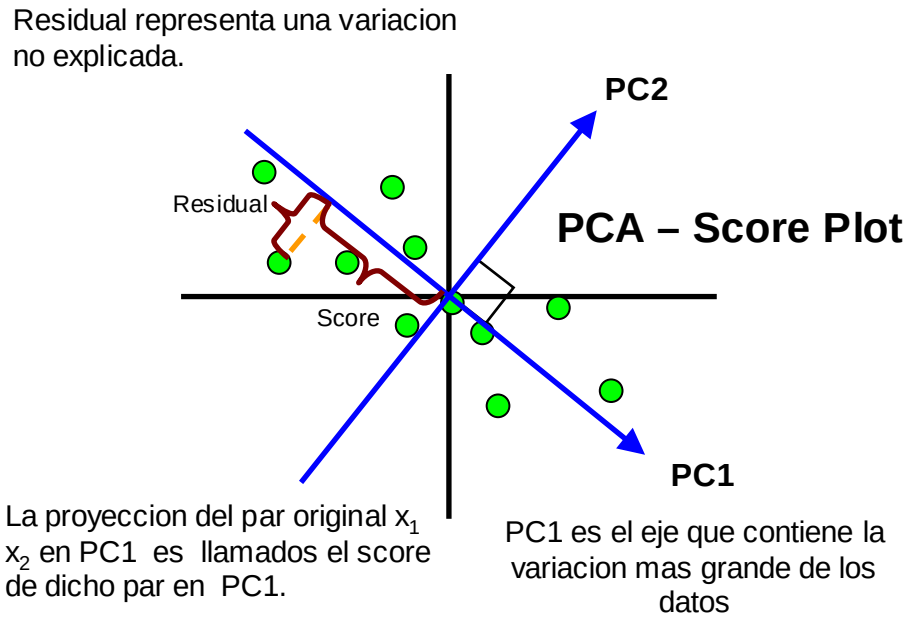
$$\frac{\tilde{X}'\tilde{X}}{n-1}$$

donde $\tilde{X}$ es la matriz que contiene las observaciones de **X** pero con las columnas centradas por la media muestral.
La matriz de Correlaciones es igual a la matriz de covarianza pero con cada entrada estandarizada. Es decir,

Corr(i,j)=cov(i,j)/sqrt(cov(i,i)*cov(j,j))=cov(i,j)/sd(i)sd(j)

La matriz V' la cual es llamada la matriz de cargas (**loadings**) es la misma que se obtiene al aplicar SVD a la matriz X, la cual ha sido centrada previamente). Esto es, X=UDV".
Como Var(Z) y Σ tienen la misma traza (igual a la suma de los valores singulares de X) y determinante (el producto de los valores propios de Σ) se concluyen que Z y X tienen la misma variabilidad total cuando p'=p.
La matriz Z es llamada la matriz de **scores** y contiene los valores originales transformados (rotados).
Las funciones **cov** y **cor** de R y los comandos COV y CORRCOEF de Matlab calculan la matriz de covarianza y de correlaciones respectivamente.
Los componentes principales son calculados de tal manera que la primera componente incluye la mayor cantidad de variabilidad en el conjunto de datos, mientras que las subsiguientes componentes principales describe progresivamente cantidades de variabilidad cada vez mas decreciente.

Geométricamente hablando la aplicación de componentes principales equivale a una rotación de los ejes coordenados. La siguiente figura muestra la transfomacion de una muestra del vector aleatorio (X1,X2) en los componentes principales PC1 y PC2.



Ejemplo: Hallar la matriz de covarianza y de correlaciones del vector $\mathbf{X}=(X_1,X_2,X_3,X_4)$ del cual se han tomado las siguientes 6 observaciones

3	4	8	7
6	7	12	5
9	15	31	6
8	11	20	13
4	9	19	21
9	13	25	14

Solución:
En R,
Primera manera: hallando la covarianza y correlacion por definición

```
> X=rbind(c(3,4,8,7),c(6,7,12,5),c( 9,15, 31, 6),c(8,11,20,13), c(4, 9, 19, 21),c( 9, 13,  25, 14))
> X
     [,1] [,2] [,3] [,4]
[1,]    3    4    8    7
[2,]    6    7   12    5
[3,]    9   15   31    6
[4,]    8   11   20   13
[5,]    4    9   19   21
[6,]    9   13   25   14
```

```
> #centrando las columnas de X
> xc=scale(X,center=T,scale=F)
> xc
     [,1]       [,2]        [,3] [,4]
[1,] -3.5 -5.8333333 -11.1666667   -4
[2,] -0.5 -2.8333333  -7.1666667   -6
[3,]  2.5  5.1666667  11.8333333   -5
[4,]  1.5  1.1666667   0.8333333    2
[5,] -2.5 -0.8333333  -0.1666667   10
[6,]  2.5  3.1666667   5.8333333    3
attr(,"scaled:center")
[1]  6.500000  9.833333 19.166667 11.000000
> #calculando la matriz de covarianza
> cov.x=t(xc)%*%xc/5
> cov.x
     [,1]     [,2]     [,3] [,4]
[1,]  6.7  9.30000 17.70000 -2.0
[2,]  9.3 16.16667 33.23333  3.6
[3,] 17.7 33.23333 70.16667  9.2
[4,] -2.0  3.60000  9.20000 38.0
> #hallando la matriz de correlacion
>xcs=scale(X,center=T,scale=T)
> cor.x=t(xcs)%*%xcs/5
> cor.x
          [,1]      [,2]      [,3]       [,4]
[1,]  1.0000000 0.8935838 0.8163388 -0.1253432
[2,]  0.8935838 1.0000000 0.9867298  0.1452448
[3,]  0.8163388 0.9867298 1.0000000  0.1781684
[4,] -0.1253432 0.1452448 0.1781684  1.0000000
>#Segunda  manera: Usando la función cov y cor.
> cov(X)
     [,1]     [,2]     [,3] [,4]
[1,]  6.7  9.30000 17.70000 -2.0
[2,]  9.3 16.16667 33.23333  3.6
[3,] 17.7 33.23333 70.16667  9.2
[4,] -2.0  3.60000  9.20000 38.0
> cor(X)
          [,1]      [,2]      [,3]       [,4]
[1,]  1.0000000 0.8935838 0.8163388 -0.1253432
[2,]  0.8935838 1.0000000 0.9867298  0.1452448
[3,]  0.8163388 0.9867298 1.0000000  0.1781684
[4,] -0.1253432 0.1452448 0.1781684  1.0000000
>
```

En Matlab,

```
» X=[3 4 8 7 ; 6 7 12 5; 9 15 31 6; 8 11 20 13; 4 9 19 21; 9 13  25 14];
```

» %Segunda manera: hallando la covarianza y correlacion directamente

» cov(X)

ans =

```
   6.7000    9.3000   17.7000   -2.0000
   9.3000   16.1667   33.2333    3.6000
  17.7000   33.2333   70.1667    9.2000
  -2.0000    3.6000    9.2000   38.0000
```

» corrcoef(X)

ans =

```
   1.0000    0.8936    0.8163   -0.1253
   0.8936    1.0000    0.9867    0.1452
   0.8163    0.9867    1.0000    0.1782
  -0.1253    0.1452    0.1782    1.0000
```

R tiene dos funciones que calculan las componentes principales **prcomp** y **printcomp.**
La primera hace uso de la svd de la matriz X (centrada) y la segunda usa los vectores y valores propios de de la matriz de covarianza o correlación de X. Las desviaciones estándar de las componentes que producen las dos funciones son iguales excepto por un factor sqrt(n/(n-1)). Las dos funciones permiten graficar el scree.plot que da una idea de cuantos componentes son convenientes usar.

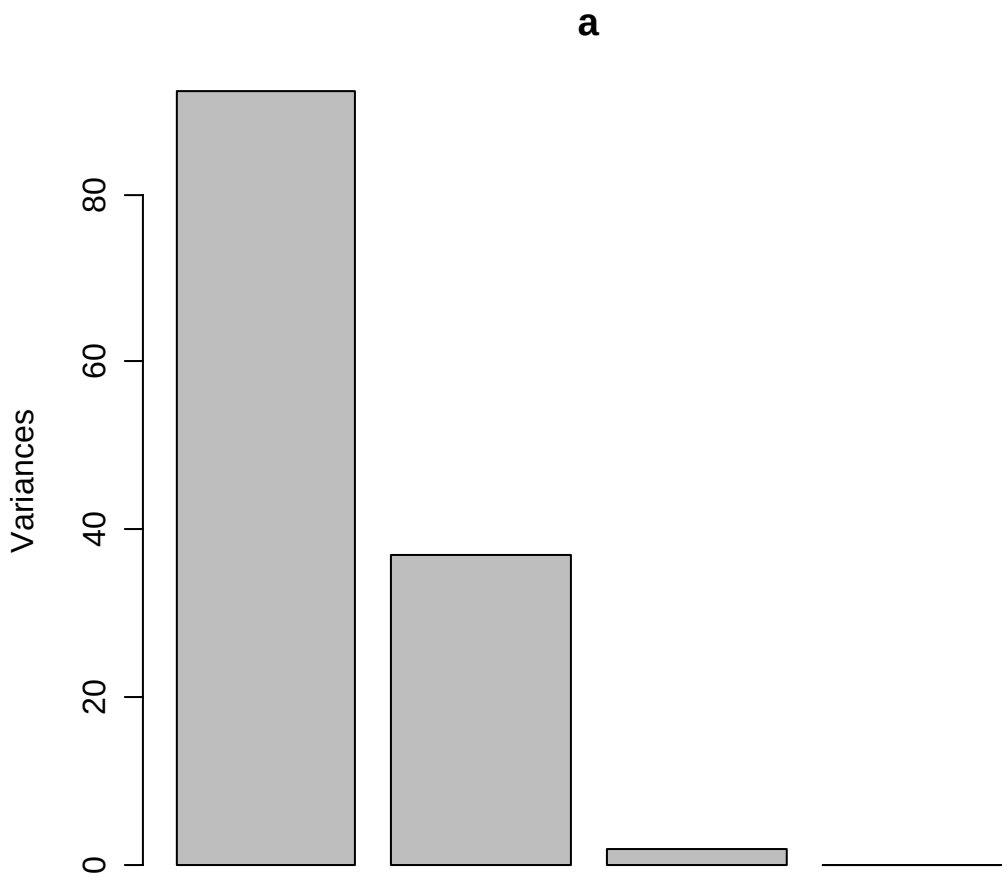
Ejemplo: Hallar los componentes principales de X

Solución:

```
> prcomp(X)
Standard deviations:
[1] 9.6059855 6.0810798 1.3298302 0.1019597

Rotation:
          PC1        PC2        PC3         PC4
[1,] 0.2205009 -0.1537242  0.8697868  0.41378629
[2,] 0.4136288 -0.0776131  0.2902832 -0.85943180
[3,] 0.8675331 -0.1116078 -0.3817417  0.29866913
[4,] 0.1663634  0.9787179  0.1161026  0.03089717
> a=prcomp(X)
>#las desviaciones estandar de las componentes principales son:
> a$sdev
[1] 9.6059855 6.0810798 1.3298302 0.1019597
> Los scores (nuevos valores) son:
```

```
>a$x
           PC1       PC2        PC3         PC4
[1,] -13.5374951 -1.677806 -0.9392012  0.10637290
[2,]  -8.4977001 -4.775685  0.7818371 -0.09768145
[3,]  12.1223264 -6.999594 -1.4235257 -0.02616639
[4,]   1.8689895  1.543294  1.5574312 -0.07130573
[5,]   0.6231022 10.254769 -1.1917201 -0.05907909
[6,]   7.4207772  1.655023  1.2151787  0.14785976
> las cargas (coeficientes de los components principales son:
>a$rot
         PC1        PC2        PC3         PC4
[1,] 0.2205009 -0.1537242  0.8697868  0.41378629
[2,] 0.4136288 -0.0776131  0.2902832 -0.85943180
[3,] 0.8675331 -0.1116078 -0.3817417  0.29866913
[4,] 0.1663634  0.9787179  0.1161026  0.03089717
>
> b=princomp(X)
> attributes(b)
$names
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "call"

$class
[1] "princomp"

> b$sdev
   Comp.1     Comp.2     Comp.3     Comp.4
8.76902493 5.55124096 1.21396331 0.09307604
> b$loadings

Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4
[1,] -0.221  0.154  0.870 -0.414
[2,] -0.414         0.290  0.859
[3,] -0.868  0.112 -0.382 -0.299
[4,] -0.166 -0.979  0.116

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
> b$scores
         Comp.1     Comp.2     Comp.3      Comp.4
[1,]  13.5374951   1.677806 -0.9392012 -0.10637290
[2,]   8.4977001   4.775685  0.7818371  0.09768145
[3,] -12.1223264   6.999594 -1.4235257  0.02616639
[4,]  -1.8689895  -1.543294  1.5574312  0.07130573
```

[5,] -0.6231022 -10.254769 -1.1917201 0.05907909
[6,] -7.4207772 -1.655023 1.2151787 -0.14785976
>plot(a)



El scree.plot recomienda que solo dos componentes principales debian ser usados reduciendo la dimensión de 4 a 2.

En Matlab,

» load X
» %Primera manera. Usando los vectores y valores propios de la matriz de covarianza.
» cov=COV(X)

cov =

```
    6.7000    9.3000   17.7000   -2.0000
    9.3000   16.1667   33.2333    3.6000
   17.7000   33.2333   70.1667    9.2000
   -2.0000    3.6000    9.2000   38.0000
```

```
» [u,d]=eig(cov)

u =

    0.4138   -0.8698   -0.1537    0.2205
   -0.8594   -0.2903   -0.0776    0.4136
    0.2987    0.3817   -0.1116    0.8675
    0.0309   -0.1161    0.9787    0.1664


d =

    0.0104         0         0         0
         0    1.7684         0         0
         0         0   36.9795         0
         0         0         0   92.2750


» varprop=diag(d)./trace(d)

varprop =

    0.0001
    0.0135
    0.2822
    0.7042

» Las cargas de las componentes principales aparecen en la matriz u
» Segunda forma: Aplicando la svd a la matriz centrada de datos

» xcent=zeros(6,4)

xcent =

     0     0     0     0
     0     0     0     0
     0     0     0     0
     0     0     0     0
     0     0     0     0
     0     0     0     0
```

```
» for i=1:4
xcent(:,i)=X(:,i)-mean(X(:,i))
end


xcent =

   -3.5000   -5.8333  -11.1667   -4.0000
   -0.5000   -2.8333   -7.1667   -6.0000
    2.5000    5.1667   11.8333   -5.0000
    1.5000    1.1667    0.8333    2.0000
   -2.5000   -0.8333   -0.1667   10.0000
    2.5000    3.1667    5.8333    3.0000


» [u,s,v]=svd(xcent./sqrt(5))

u =

   -0.6302   -0.1234   -0.3158   -0.4666   -0.1465    0.4987
   -0.3956   -0.3512    0.2629    0.4284    0.6610    0.1747
    0.5644   -0.5148   -0.4787    0.1148    0.0716    0.4111
    0.0870    0.1135    0.5238    0.3128   -0.4641    0.6261
    0.0290    0.7542   -0.4008    0.2591    0.3281    0.3083
    0.3455    0.1217    0.4087   -0.6485    0.4620    0.2545


s =

   9.6060         0         0         0
        0    6.0811         0         0
        0         0    1.3298         0
        0         0         0    0.1020
        0         0         0         0
        0         0         0         0


v =

    0.2205   -0.1537    0.8698   -0.4138
    0.4136   -0.0776    0.2903    0.8594
    0.8675   -0.1116   -0.3817   -0.2987
    0.1664    0.9787    0.1161   -0.0309

»% Las cargas   de las componentes principales aparecen en la matriz v
```

De acuerdo a la contribución a la variabilidad total solamente merecen usarse los dos primeros componentes principales. Es decir

Z1= 0.2205 X1+.4136 X2 +0.8675 X3 +0.1664X4 y
Z2=-0.1537 X1-0.0776 X2 -0.1116 X3 +0.9787
A continuación se muestran los valores transformados de X despues de aplicar componentes principales, estos son llamados **scores**

```
» scores=xcent*v

scores =

  -13.5375   -1.6778   -0.9392   -0.1064
   -8.4977   -4.7757    0.7818    0.0977
   12.1223   -6.9996   -1.4235    0.0262
    1.8690    1.5433    1.5574    0.0713
    0.6231   10.2548   -1.1917    0.0591
    7.4208    1.6550    1.2152   -0.1479

» % cotejando que ahora la correlacion es cero

»corrcoef(scores)

ans =

    1.0000    0.0000   -0.0000   -0.0000
    0.0000    1.0000    0.0000   -0.0000
   -0.0000    0.0000    1.0000   -0.0000
   -0.0000   -0.0000   -0.0000    1.0000
```