Patrones
Tarea 1 **Nota: 7.5 / 10**

Randy Osbaldo Ibarra Cayo

A. Preguntas Cortas

Problema 2 **Nota: 1/1**

Para un estudio se mide la temperatura en diferentes partes del cuerpo de una muestra de personas. Un investigador expresa **todas** las temperaturas en grados Celcius. Otro investigador expresa primero **todas** estas temperaturas en grados Fahrenheit.

- ¿Cómo se relacionan las matrices de Covarianza de ambos datos?
  Solución: Para convertir de grados Celcius a Fahrenheit se usa la transformación lineal $g : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} g(x) &= ax + b \\ &= (\frac{9}{5})x + 32 \end{aligned}$$

  Sea $X \in \mathbb{R}^n$, donde n es el número de temperaturas (Tomadas en diferentes partes del cuerpo) de una persona y $Y = aX + b\mathbf{1}$ como las temperaturas medidas en grados Fahrenheit. Entonces la matriz de covarianza $Cov(Y)$ es:

$$\begin{aligned} Cov(Y) &= E\left[[Y - E(Y)][Y - E(Y)]^T\right] \\ &= E\left[[aX + b\mathbf{1} - E(aX) - E(b\mathbf{1})][aX + b\mathbf{1} - E(aX) - E(b\mathbf{1})]^T\right] \\ &= E\left[[aX + b\mathbf{1} - aE(X) - b\mathbf{1}][aX + b\mathbf{1} - aE(x) - b\mathbf{1}]^T\right] \\ &= E\left[[aX - aE(X)][aX - aE(X)]^T\right] \\ &= E\left[a[X - E(X)]a[X - E(X)]^T\right] \\ &= E\left[a^2[X - E(X)][X - E(X)]^T\right] \\ &= a^2 E\left[[X - E(X)][X - E(X)]^T\right] \\ &= a^2 Cov(X) \end{aligned}$$ **Ok**

  Por lo que concluimos:

$$\begin{aligned} Cov(Y) &= a^2 Cov(X) \\ &= (\frac{9}{5})^2 Cov(X) \\ &= \frac{81}{25} Cov(X) \\ &= (3.24) Cov(X) \end{aligned}$$ **Ok**

- Si ambos deciden proyectar en las direcciones de máxima varianza, ¿Obtendrán las mismas direcciones de proyección?
  Solución: Se vio en clase que

$$\max_{\|l\|=1} Var(l^T X) \iff \max_{l} \frac{l^T Cov(X) l}{\|l\|}$$

es un cociente de Rayleigh, por lo que la solución es el primer vector propio de $Cov(X)$. Luego

$$\max_{\|l\|=1} Var(l^T Y) \iff \max_{l} \frac{l^T Cov(Y) l}{\|l\|}$$

es un cociente de Rayleigh, por lo que la solución es el primer vector propio de $Cov(Y)$. Por otro lado

$$l^T Cov(Y) l = l^T (3.24) Cov(X) l$$

entonces, al ser proporcionales las matrices, sus vectores propios son proporcionales, por lo que ambos obtendrían la misma dirección de máxima varianza. **Ok**

Problema 3 **Nota: 1/1**

Supongamos que $X = (X_1, X_2)$, $Var(X_1) = Var(X_2) = 1$. Calcula los vectores propios de la matriz de covarianza. ¿Qué observas?

Solución:

$$\begin{aligned} Cov(X) &= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{bmatrix} \\ &= \begin{bmatrix} 1 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & 1 \end{bmatrix} \end{aligned}$$ **Ok**

Esta matriz tiene polinomio característico $p(\lambda) = (1-\lambda)^2 - Cov(X_1, X_2)^2$, con raíces en

$$\begin{aligned} \lambda &= \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)[1 - Cov(X_1, X_2)^2]}}{2(1)} \\ &= \frac{2 \pm \sqrt{4 - 4[1 - Cov(X_1, X_2)^2]}}{2} \\ &= \frac{2 \pm \sqrt{4[1 - 1 + Cov(X_1, X_2)^2]}}{2} \\ &= \frac{2 \pm 2\sqrt{1 - 1 + Cov(X_1, X_2)^2}}{2} \\ &= 1 \pm \sqrt{Cov(X_1, X_2)^2} \\ \lambda &= 1 \pm Cov(X_1, X_2) \end{aligned}$$ **Ok**

Notemos entonces que

$$\begin{bmatrix} 1-\lambda & Cov(X_1, X_2) \\ Cov(X_2, X_1) & 1-\lambda \end{bmatrix} = \begin{bmatrix} \pm Cov(X_2, X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & \pm Cov(X_2, X_1) \end{bmatrix}$$

Por lo que al resolver los sistemas

$$\begin{bmatrix} Cov(X_2, X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Cov(X_2, X_1) \end{bmatrix} v = \mathbf{0}$$
$$\begin{bmatrix} -Cov(X_2, X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & -Cov(X_2, X_1) \end{bmatrix} v = \mathbf{0}$$

obtenemos los vectores propios

$$\begin{aligned} v_1^T &= [1, -1] \\ v_2^T &= [1, 1] \end{aligned}$$ **Ok**

Problema 5

Revisa el video sobre la maximización del cociente de Rayleigh. Haz unos pequeños cambios necesarios para demostrar que el segundo vector propio de $Cov(X)$ es la solución del problema de maximizar el cociente bajo la restricción adicional de ser ortogonal al primer vector propio.

Solución:

Falta: Nota 0/2

B. Análisis de Datos

Problema 1 **Nota: 3/3**

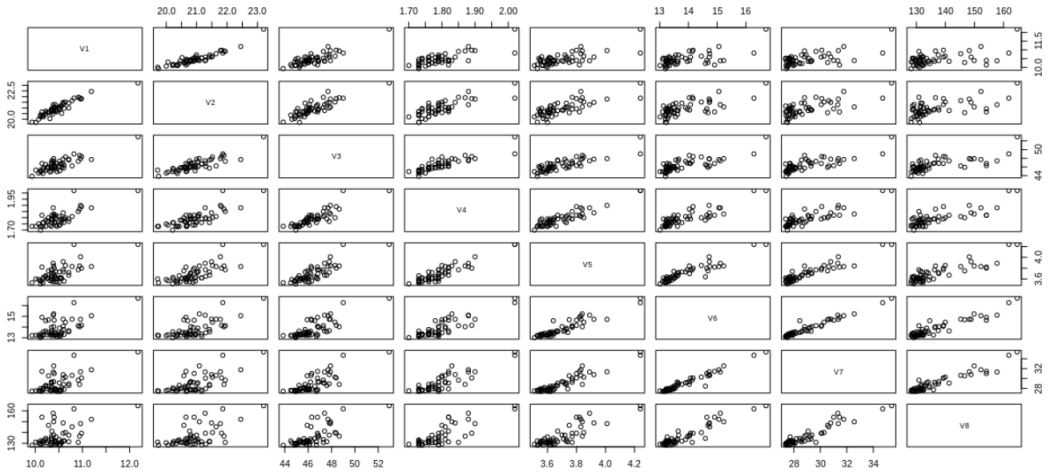
En `deport.dat` se encuentran, **por país**, el mejor tiempo obtenido (hasta 1984) en diferentes pruebas de pista (ordenadas según distancia). Da una descripción general de los datos y usa PCA para entender mejor las diferencias en desempeño deportivo entre los pases.
Solución: Se tienen los siguientes 6 resultados de los 55 países:

```
> print(head(score))
     V1    V2    V3   V4   V5    V6    V7     V8       V9
1 10.39 20.81 46.84 1.81 3.70 14.04 29.36 137.72 argentin
2 10.31 20.06 44.84 1.74 3.57 13.28 27.66 128.30 australi
3 10.44 20.81 46.82 1.79 3.60 13.26 27.72 135.90  austria
4 10.34 20.68 45.04 1.73 3.60 13.22 27.45 129.95  belgium
5 10.28 20.58 45.91 1.80 3.75 14.68 30.55 146.62  bermuda
6 10.22 20.43 45.21 1.73 3.66 13.62 28.62 133.13   brazil
```

Notemos que la novena variable es el nombre del país, la cuál no nos interesa analizar sino que la usamos para reconocer cada observación, por lo que podemos omitirla y usarla para nombrar las observaciones:

```
> print(head(score))
           V1    V2    V3   V4   V5    V6    V7     V8
argentin 10.39 20.81 46.84 1.81 3.70 14.04 29.36 137.72
australi 10.31 20.06 44.84 1.74 3.57 13.28 27.66 128.30
austria  10.44 20.81 46.82 1.79 3.60 13.26 27.72 135.90
belgium  10.34 20.68 45.04 1.73 3.60 13.22 27.45 129.95
bermuda  10.28 20.58 45.91 1.80 3.75 14.68 30.55 146.62
brazil   10.22 20.43 45.21 1.73 3.66 13.62 28.62 133.13
```

Observese que valores pequeños (Para cada una de las variables) son indicativos de un mejor desempeño. A continuación la gráfica de dispersión (Notemos que no se aprecian "huecos")



Ok

y la matriz de correlación

```
> CorX <- cor(score)
> print(CorX)
          V1        V2        V3        V4        V5        V6        V7        V8
V1 1.0000000 0.9226384 0.8411468 0.7560278 0.7002382 0.6194618 0.6325389 0.5199490
V2 0.9226384 1.0000000 0.8507270 0.8066265 0.7749513 0.6953770 0.6965391 0.5961837
V3 0.8411468 0.8507270 1.0000000 0.8701714 0.8352694 0.7786139 0.7872045 0.7049905
V4 0.7560278 0.8066265 0.8701714 1.0000000 0.9180442 0.8635939 0.8690489 0.8064764
V5 0.7002382 0.7749513 0.8352694 0.9180442 1.0000000 0.9281140 0.9346970 0.8655492
V6 0.6194618 0.6953770 0.7786139 0.8635939 0.9281140 1.0000000 0.9746354 0.9321884
V7 0.6325389 0.6965391 0.7872045 0.8690489 0.9346970 0.9746354 1.0000000 0.9431763
V8 0.5199490 0.5961837 0.7049905 0.8064764 0.8655492 0.9321884 0.9431763 1.0000000
```

La matriz de correlación muestra que todos los pares de eventos están correlacionados positivamente y que todos los valores son mayores a $\frac{1}{2}$. Por otro lado calculando los componentes principales se obtiene

```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5    PC6    PC7     PC8
Standard deviation     2.5734 0.9368 0.39915 0.35221 0.28263 0.2607 0.2155 0.15033
Proportion of Variance 0.8278 0.1097 0.01992 0.01551 0.00999 0.0085 0.0058 0.00283
Cumulative Proportion  0.8278 0.9375 0.95739 0.97289 0.98288 0.9914 0.9972 1.00000
> print(pca)
Standard deviations (1, .., p=8):
[1] 2.5733531 0.9368128 0.3991505 0.3522065 0.2826310 0.2607013 0.2154519 0.1503333

Rotation (n x k) = (8 x 8):
         PC1         PC2        PC3         PC4        PC5        PC6          PC7           PC8
V1 0.3175565  0.56687750  0.3322620 -0.12762827  0.2625555 -0.5937042  0.136241260 -0.1055416752
V2 0.3369792  0.46162589  0.3606567  0.25911576 -0.1539571  0.6561367 -0.112639528  0.0960543222
V3 0.3556454  0.24827331 -0.5604674 -0.65234077 -0.2183229  0.1566252 -0.002853707  0.0001272032
V4 0.3686841  0.01242993 -0.5324823  0.47999895  0.5400528 -0.0146918 -0.238016094  0.0381651151
V5 0.3728099 -0.13979665 -0.1534427  0.40451039 -0.4877151 -0.1578430  0.610011482 -0.1392909844
V6 0.3643741 -0.31203045  0.1897643 -0.02958755 -0.2539792 -0.1412987 -0.591298850 -0.5466969221
V7 0.3667726 -0.30685985  0.1817517 -0.08006862 -0.1331764 -0.2190168 -0.176871021  0.7967952190
V8 0.3419261 -0.43896267  0.2632087 -0.29951213  0.4979283  0.3152849  0.398822209 -0.1581638575
```

Observemos que todas las variables aportan una proporción muy parecida en la primera componente principal (Al rededor de 0.35), por lo que la primera componente nos puede indicar el desempeño general del país.

```
> VarX <- cov(score)
> print(VarX)
           V1         V2         V3          V4         V5         V6         V7         V8
V1 0.12350249 0.20902182 0.43069956 0.016920438 0.03836684 0.17441020  0.4018455  1.6860122
V2 0.20902182 0.41557024 0.79905603 0.033115455 0.07788771 0.35913859  0.8117114  3.5462096
V3 0.43069956 0.79905603 2.12290020 0.080743131 0.18974209 0.90887976  2.0734155  9.4778570
V4 0.01692044 0.03311545 0.08074313 0.004055758 0.00911532 0.04406209  0.1000493  0.4739033
V5 0.03836684 0.07788771 0.18974209 0.009115320 0.02430774 0.11592929  0.2634372  1.2451630
V6 0.17441020 0.35913859 0.90887976 0.044062088 0.11592929 0.64185811  1.4115480  6.8910485
V7 0.40184545 0.81171145 2.07341549 0.100049327 0.26343721 1.41154798  3.2678936 15.7321815
V8 1.68601222 3.54620963 9.47785704 0.473903333 1.24516296 6.89104852 15.7321815 85.1381467
```

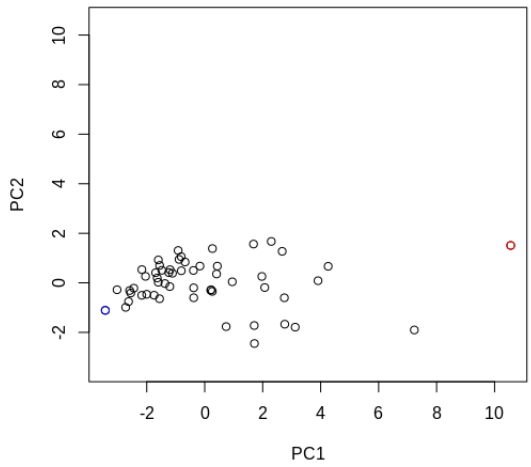
Por otro lado, en la segunda componente podemos notar que la variable 4 es la que está más cerca a 0 (Observemos que la variable 4 es la que menor varianza tiene) por lo que lo que podemos intuir que en la cuarta prueba los países tienen un desempeño parecido. Las variables 1 y 2 son las que más aportan a esta componente, e inversamente las variables 8 y 7, por lo que podemos pensar que los países tienen diferencias en su desempeño en pruebas de tiempo muy corto o tiempo muy largo.

Bien!

Nótese que con las primeras dos compontes ya se acumula más de 93% de la varianza, y con las primeras tres componentes se acumula más de 95%. Sugiero aquí indicar la referencia a la figua donde se ven estos acumulados de variabilidad.

```
> VarX <- cov(score)
> print(VarX)
           V1         V2         V3          V4         V5         V6         V7         V8
V1 0.12350249 0.20902182 0.43069956 0.016920438 0.03836684 0.17441020  0.4018455  1.6860122
V2 0.20902182 0.41557024 0.79905603 0.033115455 0.07788771 0.35913859  0.8117114  3.5462096
V3 0.43069956 0.79905603 2.12290020 0.080743131 0.18974209 0.90887976  2.0734155  9.4778570
V4 0.01692044 0.03311545 0.08074313 0.004055758 0.00911532 0.04406209  0.1000493  0.4739033
V5 0.03836684 0.07788771 0.18974209 0.009115320 0.02430774 0.11592929  0.2634372  1.2451630
V6 0.17441020 0.35913859 0.90887976 0.044062088 0.11592929 0.64185811  1.4115480  6.8910485
V7 0.40184545 0.81171145 2.07341549 0.100049327 0.26343721 1.41154798  3.2678936 15.7321815
V8 1.68601222 3.54620963 9.47785704 0.473903333 1.24516296 6.89104852 15.7321815 85.1381467
```

Podemos elegir entonces representar los datos con las primeras dos componentes principales, por lo que las observaciones proyectadas quedan como sigue



Ok

Podemos observar que el punto con el mayor valor en la primera componente pertenece al país "cookis", el cuál tiene el menor puntaje en todos los eventos, mientras que el punto con menor valor en la primera componente pertenece al país "usa", el cuál tiene el mejor puntaje en la mayoría de los eventos.

Comentarios

- Las variables están ordenadas de acuerdo a la distancia de la prueba por lo que parece que las primeras tres variables están en una misma unidad de medida de tiempo (Posiblente segundos), mietras que las siguientes cinco variables están en otra unidad de medida de tiempo (Posiblemente Minutos), entonces poodemos considerar transformar estas variables. Buen punto. La escala puede sesgar el PCA. Lo mejor es estandarizar los datos antes de aplicar PCA.
- Valores pequeños son indicativos de un mejor desempeño, por lo que podríamos cambiar la "dirección" haciendo que valores grandes sean indicativos de un mejor desempeño.

Problema 2 **Nota: 2.5/3**

Considera los datos `oef2.data`. Se trata de los promedios mensuales de la temperatura (en Celsius) en 35 estaciones canadienses de monitoreo. El interés es comparar las estaciones entre sí en base de sus curvas de temperatura.

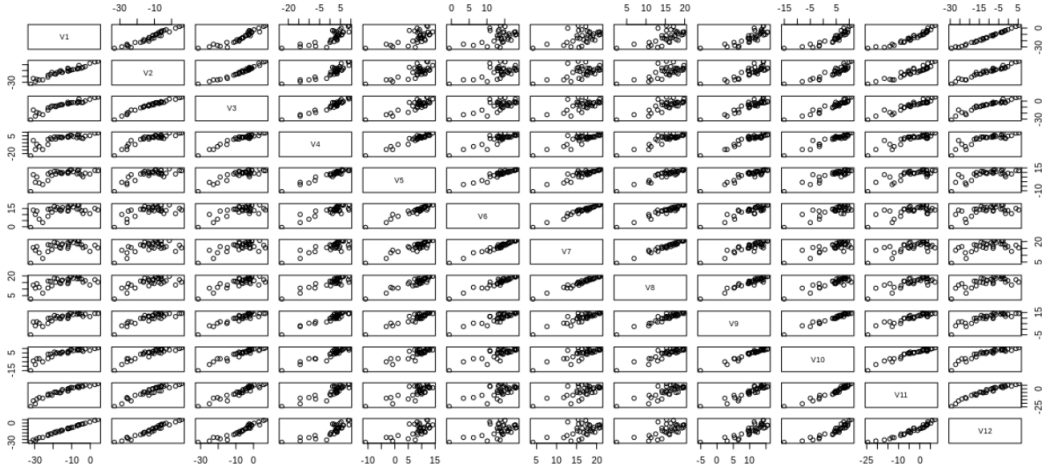
Considerando las 12 mediciones por estación como un vector X, aplica un análisis de componentes principales. Como X representa (un muestreo de) una curva, este tipo de datos se llama datos funcionales. Interpreta y dibuja (como curva) los primeros dos componentes, p_1, p_2 es decir grafica $\{(i, p1_i)\}$ y $\{(i, p2_i)\}$. Agrupa e interpreta las estaciones en el biplot (ten en mente un mapa de Canada).

Solución: Se tienen los siguientes 6 estaciones de las 35 :

```
> print(head(temp))
             [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
St. John_s   -3.9 -4.5 -2.3  1.2  5.4 10.9 15.5 15.3 11.6   6.9   3.4  -1.5
Charlottetown -7.1 -7.5 -3.1  2.3  8.5 14.5 18.3 17.8 13.5   8.1   2.9  -3.9
Halifax      -6.0 -6.1 -1.6  3.3  9.2 14.8 18.2 18.1 13.8   8.6   3.4  -2.9
Sydney       -4.7 -5.9 -2.5  2.0  7.4 13.2 17.7 17.6 13.5   8.4   3.8  -1.8
Yarmouth     -2.7 -3.2  0.3  4.7  9.2 13.4 16.3 16.4 13.6   9.5   5.2  -0.3
Fredericton  -9.2 -8.4 -2.4  4.1 10.8 16.2 19.3 18.2 13.2   7.5   1.4  -6.5
```

A continuación la gráfica de dispersión (Notemos que no se aprecian "huecos")

Pregunta: ¿a qué te refieres exactamente con huecos?



Ok

y la matriz de correlación

```
> CorX <- cor(temp)
> print(CorX)
          V1        V2        V3        V4        V5        V6        V7        V8        V9       V10       V11       V12
V1  1.0000000 0.9690128 0.9308339 0.7896832 0.6455956 0.5018853 0.4818370 0.6195500 0.8113476 0.8724211 0.9530715 0.9961394
V2  0.9690128 1.0000000 0.9782724 0.8799001 0.7472193 0.5806825 0.5308403 0.6545622 0.8332347 0.8944095 0.9328381 0.9762553
V3  0.9308339 0.9782724 1.0000000 0.9529004 0.8545031 0.7163491 0.6703329 0.7721727 0.9105419 0.9480522 0.9498538 0.9497398
V4  0.7896832 0.8799001 0.9529004 1.0000000 0.9652815 0.8627565 0.8096495 0.8712317 0.9410571 0.9444381 0.8725745 0.8257588
V5  0.6455956 0.7472193 0.8545031 0.9652815 1.0000000 0.9541913 0.9019899 0.9208299 0.9231298 0.8849489 0.7662805 0.6869151
V6  0.5018853 0.5806825 0.7163491 0.8627565 0.9541913 1.0000000 0.9778851 0.9549272 0.8813140 0.7914400 0.6515181 0.5416547
V7  0.4818370 0.5308403 0.6703329 0.8096495 0.9019899 0.9778851 1.0000000 0.9804777 0.8826568 0.7850964 0.6508242 0.5247458
V8  0.6195500 0.6545622 0.7721727 0.8712317 0.9208299 0.9549272 0.9804777 1.0000000 0.9504771 0.8811411 0.7728668 0.6618252
V9  0.8113476 0.8332347 0.9105419 0.9410571 0.9231298 0.8813140 0.8826568 0.9504771 1.0000000 0.9768841 0.9205469 0.8440649
V10 0.8724211 0.8944095 0.9480522 0.9444381 0.8849489 0.7914400 0.7850964 0.8811411 0.9768841 1.0000000 0.9661882 0.9042817
V11 0.9530715 0.9328381 0.9498538 0.8725745 0.7662805 0.6515181 0.6508242 0.7728668 0.9205469 0.9661882 1.0000000 0.9692705
V12 0.9961394 0.9762553 0.9497398 0.8257588 0.6869151 0.5416547 0.5247458 0.6618252 0.8440649 0.9042817 0.9692705 1.0000000
```

La matriz de correlación muestra que todos los pares de eventos están correlacionados positivamente y que todos los valores son mayores a $\frac{1}{2}$. Por otro lado calculando los componentes principales se obtiene

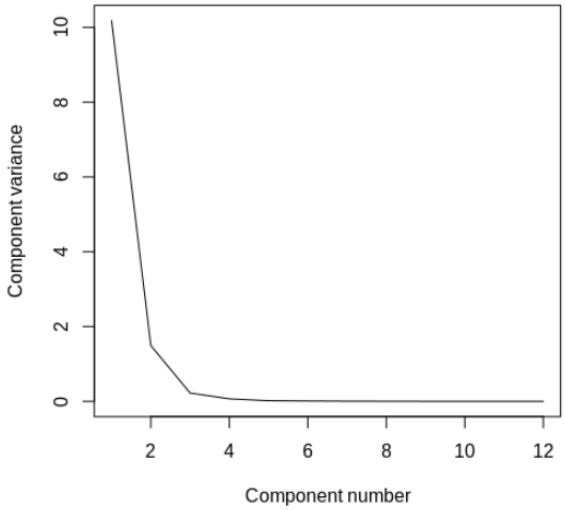
```
> print(pca)
Standard deviations (1, .., p=12):
 [1] 3.19067283 1.22136331 0.47193311 0.25815821 0.13157115 0.09983492 0.07084035 0.05640322 0.03659307 0.02895496 0.02159929 0.01975896

Rotation (n x k) = (12 x 12):
           PC1         PC2        PC3         PC4          PC5         PC6          PC7         PC8         PC9         PC10        PC11        PC12
V1  -0.2724928  0.38864197  0.1669129  0.40164814  0.274420170 -0.12762700  0.028552818 -0.11721833  0.29941821 -0.075413658  0.54819355  0.28899623
V2  -0.2840193  0.32068851 -0.2823290  0.26670227 -0.299948115 -0.20640219  0.150351377  0.41235814 -0.32545303 -0.083821192  0.11475172 -0.46295034
V3  -0.3024111  0.17776321 -0.2837999  0.04638907 -0.224268416  0.48497012 -0.213915326  0.15227393 -0.19608812  0.259967997 -0.11107164  0.56695958
V4  -0.3043605 -0.05148195 -0.4575479 -0.25475797 -0.231240354  0.13659367 -0.200254081 -0.25311593  0.58353316 -0.165818473  0.11571749 -0.26975174
V5  -0.2906974 -0.24656985 -0.4464331 -0.12017579  0.322165895 -0.27939578  0.232302353 -0.41131619 -0.43876161 -0.092727101  0.06198342  0.17607754
V6  -0.2663673 -0.41940374 -0.1242702  0.32158463  0.495860412  0.08865122 -0.001904418  0.35886459  0.24611146  0.356486456 -0.17010700 -0.18755097
V7  -0.2600561 -0.44041802  0.2575670  0.28104053 -0.215566837  0.27907596  0.187454507  0.05388226 -0.03397013 -0.643713629 -0.05484568  0.11024548
V8  -0.2843447 -0.31342729  0.3382340  0.06846111 -0.467566507 -0.13452413  0.073008758 -0.29483078 -0.06300819  0.558869510  0.22108027 -0.08757823
V9  -0.3086486 -0.08632847  0.2394660 -0.16179970  0.078986459 -0.36416003 -0.770532774  0.14881321 -0.16550818 -0.171414175 -0.04295856  0.02053361
V10 -0.3083248  0.04508856  0.1589914 -0.56535655 -0.029531571 -0.24542489  0.437435663  0.42188890  0.23616484  0.013942825 -0.05898865  0.26096550
V11 -0.2960035  0.22103108  0.3196341 -0.33295658  0.330261895  0.53798523  0.053837609 -0.12394283 -0.25955018  0.015516405  0.12228512 -0.38881757
V12 -0.2812635  0.35302259  0.1492018  0.20566538 -0.004058257 -0.15192249  0.091164969 -0.35450651  0.12933630  0.002531649 -0.74541720 -0.04735601
```

Observemos que todas las variables aportan una proporción muy parecida en la primera componente principal (Al rededor de -0.3) Por otro lado, las temperaturas de los meses 1, 6, 7 y 12, correspondientes a las estaciones de Invierno y Verano, son las que mayor aportan a la segunda componente.

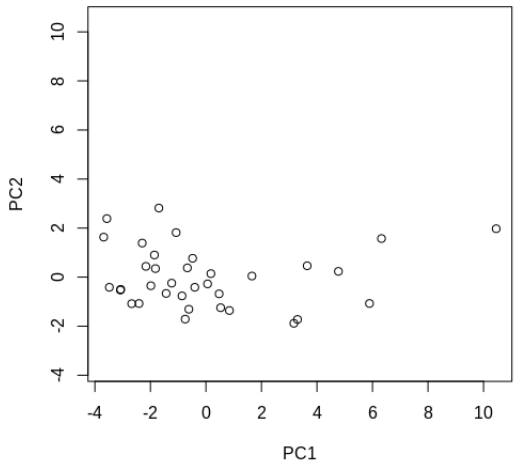
```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     3.1907 1.2214 0.47193 0.25816 0.13157 0.09983 0.07084 0.05640 0.03659 0.02895 0.02160 0.01976
Proportion of Variance 0.8484 0.1243 0.01856 0.00555 0.00144 0.00083 0.00042 0.00027 0.00011 0.00007 0.00004 0.00003
Cumulative Proportion  0.8484 0.9727 0.99124 0.99679 0.99823 0.99906 0.99948 0.99975 0.99986 0.99993 0.99997 1.00000
```

Nótese que con las primeras dos compontes ya se acumula más de 97% de la varianza, y con las primeras tres componentes se acumula más de 99%.



Ok

Podemos elegir entonces representar los datos con las primeras dos componentes principales, por lo que las observaciones proyectadas quedan como sigue

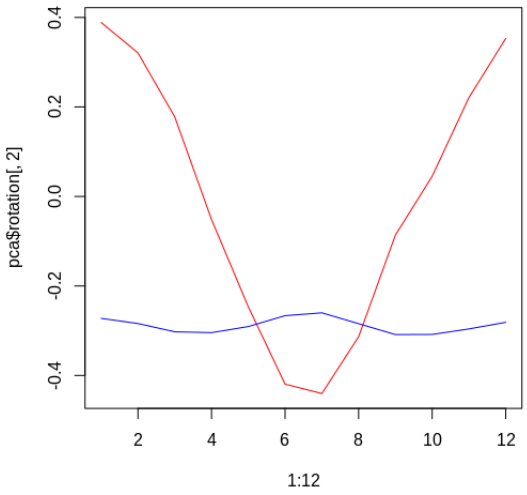


Ok

Observemos que la estación con mayor valor en la primera componente principal está ubicada en "Resolute" (Lo más al norte de Canadá), mientras que las menores están en "Vancouver" y "Victoria" (Sur de Canadá). Por otro lado, los extremos en la segunda componente son "Prince Rupert" con el mayor valor y "Dawson" con el menor valor en la segunda componente, ambas estaciones están del lado de Alaska.

De hecho, se puede mostrar que hay cierta correlación entre la distribución geográfica de las escationes y lo que ocurre en las primeras componentes principales.

Por ultimo, graficamos las primeras dos componentes principales



Ok

Falta el biplot (-0.5).