UFM UNIVERSIDAD FRANCISCO MARROQUÍN
VERITAS • LIBERTAS • IVSTITIA

FACULTAD de CIENCIAS ECONÓMICAS

Interpretación del PCA

Alan Reyes-Figueroa

Elements of Machine Learning (Aula 07) 03.febrero.2025

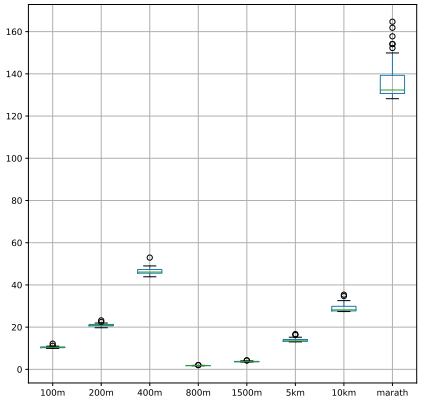
Interpretación de PCA

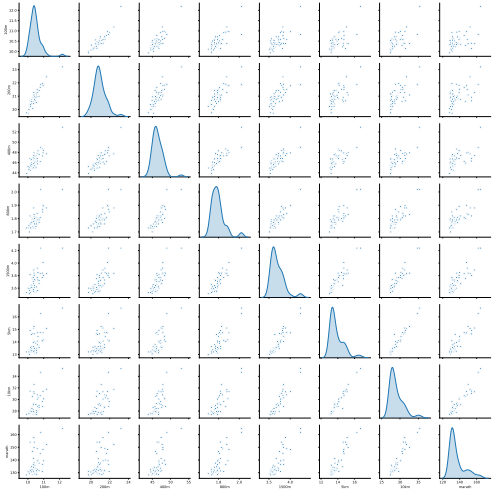
Ejemplo: (Datos de atletismo *deport.csv*) Recordemos los datos de deportes

country	100m	200m	400m	800m	1500m	5km	10km	marath
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

Vimos que los datos relativamente siguen una distribución cercana a una normal multivariada.

Interpretación de PCA





UFM
UNIVERSIDAD
FRANCISCO
MARROQUÍN

Interpretación de PCA

Para ese conjunto de datos, obtuvimos las siguientes componentes principales

C1	C2	C3	C4	C5	C6	C7	C8
0.318	0.567	0.332	0.128	0.263	0.594	0.136	0.106
0.337	0.462	0.361	-0.259	-0.154	-0.656	-0.113	-0.096
0.356	0.248	-0.560	0.652	-0.218	-0.157	-0.003	0.
0.369	0.012	-0.532	-0.480	0.540	0.015	-0.238	-0.038
0.373	-0.140	-0.153	-0.405	-0.488	0.158	0.610	0.139
0.364	-0.312	0.190	0.03	-0.254	0.141	-0.591	0.547
0.367	-0.307	0.182	0.08	-0.133	0.219	-0.177	-0.797
0.342	-0.439	0.263	0.300	0.498	-0.315	0.399	0.158

Interpretación de PCA

Por ejemplo R hace una simplificación, y usualmente descarta aquellos coeficientes cercanos a cero.

C1	C2	C3	C4	C5	C6	C7	C8
0.318	0.567	0.332	0.128	0.263	0.594	0.136	0.106
0.337	0.462	0.361	-0.259	-0.154	-0.656	-0.113	
0.356	0.248	-0.560	0.652	-0.218	-0.157		
0.369		-0.532	-0.480	0.540		-0.238	
0.373	-0.140	-0.153	-0.405	-0.488	0.158	0.610	0.139
0.364	-0.312	0.190		-0.254	0.141	-0.591	0.547
0.367	-0.307	0.182		-0.133	0.219	-0.177	-0.797
0.342	-0.439	0.263	0.300	0.498	-0.315	0.399	0.158

Interpretación de PCA

En ocasiones, es posible hacer una interpretación de los signos en las componentes principales.

Clasificamos los coeficientes de cada componente principal de acuerdo a sus signos: $+$, $-$ ó 0. Por ejemplo, la primer componente principal C_1 es

$$C_1 = (0.318, 0.337, 0.356, 0.369, 0.373, 0.364, 0.367, 0.342)$$

esto significa que

$$C_1 = 0.318x_1 + 0.337x_2 + 0.356x_3 + 0.369x_4 + 0.373x_5 + 0.364x_6 + 0.367x_7 + 0.342x_8,$$

Como C_1 tiene la signatura $(+, +, +, +, +, +, +, +)$, podemos interpretar esta componente como un promedio ponderado de las 8 variables que conforman nuestros datos.

Interpretación de PCA

Esto es: la primer componente C_1 es el tiempo promedio de desempeño en las pruebas de atletismo.

Consideremos ahora la segunda componente principal C_2

$$C_2 = (0.567, 0.462, 0.248, 0.012, -0.14, -0.312, -0.307, -0.439),$$

o bien

$$C_2 = 0.567x_1 + 0.462x_2 + 0.248x_3 + 0.01x_4 - 0.14x_5 - 0.312x_6 - 0.307x_7 - 0.439x_8,$$

con la signatura $(+, +, +, 0, -, -, -, -)$. Esta componente puede interpretarse como un contraste entre las primeras tres variables (100m, 200m, 400m) y las últimas cuatro (1500m, 5km, 10km, maratón).

Interpretación de PCA

Esta componente separa los grupos en función de la diferencia de desempeño entre las pruebas de velocidad y de resistencia.

Finalmente, la tercer componente principal es

$$C_3 = (-0.332, -0.361, 0.56, 0.532, 0.153, -0.19, -0.182, -0.263),$$

con signatura $(-,-,+,+,+,-,-,-)$. Esta componente puede interpretarse como un contraste entre las variables x_3, x_4, x_5 (400m, 800m, 1500m) y el resto de variables.

De alguna manera, esta componente separa los grupos en función de su desempeño en las pruebas de medio fondo.

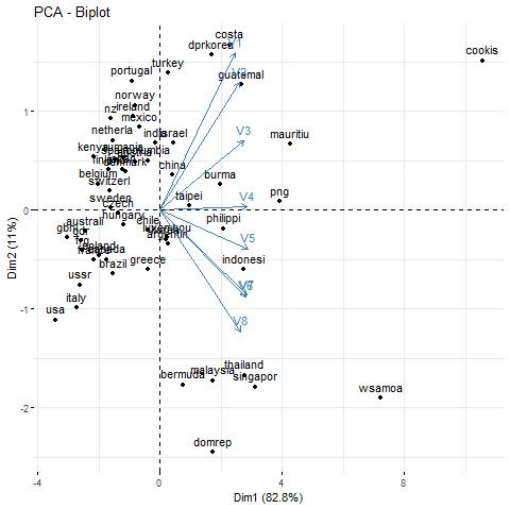
Interpretación de PCA

El resto de componentes se interpretan igual

	C1	C2	C3	C4	C5	C6	C7	C8
V1	0.318	0.567	0.332	0.128	0.263	0.594	0.136	0.106
V2	0.337	0.462	0.361	-0.259	-0.154	-0.656	-0.113	
V3	0.356	0.248	-0.560	0.652	-0.218	-0.157		
V4	0.369		-0.532	-0.480	0.540		-0.238	
V5	0.373	-0.140	-0.153	-0.405	-0.488	0.158	0.610	0.139
V6	0.364	-0.312	0.190		-0.254	0.141	-0.591	0.547
V7	0.367	-0.307	0.182		-0.133	0.219	-0.177	-0.797
V8	0.342	-0.439	0.263	0.300	0.498	-0.315	0.399	0.158

Esta información se resume en el **biplot**.

Interpretación de PCA



UFM
UNIVERSIDAD
FRANCISCO
MARROQUÍN

Errores comunes

0. No centrar los datos

En PCA, es mandatorio centrar los datos. Esto tiene varios motivos:

- Queremos proyectar a un subespacio de menor dimensión (subespacio vectorial, de modo que debe contener al origen).
- La definición de covarianza implica centrar los datos, pues

$$Cov(X) = \mathbb{E}[(\mathbb{X} - \mu)^T(\mathbb{X} - \mu)].$$

- Podríamos pensar en proyectar hacia un subespacio afín (que no pasa por el origen). En ese caso, en lugar de obtener la dirección principal, el PCA va a devolver el vector que apunta al centroide de la distribución.

Errores comunes

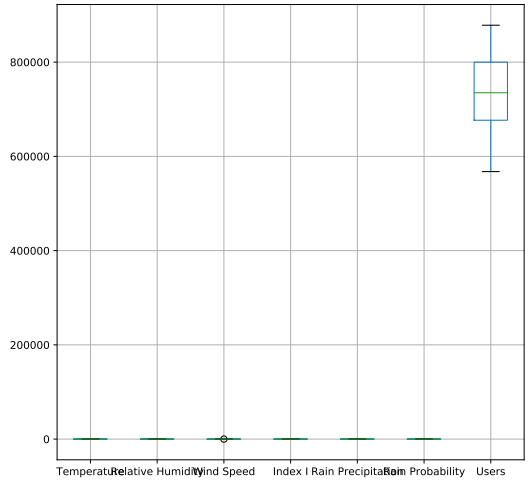
1. Escala de los datos

Algo que hay que tener en cuenta a la hora de hacer PCA es que la escala de los datos puede influenciar o sesgar el resultado.

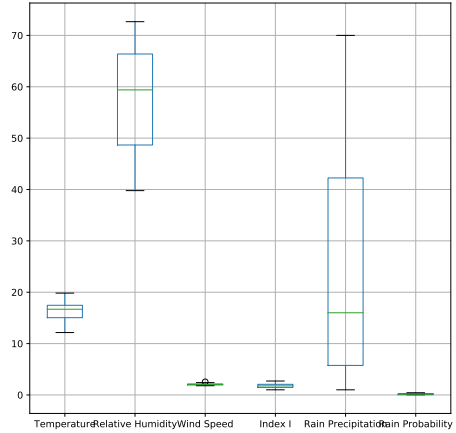
Por ejemplo, en los datos de *meteo-users.csv*

	Temp.	Rel.Humidity	W.Speed	Index.I	Rain.Prec	Rain.Prob	Users
2013-11	14.747	66.516	2.001	1.314	1	0.033	656292
2013-12	14.418	57.069	1.885	1.708	2	0.033	580750
2014-01	12.669	51.606	1.910	1.794	10	0.066	664949
2014-02	16.324	40.144	1.875	1.272	5	0.033	691331
2014-03	17.748	42.042	2.286	1.794	6	0.066	746663
2014-04	18.659	42.776	2.153	2.161	14	0.133	643804

Errores comunes

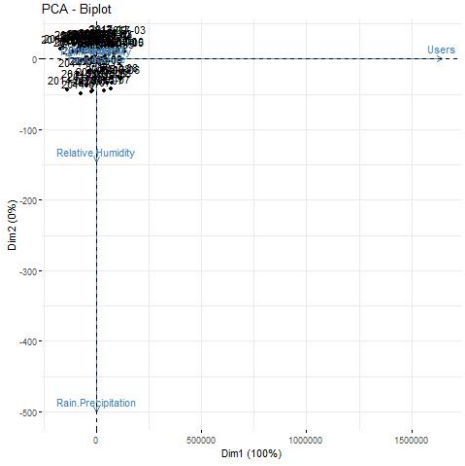


Errores comunes



Errores comunes

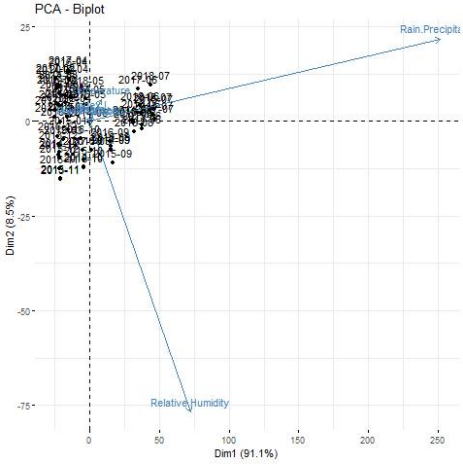
Al hacer PCA de los datos crudos (sin escalar), se obtiene



UFM UNIVERSIDAD FRANCISCO MARROQUÍN

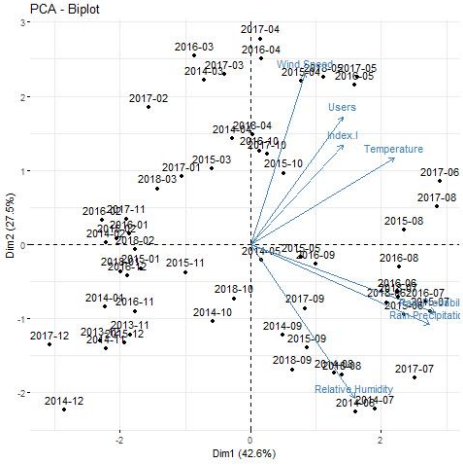
Errores comunes

Ahora, dividiendo “Users” entre 10^6, se obtiene



Errores comunes

Haciendo PCA con los datos estandarizados, obtenemos más coherencia.



UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Errores comunes

3. No normalidad

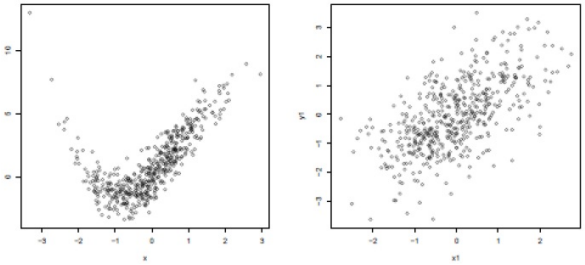
PCA asume normalidad de los datos. Como ya indicamos, PCA basa su análisis en estudiar la covarianza $Cov(X)$, de modo que sólo es capaz de detectar estructuras lineales subyacentes.

Esto no quiere decir que está prohibido usar PCA con datos no normales. Sin embargo, se debe ser cuidadoso en su uso en estos casos: a medida que los datos se alejan de una normal multivariada, el PCA es cada vez menos informativo.

4. Datos no cuantitativos

No tiene sentido aplicar PCA a datos no cuantiativos.

Cuidados en uso de PCA

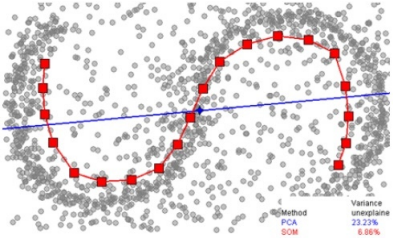


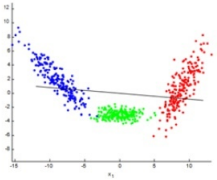
Dos veces misma correlación. =(

Obs.

- Cuidado con desviaciones fuertes de normalidad.
- Lo ideal es investigar la normalidad de los datos en la práctica, al menos ver si escala es continua, distribución unimodal, simétrica, ...

Cuidados en uso de PCA





En el caso de querer detectar correlaciones no lineales, o comportamiento no-lineales, es mejor usar:

- Información mutua
- Coeficiente de correlación rank de Pearson
- PPS (*predictive power score*)