Ciencia de Datos

Victor Muñiz

victor_m@cimat.mx

Asistente:

Víctor Gómez

victor.gomez@cimat.mx

Maestría en Cómputo Estadístico.
Centro de Investigación en Matemáticas.
Unidad Monterrey.

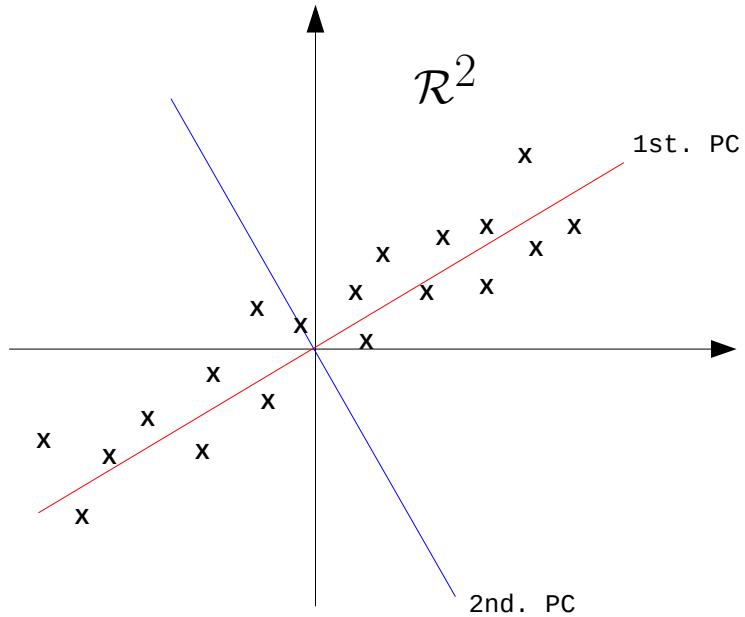
Enero-Junio 2021

Análisis de Componentes Principales (PCA)

Pearson (1901), Hotelling (1933).

PCA

- Quizá el método más conocido y usado para reducir la dimensión de los datos y ayudar a comprender su estructura.
- Se basa en proyectar los datos en "direcciones interesantes". En este caso, estas direcciones están dadas por la estructura de covarianzas de los datos.
- Estas nuevas direcciones definen un nuevo espacio coordenado, que reemplazará a las variables o coordenadas originales.



PCA

- Los componentes principales son combinaciones lineales de las variables originales que **rotan** el sistema de coordenadas original en d dimensiones.
- ¿Cómo encontrar estas direcciones?
- Pearson: como un problema de mínimos cuadrados.
- Hotelling: como un problema de optimización (relacionado a análisis de factores).

PCA

- Los componentes principales son combinaciones lineales de las variables originales que **rotan** el sistema de coordenadas original en d dimensiones.
- ¿Cómo encontrar estas direcciones?
- Pearson: como un problema de mínimos cuadrados.
- Hotelling: como un problema de optimización (relacionado a análisis de factores).

PCA

- Consideraremos $\mathbf{X}_{(n\times d)}$ nuestra matriz de datos con n observaciones en d dimensiones.
- Por simplicidad, supondremos que nuestros datos están centrados por columnas, es decir $X_{i,j} = x_{i,j} - \bar{x}_j$, para $i = 1, \ldots, n$ y $j = 1, \ldots, d$.
- La estimación de la matriz de covarianzas de nuestros datos está dada por

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

- En PCA, queremos encontrar direcciones de proyección que sean solución al siguiente problema de optimización:

$$\max_{\mathbf{u}} \text{Var}(\mathbf{u}^T\mathbf{X}) = \mathbf{u}^T\mathbf{S}\mathbf{u} \quad \text{sujeto a } \|\mathbf{u}\|^2 = 1, \qquad (1)$$

donde la restricción de norma unitaria se impone para evitar que $\mathbf{u}$ crezca en forma arbitraria.

PCA

- Utilizando el método de Lagrange, la solución al problema de optimización se expresa en términos del Lagrangiano:

$$\mathcal{L}(\mathbf{u}, \lambda) = \mathbf{u}^T \mathbf{S} \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1),$$

donde λ es el multiplicador de Lagrange. La solución debe cumplir

$$\nabla \mathcal{L}(\mathbf{u}, \lambda) = \mathbf{S}\mathbf{u} - \lambda \mathbf{u} = 0,$$

entonces

$$\mathbf{S}\mathbf{u} = \lambda \mathbf{u},$$

es decir, la solución corresponde a un vector propio de $\mathbf{S}$.

PCA

- Sustituyendo lo anterior en la función a maximizar, tenemos que

$$\mathbf{u}^T\mathbf{S}\mathbf{u} = \lambda\mathbf{u}^T\mathbf{u} = \lambda,$$

  entonces, para maximizar (1) debemos escoger el valor λ más grande.
- Sean $(\lambda_1, \mathbf{u}_1), (\lambda_2, \mathbf{u}_2), \ldots, (\lambda_t, \mathbf{u}_t)$ los pares eigenvalor-eigenvector de $\mathbf{S}$, con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_t \geq 0$, entonces

el primer componente principal de PCA corresponde a $\mathbf{u}_1$, el primer vector propio de $\mathbf{S}$.

PCA

- La función de proyección de un punto $\mathbf{x}$ en la dirección de $\mathbf{u}$ como

$$P_{\mathbf{u}}(\mathbf{x}) = \langle \mathbf{u}, \mathbf{x} \rangle = \mathbf{u}^T \mathbf{x}, \tag{2}$$

  entonces, **la proyección de los datos en el primer componente principal** estará dada por

$$\mathbf{y}_1 = \mathbf{u}_1^T \mathbf{X}.$$

- Para encontrar el segundo componente principal añadimos la restricción $\mathbf{u}_1^T \mathbf{u}_2 = 0$ para imponer la decorrelación entre las proyecciones $\mathbf{u}_1^T \mathbf{X}$ y $\mathbf{u}_2^T \mathbf{X}$.

PCA

- Como antes, formulamos el Lagrangiano añadiendo la nueva restricción y derivando respecto a $\mathbf{u}_2$ obtenemos

$$\mathbf{S}\mathbf{u}_2 - \lambda_2\mathbf{u}_2 - \pi\mathbf{u}_1 = 0, \tag{3}$$

donde π es el multiplicador de Lagrange relacionado con la nueva restricción.
Multiplicamos por la izquierda por $\mathbf{u}_1^T$, y por la restricción de ortogonalidad entre $\mathbf{u}_1$ y $\mathbf{u}_2$ y la decorrelación de las proyecciones, obtenemos

$$\cancelto{0}{\mathbf{u}_1^T\mathbf{S}\mathbf{u}_2} - \lambda_2\cancelto{0}{\mathbf{u}_1^T\mathbf{u}_2} - \pi\cancelto{1}{\mathbf{u}_1^T\mathbf{u}_1} = 0.$$

Entonces, $\pi = 0$.

PCA

- Sustituyendo este valor en (3), tenemos que

$$\mathbf{S}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2,$$

siendo la solución $\mathbf{u}_2$, el eigenvector relacionado con λ_2, el segundo eigenvalor más grande de $\mathbf{S}$.

- Para obtener todos los componentes principales realizamos el mismo procedimiento. Entonces, la proyección de los datos en el $i-$ésimo componente principal está dado por

$$\mathbf{y}_i = \mathbf{u}_i^T \mathbf{X} \tag{4}$$

con $var(\mathbf{y}_i) = \lambda_i$, donde $(\mathbf{u}_i, \lambda_i)$, es el $i-$ésimo par eigenvector-eigenvalor de la matriz de covarianzas $\mathbf{S}$, con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_t \geq 0$.

PCA

Entonces, el cálculo de los componentes principales, se basa en la descomposición espectral:

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}',$$

con las columnas de $\mathbf{U}$ los eigenvectores normalizados de $\mathbf{S}$, $\mathbf{U}\mathbf{U}' = \mathbf{U}'\mathbf{U} = \mathbf{I}$ y $\text{diag}(\mathbf{\Lambda}) = (\lambda_1, \lambda_2, \cdots, \lambda_d)$, los valores propios ordenados.

PCA

Es evidente que

$$\text{Var. Total} = \text{tr}(\mathbf{S}) = \text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}') = \text{tr}(\mathbf{\Lambda}\mathbf{U}'\mathbf{U}) = \text{tr}(\mathbf{\Lambda}),$$

Por supuesto, usaremos un número "pequeño" de componentes principales. La variación proporcional de cada componente será

$$\frac{\lambda_i}{\text{tr}(\mathbf{\Lambda})},$$

y la **varianza acumulada** (información "retenida") la obtenemos mediante

$$\frac{\sum_{i=1}^{p} \lambda_i}{\text{tr}(\mathbf{\Lambda})},$$

con $p < d$.

PCA

Es evidente que

$$\text{Var. Total} = \text{tr}(\mathbf{S}) = \text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}') = \text{tr}(\mathbf{\Lambda}\mathbf{U}'\mathbf{U}) = \text{tr}(\mathbf{\Lambda}),$$

Por supuesto, usaremos un número "pequeño" de componentes principales. La variación proporcional de cada componente será

$$\frac{\lambda_i}{\text{tr}(\mathbf{\Lambda})},$$

y la **varianza acumulada** (información "retenida") la obtenemos mediante

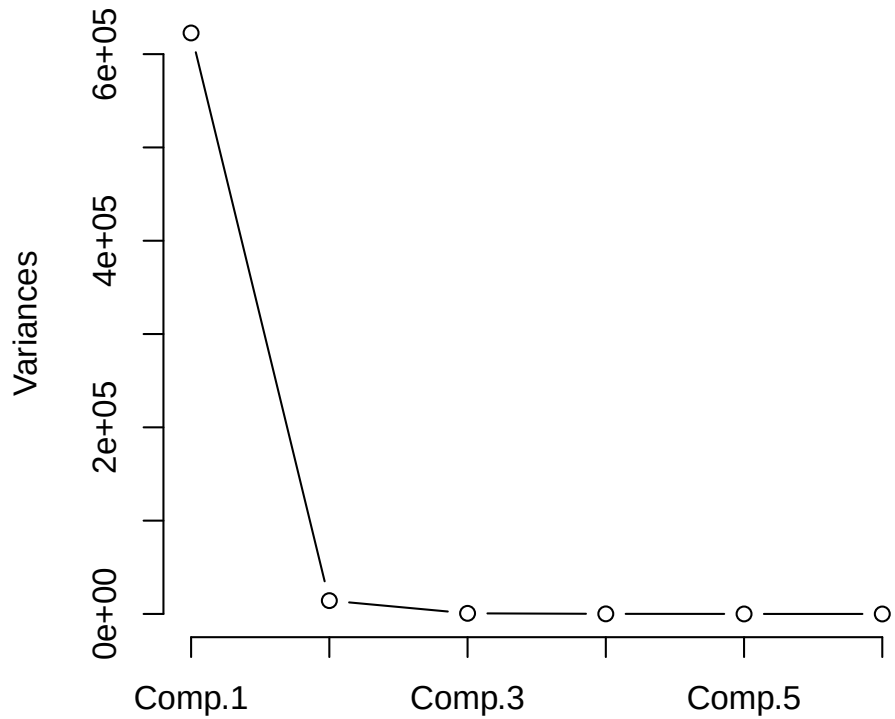
$$\frac{\sum_{i=1}^{p} \lambda_i}{\text{tr}(\mathbf{\Lambda})},$$

con $p < d$.

PCA

Sobre la selección del número de componentes

- Es un aspecto muy importante. El objetivo es reducir la dimensionalidad de los datos pero manteniendo una porción aceptable de variabilidad en $\mathbf{X}$.
- A "ojo". El screeplot:



PCA

Sobre la selección del número de componentes

- El criterio más usado es el porcentaje acumulado de variación total, donde se selecciona un porcentaje de variación que los componentes deben retener. Este porcentaje está dado por

$$\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=n}^{d} \lambda_i} \times 100$$

donde $p < d$ es el número de componentes seleccionados. Generalmente, nos quedamos con los componentes que acumulan al menos 80 % de la varianza.

PCA

Sobre la selección del número de componentes

- Otro criterio es la llamada *regla de Kaiser*, que consiste en retener los componentes principales cuyas varianzas $\lambda_i \geq \bar{\lambda}$, donde $\bar{\lambda}$ es el promedio de las varianzas.
- Generalmente, se utilizan los primeros p componentes principales, sin embargo, es posible que los últimos componentes principales sean posean cierta información que se excluya al usar este criterio.

PCA

¿Covarianza o correlación?

- Es muy fácil ver que todos los resultados anteriores para PCA son válidos usando la matriz de correlación

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2},$$

con $\mathbf{D}^{1/2}$ =diag$(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{dd}})$.

- Usar $\mathbf{R}$ es equivalente a usar datos estandarizados (centrados y con varianza 1 en cada variable).
- ¿Qué implicación tiene en el resultado?

Considera un ejemplo sencillo:

$$\mathbf{S}_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \qquad \mathbf{S}_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

PCA

¿Covarianza o correlación?

- Es muy fácil ver que todos los resultados anteriores para PCA son válidos usando la matriz de correlación

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2},$$

con $\mathbf{D}^{1/2}$ =diag$(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{dd}})$.

- Usar $\mathbf{R}$ es equivalente a usar datos estandarizados (centrados y con varianza 1 en cada variable).
- ¿Qué implicación tiene en el resultado?

Considera un ejemplo sencillo:

$$\mathbf{S}_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \qquad \mathbf{S}_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

PCA

¿Covarianza o correlación?

- Es muy fácil ver que todos los resultados anteriores para PCA son válidos usando la matriz de correlación

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2},$$

con $\mathbf{D}^{1/2}$ =diag$(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{dd}})$.

- Usar $\mathbf{R}$ es equivalente a usar datos estandarizados (centrados y con varianza 1 en cada variable).
- ¿Qué implicación tiene en el resultado?

Considera un ejemplo sencillo:

$$\mathbf{S}_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \qquad \mathbf{S}_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

PCA

¿Covarianza o correlación?

- Es muy fácil ver que todos los resultados anteriores para PCA son válidos usando la matriz de correlación

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2},$$

con $\mathbf{D}^{1/2}$ =diag$(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{dd}})$.

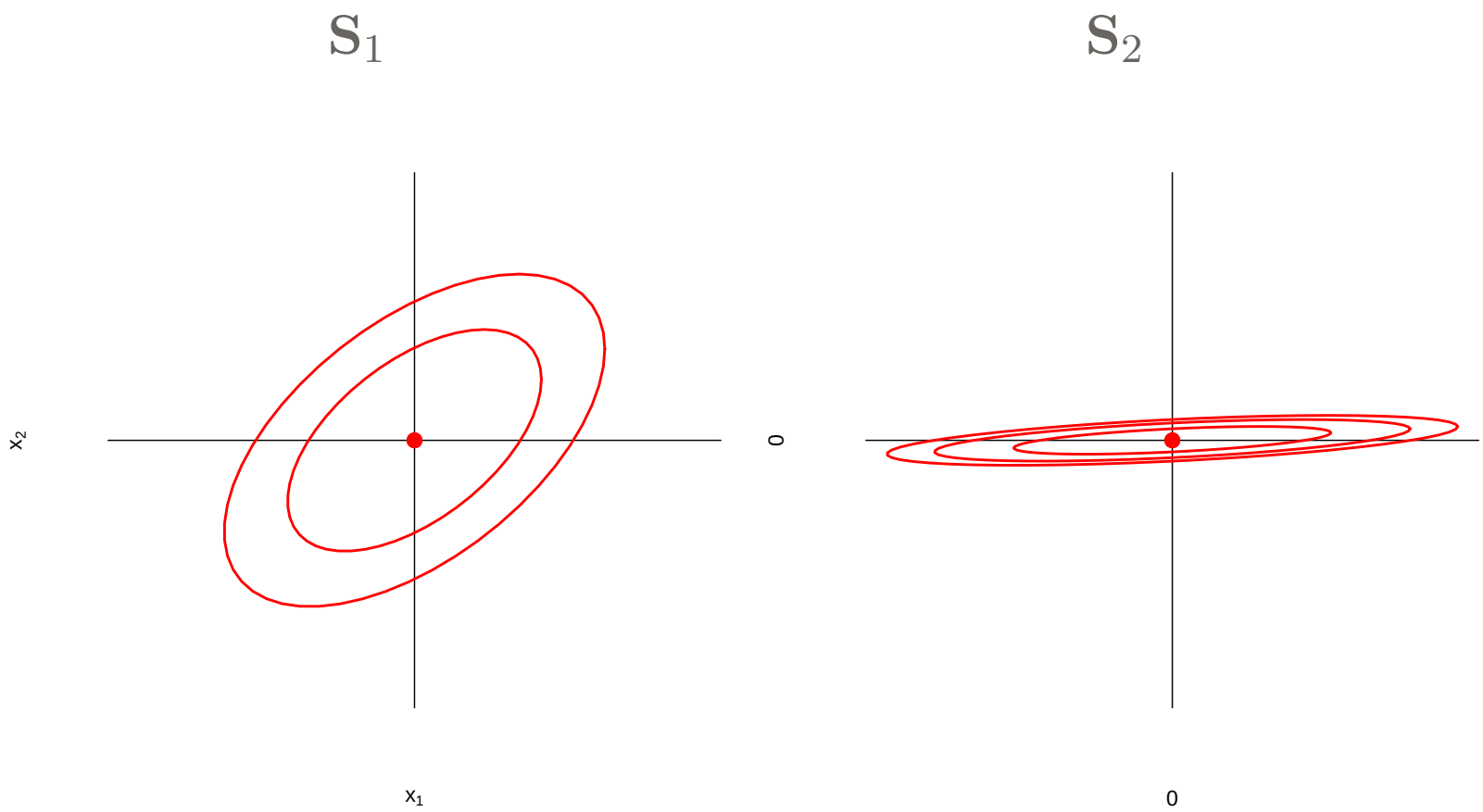
- Usar $\mathbf{R}$ es equivalente a usar datos estandarizados (centrados y con varianza 1 en cada variable).
- ¿Qué implicación tiene en el resultado?

Considera un ejemplo sencillo:

$$\mathbf{S}_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \qquad \mathbf{S}_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

PCA

¿Covarianza o correlación?
Las elipses de distancias (probabilidades) constantes son:



Interpreta...

PCA

¿Covarianza o correlación?

- El porcentaje de varianza obtenida por los componentes de $\mathbf{R}$ y $\mathbf{S}$ difieren
- Los coeficientes de los componentes principales (y en consecuencia la proyección de los datos en ellos) varían entre $\mathbf{R}$ y $\mathbf{S}$
- Si los datos tienen varianzas muy diferentes, es conveniente estandarizar, de lo contrario, las variables con varianzas más grandes, dominarán los coeficientes de los componentes principales
- Si los datos se miden en escalas diferentes, o tienen rangos de valores muy diferentes, es conveniente estandarizar

PCA como un modelo interpretativo

PCA

Muchas veces, recopilamos datos que son (o esperamos que sean) representativos de algún fenómeno que nos interesa analizar.
Mas que el simple hecho de *simplificación computacional* del problema, a nosotros nos interesaría inferir ciertos comportamientos, comprobar supuestos y en general, sacar conclusiones. Esta es la función mas tradicional de PCA.

En este caso, podemos ver PCA como un método para

- Extraer información importante de nuestros datos
- Simplificar la información
- Simplificar la descripción de los datos
- Analizar la estructura y relación entre las observaciones y las variables.

PCA

Ejemplo: US air pollution data.

`notebooks/3-visualizacion.ipynb`

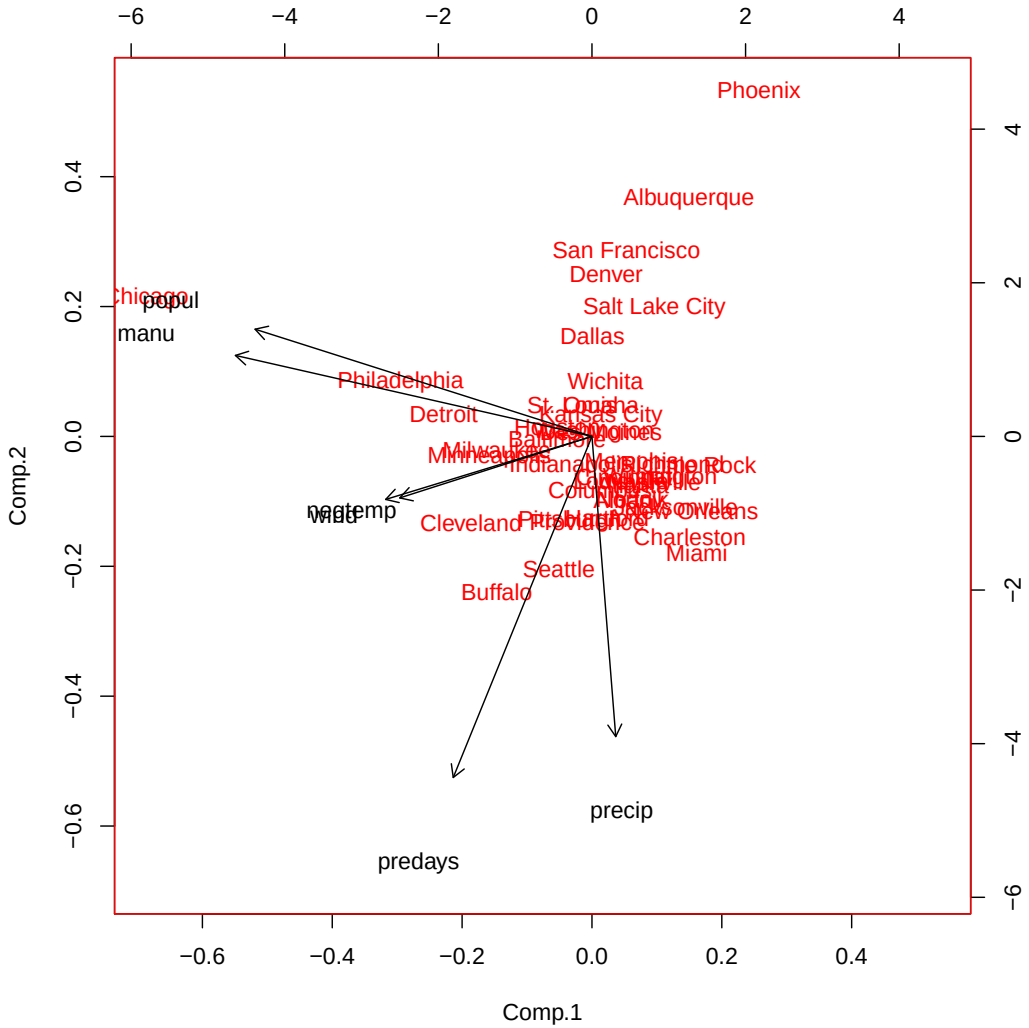
PCA

El biplot.

- Es una gráfica de los datos en una matriz de $n \times d$.
- Nos permite visualizar las distancias generalizadas entre los objetos, además de las varianzas y covarianzas entre las variables de los datos. Para más detalles respecto a la construcción de este gráfico, puede verse: Gabriel, K. R. (1971), *The biplot graphic display of matrices with application to principal component analysis*, Biometrika, 58, 453-467.
- En la gráfica pueden verse la representación espacial de los datos (en 2 dimensiones).
- La longitud del vector que parte del inicio y se dirige hacia la coordenada de alguna variable representa la **varianza** de dicha variable
- El ángulo entre dos vectores refleja la **correlación** entre las variables correspondientes.

PCA

Para los datos de contaminación:



PCA

Nos concentramos principalmente en analizar los primeros componentes principales, pero, ¿Qué pasa con los últimos?

- Si la varianza de un componente es prácticamente cero, tal componente representa una combinación lineal de las variables "constante".
- Sugiere presencia de colinearidad o dependencia entre variables.

PCA

Nos concentramos principalmente en analizar los primeros componentes principales, pero, ¿Qué pasa con los últimos?

- Si la varianza de un componente es prácticamente cero, tal componente representa una combinación lineal de las variables "constante".
- Sugiere presencia de colinearidad o dependencia entre variables.

PCA

- Por ejemplo, supón $\mathbf{x} \in \mathbb{R}^5$, donde $x_5 = \sum_{i=1}^{4} x_i/4$.
- ¿Cuál será el valor de las proyecciones en el PC 5?
- ¿Cómo son los loadings para PC 5?
- Puede mostrarse que, en este caso, debe ser proporcional a $(1, 1, 1, 1, -4)$.

`notebooks/3-visualizacion.ipynb`

- Ver Libin Yang, William Rea and Alethea Rea *"Financial Insights from the Last Few Components of a Stock Market PCA"*.

PCA

- Por ejemplo, supón $\mathbf{x} \in \mathbb{R}^5$, donde $x_5 = \sum_{i=1}^{4} x_i/4$.
- ¿Cuál será el valor de las proyecciones en el PC 5?
- ¿Cómo son los loadings para PC 5?
- Puede mostrarse que, en este caso, debe ser proporcional a $(1, 1, 1, 1, -4)$.

`notebooks/3-visualizacion.ipynb`

- Ver Libin Yang, William Rea and Alethea Rea *"Financial Insights from the Last Few Components of a Stock Market PCA"*.

PCA

- Por ejemplo, supón $\mathbf{x} \in \mathbb{R}^5$, donde $x_5 = \sum_{i=1}^{4} x_i/4$.
- ¿Cuál será el valor de las proyecciones en el PC 5?
- ¿Cómo son los loadings para PC 5?
- Puede mostrarse que, en este caso, debe ser proporcional a $(1, 1, 1, 1, -4)$.

`notebooks/3-visualizacion.ipynb`

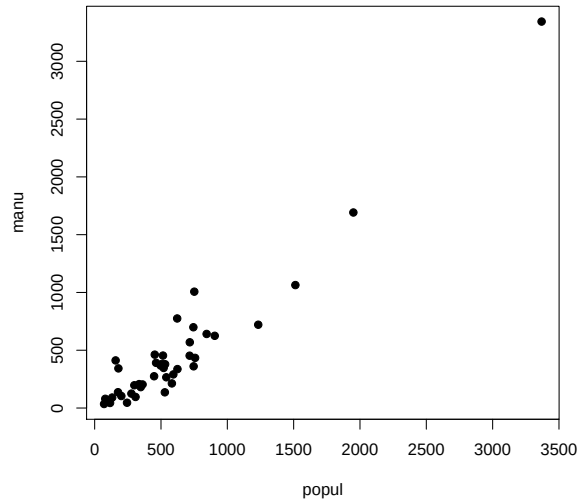
- Ver Libin Yang, William Rea and Alethea Rea *"Financial Insights from the Last Few Components of a Stock Market PCA"*.

PCA como un modelo predictivo

PCA como un modelo predictivo

¿Cómo usar PCA para predecir valores de SO2?

- Una opción es, por supuesto, usar regresión lineal multivariada, pero hay un problema con estos datos:



Hay una alta correlación entre estas dos variables.

- En general, hay muchos casos en que aparece este fenómeno de **colinearidad** en las covariables, y muchas veces no es posible (o no es recomendable) eliminar alguna de ellas para hacer regresión lineal.

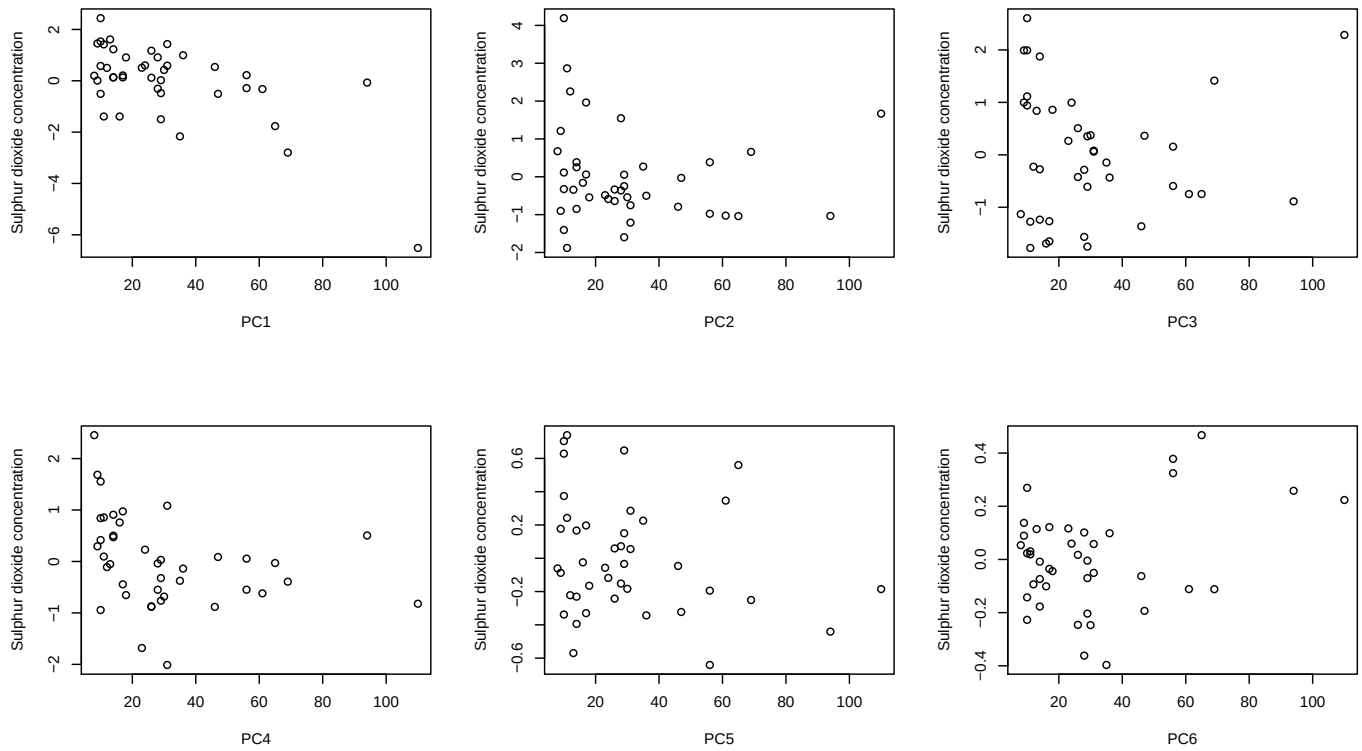
PCA como un modelo predictivo

Posibles soluciones.

- Métodos de "encogimiento" (shrinkage methods): elimina o reduce el efecto de variables "redundantes" mediante regularización. Ejemplos: LASSO (least absolute shrinkage and selection operator), Ridge Regression (RR).
- Métodos de regresión con componentes ortogonales. Ejemplos: QR regression, Principal Component Regression (PCR), Partial Least Squares Regression (PLS). PCR y PLS utilizan un número pequeño de combinaciones lineales $\mathbf{z}_j$, $j = 1, \ldots p$ de las variables originales, que son usadas como las variables para realizar la regresión.

PCA como un modelo predictivo

- Podemos hacer la regresión en los componentes principales definidos por las variables restantes. Recuerda que los PC están decorrelacionados.



PCA como un modelo predictivo

PCR: Principal Components Regression

- Sea $\mathbf{z}_j = \mathbf{X}\mathbf{u}_j$ un componente principal obtenido como vimos anteriormente, mediante la descomposición espectral de la matriz de covarianzas o de correlación (o equivalentemente, a través de la descomposición SVD de $\mathbf{X}$).
- En PCR, se realiza la regresión de $\mathbf{y}$ en $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p$ para algún $p \leq d$, obteniendo así una secuencia de modelos de regresión $\hat{\mathbf{y}}_0 \ldots \hat{\mathbf{y}}_p$. Como las $\mathbf{z}$'s son ortogonales, la regresión es la suma de regresiones univariadas:

$$\hat{\mathbf{y}} = \bar{y}\mathbf{1} + \sum_{j=1}^{p} \hat{\beta}_j \mathbf{z}_j$$

donde $\hat{\beta}_j = \langle \mathbf{z}_j, \mathbf{y} \rangle / \langle \mathbf{z}_j, \mathbf{z}_j \rangle$.

PCA como un modelo predictivo

PLS: Partial Least Squares. (Frank and Friedman. *A Statistical View of Some Chemometrics Regression Tools*. 1993)

- Al igual que PCR, construye un conjunto de combinaciones lineales de las variables de entrada, pero además de usar $\mathbf{X}$, usa también $\mathbf{y}$ para construirlas.
- Algoritmo PLS (solo como referencia):
  - Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$ and $\mathbf{x}_j^{(0)} = \mathbf{x}_j$, $j = 1, \dots d$.
  - for $m = 1, 2 \dots, d$
    - $\mathbf{z}_m = \sum_{j=1}^{d} \hat{\varphi}_{m,j} \mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{m,j} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.
    - $\hat{\beta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$
    - $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\beta}_m \mathbf{z}_m$
    - Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$
  - Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^d$.

PCA como un modelo predictivo

- PLS y PCR tratan de obtener los coeficientes de regresión lejos de las soluciones dadas por mínimos cuadrados ordinarios (OLS) hacia direcciones de mayor dispersión en el espacio de los predictores.
- Análisis: construir un subespacio $p-$dimensional del espacio $d-$dimensional original para realizar la regresión, con la restricción:

$$\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j \mathbf{u}_j$$

con $\mathbf{u}_j$, $j = 1, \ldots, p$ expandiendo el subsubespacio y $||\mathbf{u}_j||^2 = 1$.

PCA como un modelo predictivo

- Para OLS:

$$\begin{aligned} \mathbf{u}_{OLS} &= \max_{\mathbf{u}} \operatorname{corr}^2(\mathbf{y}, \mathbf{X}\mathbf{u}) \\ \text{s. a.} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- Para PCR:

$$\begin{aligned} \mathbf{u}_p(PCR) &= \max_{\mathbf{u}} \operatorname{var}(\mathbf{X}\mathbf{u}) \\ \text{s. a.} \quad & \|\mathbf{u}\|^2 = 1 \\ & \mathbf{u}_j^T \mathbf{S}\mathbf{u} = 0,\ j = 1, \ldots, p-1 \end{aligned}$$

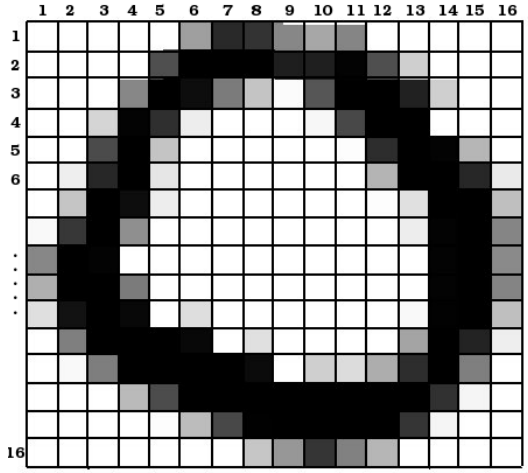
- Para PLS:

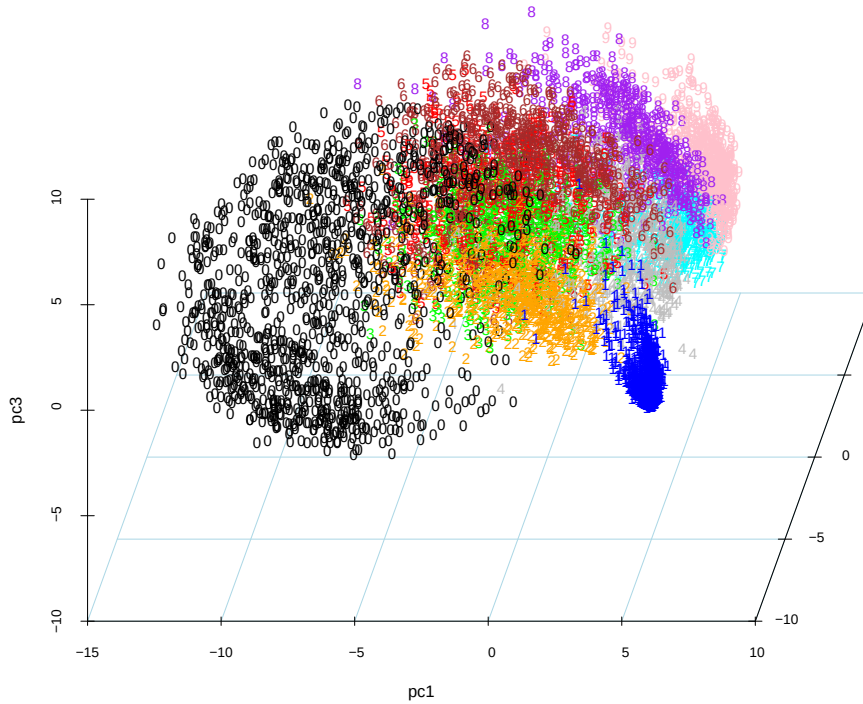
$$\begin{aligned} \mathbf{u}_p(PLS) &= \max_{\mathbf{u}} \operatorname{corr}^2(\mathbf{y}, \mathbf{X}\mathbf{u}) \operatorname{var}(\mathbf{X}\mathbf{u}) \\ \text{s. a.} \quad & \|\mathbf{u}\|^2 = 1 \\ & \mathbf{u}_j^T \mathbf{S}\mathbf{u} = 0,\ j = 1, \ldots, p-1 \end{aligned}$$

PCA como método de reducción de dimensión

PCA como método de reducción de dimensión

Ejemplo: Dígitos escritos a mano y escaneados.





`notebooks/4-visualizacion.ipynb`

PCA

Ejemplo: Eigenfaces.

Objetivo: reconocer (clasificar) un rostro según una base de datos existente.

`notebooks/4-visualizacion.ipynb`