# MAT-269: Componentes Principales Probabilisticas

**Felipe Osorio**

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

# Componentes principales probabilísticas

## Objetivo:

Introduce un modelo estadístico para PCA que tiene una cercanía con el modelo de análisis factorial.

## Características:

- Estimación máximo verosímil de las PC (así como de sus errores estándar).
- Estimación puede ser llevada a cabo eficientemente via un algoritmo EM.
- Permite usar la maquinaria de modelamiento estadístico para, por ejemplo, desarrollar test de hipótesis y aplicar métodos Bayesianos.

# Componentes principales probabilísticas

El método fue introducido independientemente por Roweis (1998) y Tipping y Bishop (1999), esencialmente aprovechando la relación con el modelo de análisis factorial

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon},$$

donde $\boldsymbol{W} \in \mathbb{R}^{p \times q}$ permite relacionar los dos conjuntos de variables. Asumiremos que

$$\boldsymbol{x} \sim \mathsf{N}_q(\boldsymbol{0}, \boldsymbol{I}), \qquad \boldsymbol{\epsilon} \sim \mathsf{N}_p(\boldsymbol{0}, \phi\boldsymbol{I}).$$

Esto lleva al modelo marginal,

$$\boldsymbol{y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \phi\boldsymbol{I}). \tag{1}$$

*Observación:*

PCA es un caso límite del modelo en (1) tomando $\lim_{\phi \to 0} \phi\boldsymbol{I}$.

# Componentes principales probabilísticas

Las PCA probabilísticas (PPCA) consideran el modelo

$$\boldsymbol{y}_i|\boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathsf{N}_p(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{x}, \phi\boldsymbol{I}), \qquad \boldsymbol{x}_i \overset{\text{ind}}{\sim} \mathsf{N}_q(\boldsymbol{0}, \boldsymbol{I}),$$

para $i = 1, \ldots, n$. Lo que lleva al modelo

$$\boldsymbol{y}_i \overset{\text{ind}}{\sim} \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^\top + \phi\boldsymbol{I},$$

con función de log-verosimilitud

$$\ell(\boldsymbol{\theta}) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}(\boldsymbol{\mu}),$$

donde $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \operatorname{vec}^\top \boldsymbol{W}, \phi)^\top$, con

$$\boldsymbol{S}(\boldsymbol{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top.$$

# Componentes principales probabilísticas

El MLE de $\boldsymbol{\mu}$ en PPCA es dado por $\overline{\boldsymbol{y}}$, en cuyo caso $\boldsymbol{S} = \boldsymbol{S}(\widehat{\boldsymbol{\mu}})$ es la matriz de covarianza muestral de $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$.

Estimaciones de $\boldsymbol{W}$ y $\phi$ pueden ser obtenidos iterativamente usando un algoritmo EM (Rubin y Thayer, 1982).

Además,

$$\boldsymbol{x}_i | \boldsymbol{y}_i \overset{\text{ind}}{\sim} \mathsf{N}_q(\boldsymbol{M}^{-1}\boldsymbol{W}^\top(\boldsymbol{y}_i - \boldsymbol{\mu}), \phi\boldsymbol{M}^{-1}), \qquad i = 1, \ldots, n,$$

con $\boldsymbol{M} = \boldsymbol{W}^\top\boldsymbol{W} + \phi\boldsymbol{I}$.

# Algoritmo EM para PPCA

Asumiremos que $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ son no observables (missing). De este modo, la función de log-verosimilitud de datos completos adopta la forma:

$$\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) = -\frac{np}{2} \log 2\pi\phi - \frac{1}{2\phi} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x}_i\|^2 - \frac{nq}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2.$$

En la etapa E del algoritmo debemos calcular (desconsiderando términos que no involucran $\boldsymbol{\theta}$)

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = -\frac{np}{2} \log \phi - \frac{1}{2\phi} \sum_{i=1}^{n} \mathsf{E}\left[\|\boldsymbol{y}_i - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x}_i\|^2 | \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)}\right].$$

Sea,

$$\boldsymbol{x}_i^{(k)} = \mathsf{E}[\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{(k)}] = [\boldsymbol{M}^{(k)}]^{-1} \boldsymbol{W}^{(k)\top} (\boldsymbol{y}_i - \boldsymbol{\mu}^{(k)}),$$

esto permite escribir

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = -\frac{np}{2} \log \phi - \frac{n\phi^{(k)}}{2\phi} \operatorname{tr}[\boldsymbol{M}^{(k)}]^{-1} \boldsymbol{W}^{\top}\boldsymbol{W} - \frac{1}{2\phi} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x}_i^{(k)}\|^2.$$

# Algoritmo EM para PPCA

Substituyendo $\boldsymbol{\mu}^{(k)}$ por $\overline{\boldsymbol{y}}$ sigue que

$$\boldsymbol{W}^{(k+1)} = \Big\{ \sum_{i=1}^{n} (\boldsymbol{y}_i - \overline{\boldsymbol{y}}) \boldsymbol{x}_i^{(k)\top} \Big\} \Big( \sum_{i=1}^{n} \boldsymbol{x}_i^{(k)} \boldsymbol{x}_i^{(k)\top} \Big)^{-1}$$

$$\phi^{(k+1)} = \frac{1}{np} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{\mu} - \boldsymbol{W}^{(k+1)} \boldsymbol{x}_i^{(k)}\|^2 + \frac{\phi^{(k)}}{p} \operatorname{tr}[\boldsymbol{M}^{(k)}]^{-1} \boldsymbol{W}^{(k)\top} \boldsymbol{W}^{(k)}.$$

Usando la definición de $\boldsymbol{x}_i^{(k)}$ (y de $\operatorname{Cov}(\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{\theta})$), y haciendo $\boldsymbol{W} = \boldsymbol{W}^{(k)}$, $\phi = \phi^{(k)}$, podemos reescribir la etapa M del algoritmo como:

$$\boldsymbol{W}^{(k+1)} = \boldsymbol{S}\boldsymbol{W}(\phi \boldsymbol{I} + \boldsymbol{M}^{-1} \boldsymbol{W}^\top \boldsymbol{S}\boldsymbol{W})^{-1}$$

$$\phi^{(k+1)} = \frac{1}{p} \operatorname{tr}(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W}\boldsymbol{M}^{-1} \boldsymbol{W}^{(k)\top}),$$

donde

$$\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^\top.$$

# Referencias


Chen, T., Martin, E., Montague, G. (2009).
Robust probabilistic PCA with missing data and contribution analysis for outlier detection.
*Computacional Statistics and Data Analysis* **53**, 3706-3716.

Roweis, S. (1998).
EM algorithms for PCA and SPCA.
*Advances in Neural Information Processing Systems* 626-632.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J. (2007).
pcaMethods - a Bioconductor package providing PCA methods for incomplete data.
*Bioinformatics* **23**, 1164-1167.

Tipping, M.E., Bishop, C.M. (1999).
Probabilistic principal component analysis.
*Journal of the Royal Statistical Society, Series B* **61**, 611-622.

Tipping, M.E., Bishop, C.M. (1999).
Mixtures of probabilistic principal component analysers.
*Neural Computation* **11**, 443-482.


# PCA para datos de contaminación de aire

**Felipe Osorio**

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

# Datos de contaminación del aire

Contaminación del aire en 41 ciudades americanas, en donde se midió las siguientes variables:

- SO2: dióxido de azufre en el aire en microgramos por metro cúbico.
- Temp: temperatura anual media (ºF).
- Manuf: Número de empresas manufacturadoras con más de 20 trabajadores.
- Pop: población en miles de habitantes (censo de 1970).
- Wind: velocidad del viento (en millas por hora).
- Precip: promedio de precipitaciones anual (en pulgadas).
- Days: promedio de días con precipitaciones por año.

# Conjunto de datos

```
> pollution
                SO2 Neg.Temp Manuf  Pop Wind Precip Days
Phoenix          10    -70.3   213  582  6.0   7.05   36
Little Rock      13    -61.0    91  132  8.2  48.52  100
San Francisco    12    -56.7   453  716  8.7  20.66   67
Denver           17    -51.9   454  515  9.0  12.95   86
Hartford         56    -49.1   412  158  9.0  43.37  127
Wilmington       36    -54.0    80   80  9.0  40.25  114
Washington       29    -57.3   434  757  9.3  38.89  111
Jacksonville     14    -68.4   136  529  8.8  54.47  116
Miami            10    -75.5   207  335  9.0  59.80  128
Atlanta          24    -61.5   368  497  9.1  48.34  115
Chicago         110    -50.6  3344 3369 10.4  34.44  122
Indianapolis     28    -52.3   361  746  9.7  38.74  121
Des Moines       17    -49.0   104  201 11.2  30.85  103
Wichita           8    -56.6   125  277 12.7  30.58   82

...

Charleston       31    -55.2    35   71  6.5  40.75  148
Milwaukee        16    -45.7   569  717 11.8  29.07  123
```

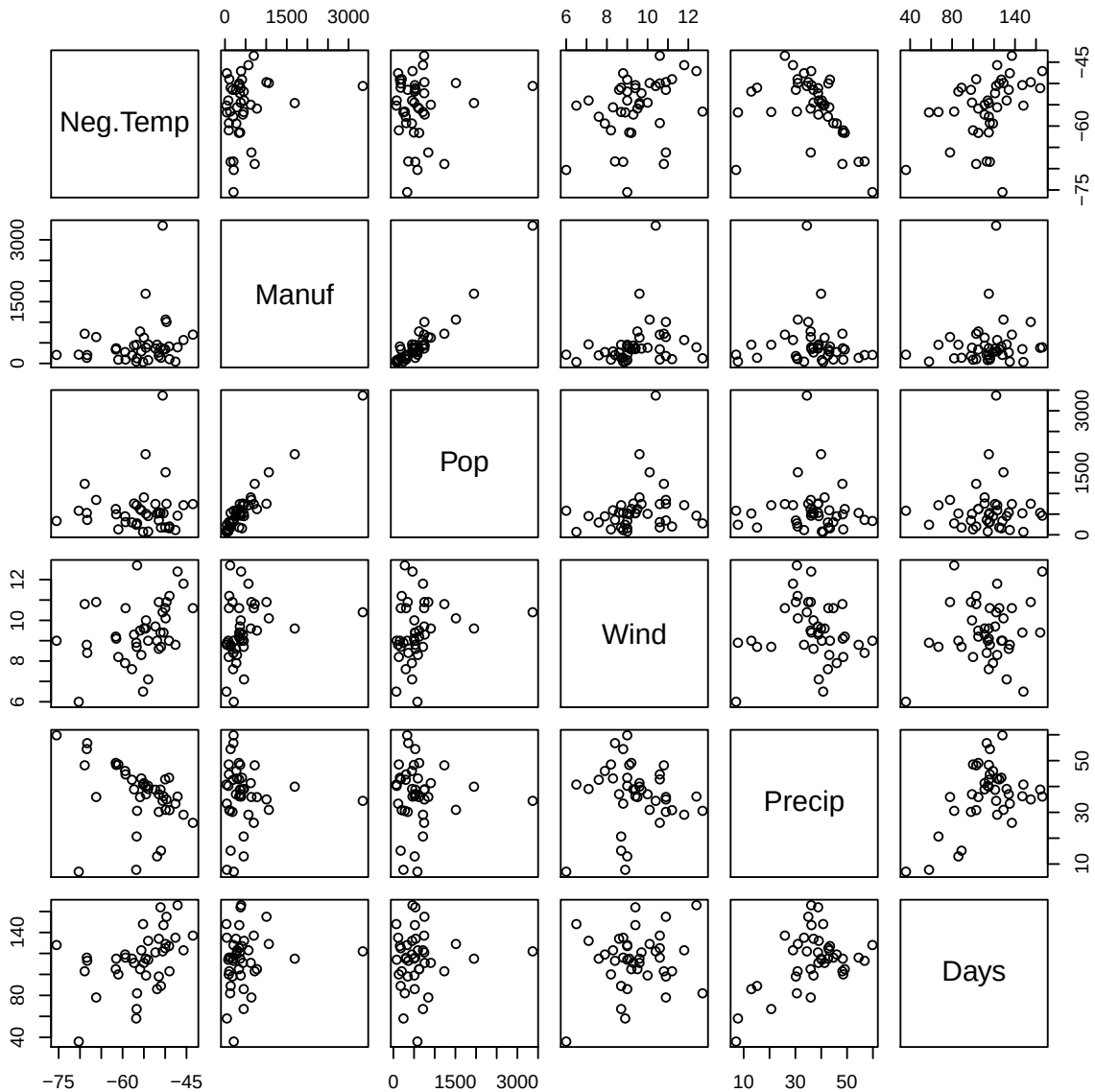
# Datos de contaminación del aire

En los análisis siguientes ignoraremos SO2. Note además:

- Pop, Manuf: están relacionadas a la ecología humana.
- Manuf: Número de empresas manufacturadoras con más de 20 trabajadores.
- Temp,Wind,Precip,Days: están asociadas al clima.

Primeramente realizamos un análisis exploratorio de los datos mediante el comando `pairs`.

# Gráfico de dispersión multivariado



# Datos de contaminación del aire

Podemos apreciar que los datos tienen muy diferentes escalas de modo que, obtenemos las PC basados en la matriz de correlación. Considere:

```
> names(pollution)
[1] "SO2"  "Neg.Temp"  "Manuf"  "Pop"  "Wind"  "Precip"  "Days"

> db <- pollution[,-1]
> cor(db)

          Neg.Temp    Manuf      Pop    Wind  Precip    Days
Neg.Temp    1.0000   0.1900   0.0627   0.350 -0.3863  0.4302
Manuf       0.1900   1.0000   0.9553   0.238 -0.0324  0.1318
Pop         0.0627   0.9553   1.0000   0.213 -0.0261  0.0421
Wind        0.3497   0.2379   0.2126   1.000 -0.0130  0.1641
Precip     -0.3863  -0.0324  -0.0261  -0.013  1.0000  0.4961
Days        0.4302   0.1318   0.0421   0.164  0.4961  1.0000
```

# Datos de contaminación del aire

Podemos apreciar que los datos tienen muy diferentes escalas de modo que, obtenemos las PC basados en la matriz de correlación. Considere:

```
> pc <- princomp(db, cor = TRUE)
> summary(pc)
   Importance of components:
                             Comp.1    Comp.2    Comp.3
   Standard deviation     1.4819456 1.2247218 1.1809526
   Proportion of Variance 0.3660271 0.2499906 0.2324415
   Cumulative Proportion  0.3660271 0.6160177 0.8484592
                             Comp.4    Comp.5    Comp.6
   Standard deviation     0.8719099 0.3384829 0.1855998
   Proportion of Variance 0.1267045 0.0190951 0.0057412
   Cumulative Proportion  0.9751637 0.9942588 1.0000000
```

# Datos de contaminación del aire

En efecto, note que

```
> R <- cor(db)
> rs <- eigen(R)
> prop <- rs$values / sum(rs$values)

> sqrt(rs$values) # standard deviation
[1] 1.48195 1.22472 1.18095 0.87191 0.33848 0.18560

> prop # proportion of variance
[1] 0.36603 0.24999 0.23244 0.12670 0.01910 0.00574

> cumsum(prop) # cumulative proportion
[1] 0.36603 0.61602 0.84846 0.97516 0.99426 1.00000
```

# Datos de contaminación del aire

```
> summary(pc, loadings = TRUE)
  Importance of components:
                            Comp.1    Comp.2    Comp.3
  Standard deviation     1.4819456 1.2247218 1.1809526
  Proportion of Variance 0.3660271 0.2499906 0.2324415
  Cumulative Proportion  0.3660271 0.6160177 0.8484592
                            Comp.4    Comp.5    Comp.6
  Standard deviation     0.8719099 0.3384829 0.1855998
  Proportion of Variance 0.1267045 0.0190951 0.0057412
  Cumulative Proportion  0.9751637 0.9942588 1.0000000

  Loadings:
            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
  Neg.Temp -0.330  0.128  0.672  0.306  0.558  0.136
  Manuf    -0.612 -0.168 -0.273  0.137  0.102 -0.703
  Pop      -0.578 -0.222 -0.350                0.695
  Wind     -0.354  0.131  0.297 -0.869 -0.113
  Precip          0.623 -0.505 -0.171  0.568
  Days     -0.238  0.708         0.311 -0.580
```

# Datos de contaminación del aire

Desde la descomposición espectral de $\boldsymbol{R}$, tenemos que:

```
> rs$vectors
           [,1]   [,2]    [,3]    [,4]    [,5]    [,6]
[1,] -0.3296  0.128  0.6717  0.3065  0.5581  0.1362
[2,] -0.6115 -0.168 -0.2729  0.1368  0.1020 -0.7030
[3,] -0.5778 -0.222 -0.3504  0.0725 -0.0781  0.6946
[4,] -0.3538  0.131  0.2973 -0.8694 -0.1133 -0.0245
[5,]  0.0408  0.623 -0.5046 -0.1711  0.5682  0.0606
[6,] -0.2379  0.708  0.0931  0.3113 -0.5800 -0.0220

> apply(rs$vectors, 2, function(x) sum(x^2)) # sums of squares
 [1] 1 1 1 1 1 1
```

## Interpretación:

- 1ª Comp: "calidad de vida" con valores altos para variables Pop y Manuf.
- 2ª Comp: "clima húmedo" destacándo las variables Precip y Days.
- 3ª Comp: "tipo de clima" provee un contraste entre Precip y Neg.Temp.

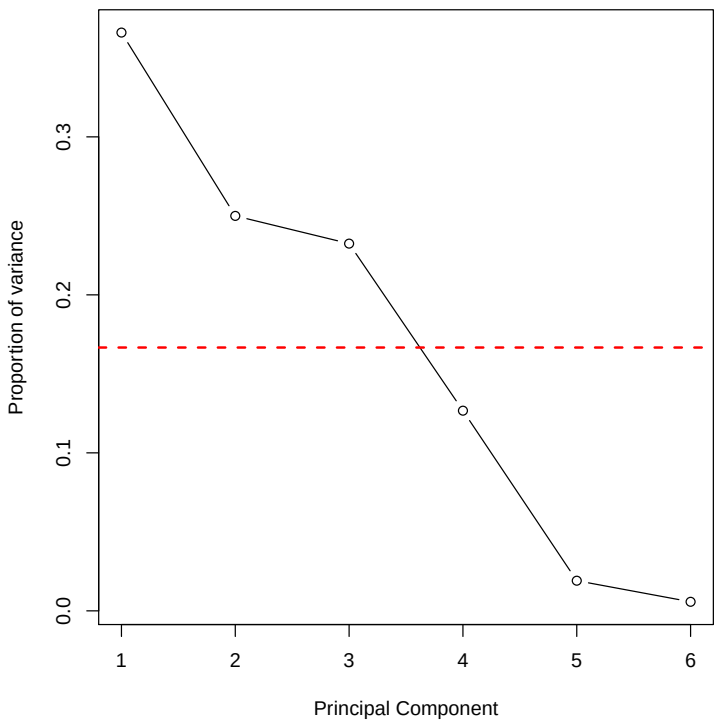
# Datos de contaminación del aire

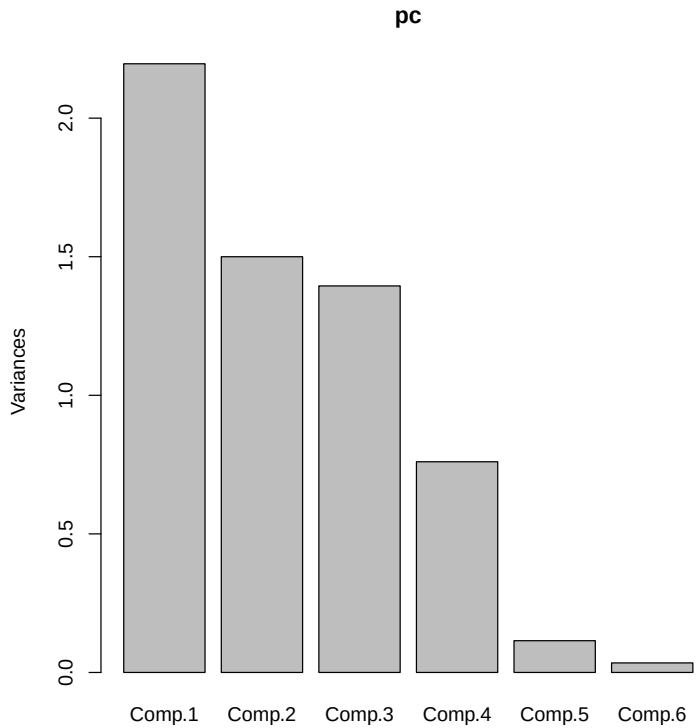
Elección del número de componentes:

- Se ha sugerido realizar el gráfico de la proporción de varianza acumulada.
- Kaiser (1958) propuso considerar aquellas componentes cuyos valores propios sean mayor que su promedio.

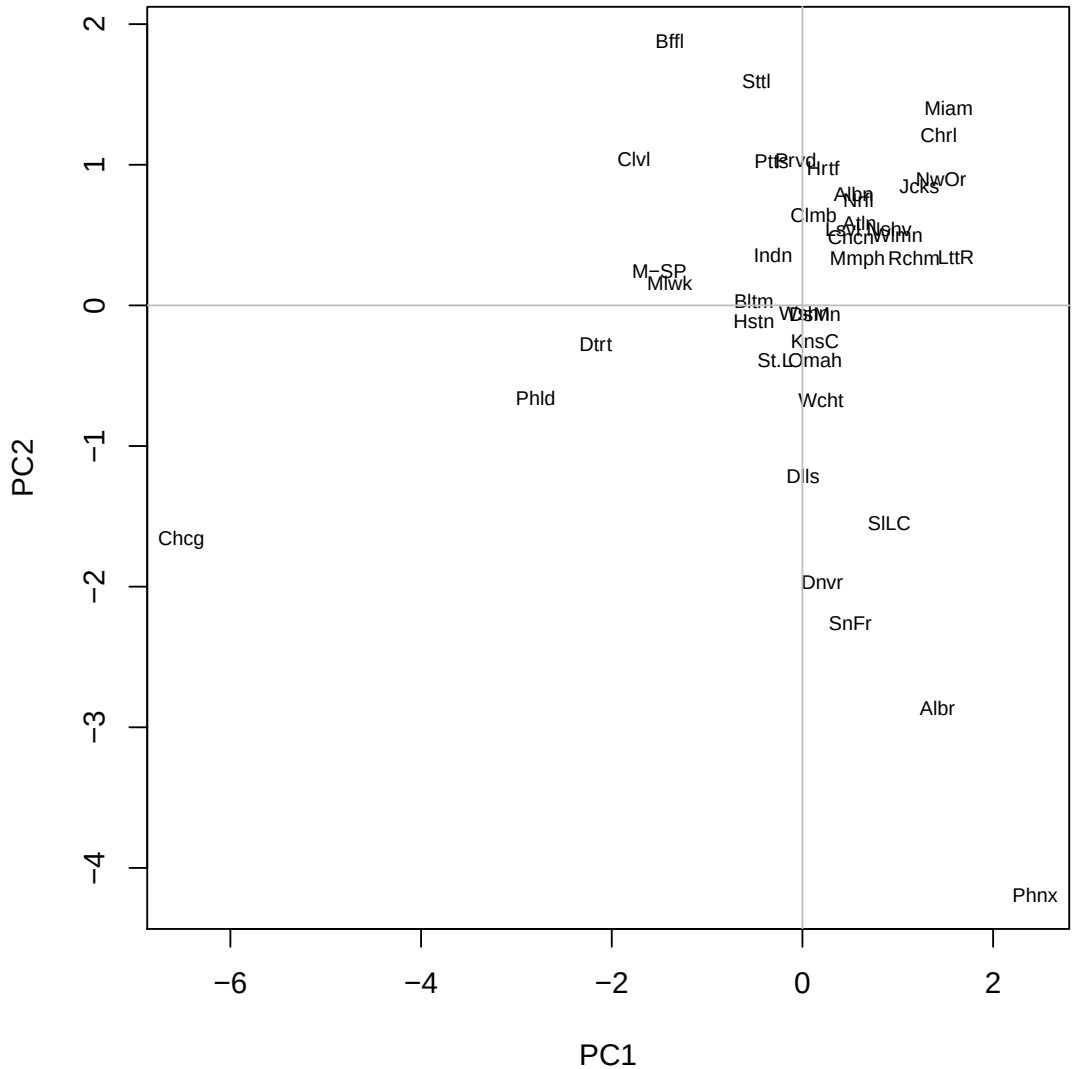
```
# 1er panel
> plot(prop, xlab = "Principal Component",
     + ylab = "Proportion of variance", type = "b")
> abline(h = mean(prop), col = "red", lwd = 2, lty = 2)

# 2do panel
> plot(pc)
```

# Datos de contaminación del aire

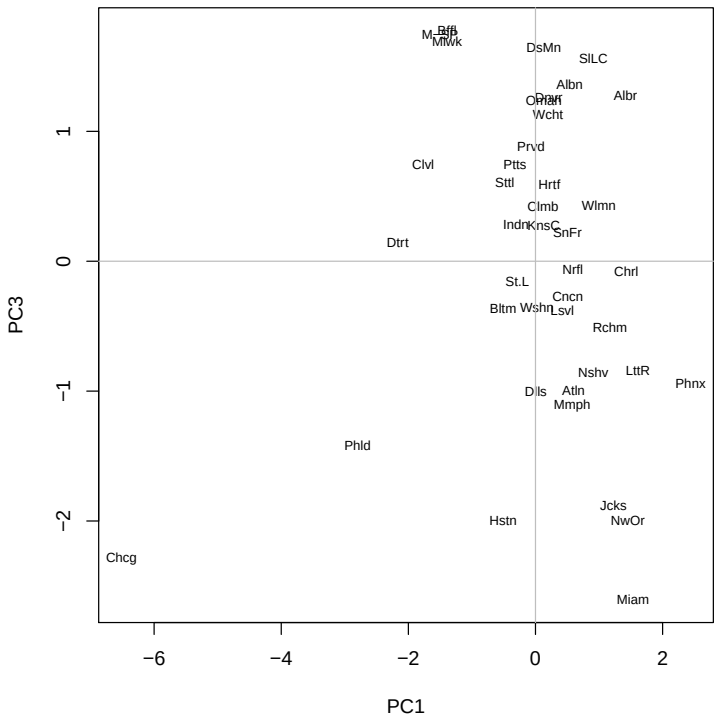


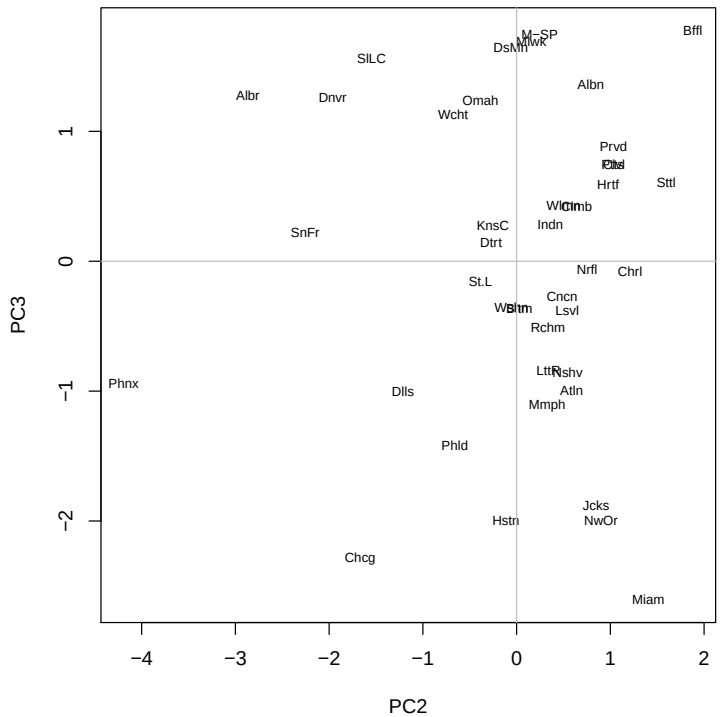



# Datos de contaminación del aire



# Datos de contaminación del aire



# Datos de contaminación del aire

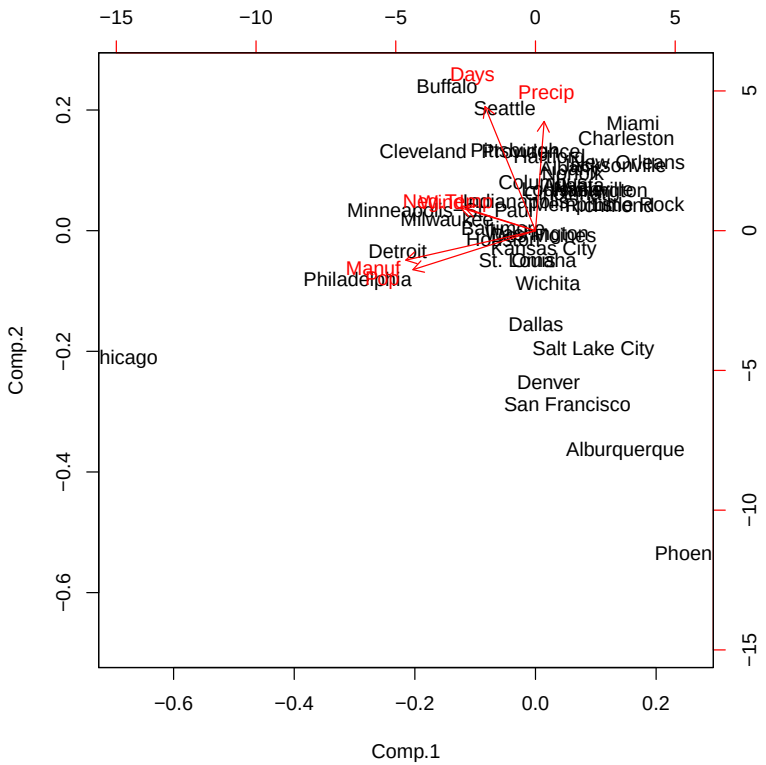
Gráficos anteriores se realizaron mediante los comandos:

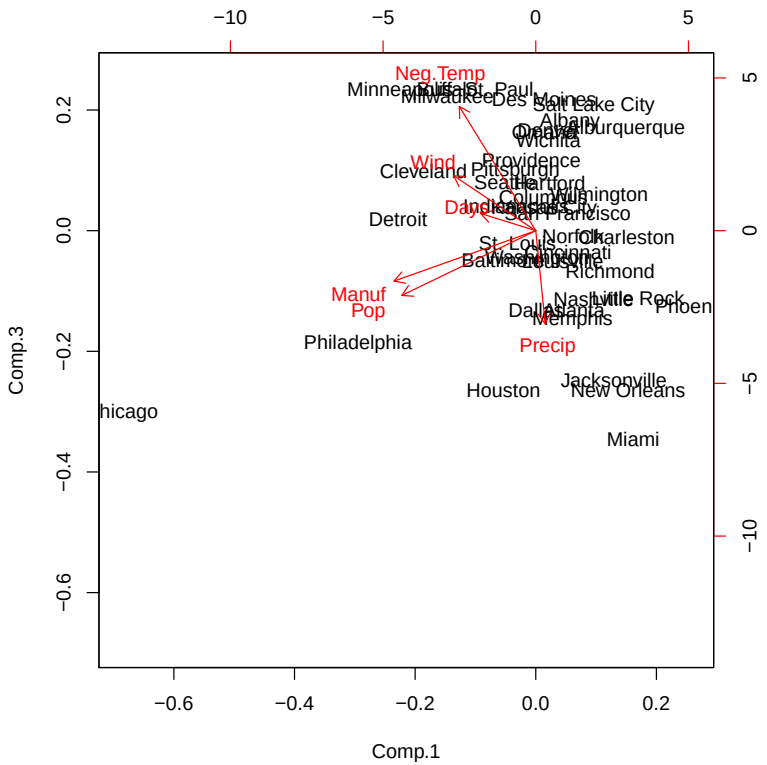
```
> par(pty = "s")
> plot(pc$scores[,1], pc$scores[,2], type = "n", xlab = "PC1",
      + ylab = "PC2")
> text(pc$scores[,1], pc$scores[,2], labels = abbreviate(row.names(db)),
      + cex = 0.7, lwd = 2)
> abline(h = 0, col = "gray")
> abline(v = 0, col = "gray")
```

Note que información equivalente puede ser obtenida con la función `biplot`:

```
> biplot(pc, choices = 1:2)
> biplot(pc, choices = c(1,3))
```

# Datos de contaminación del aire



# Datos de contaminación del aire

Note que las componentes principales pueden ser calculadas como:

$$\boldsymbol{Y} = (\boldsymbol{X} - \mathbf{1}_n \overline{\boldsymbol{x}}^\top)\boldsymbol{T},$$

donde $\boldsymbol{T}$ es matriz ortogonal (en nuestro ejemplo, obtenida desde la descomposición espectral $\boldsymbol{R} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^\top$). En los siguientes comandos la matriz $\boldsymbol{X}$ es centrada y escalada

```
> z <- scale(db) # centrado y escalado de la matriz de datos
> scores <- z %*% rs$vectors
> scores
```

Los vectores propios corresponden a los loadings obtenidas desde la función `princomp` (salvo un factor de escala y/o un signo).

# Otras herramientas para PCA

- `prcomp`: utiliza SVD para realizar los cálculos (es un poco más eficiente que su rutina hermana `princomp`).
- `princomp`: basado en la descomposición espectral (`eigen`), puede no ser recomendable para problema de gran dimensión.
- `PCA`: desde el paquete `FactoMineR`, posiblemente la mejor herramienta disponible en R para PCA.
- `ppca`: desde el paquete `pcaMethods`,[1] para Bioconductor[2] ofrece una multitud de procedimientos para PCA, entre ellos probabilistic PCA.

[1] URL: `https://bioconductor.org/packages/pcaMethods/`

[2] Paquetes de Bioconductor pueden ser instalados sin mucha dificultad en R.