UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Variables latentes: ICA

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 13) 18.febrero.2021

Variables latentes

Idea: En estadística, las variables latentes (o variables ocultas, en contraposición a las variables observables), son las variables que no se observan directamente sino que son inferidas (a través de un modelo matemático) a partir de otras variables que se observan.

De alguna forma, al hacer PCA, estamos construyendo variables latentes (la direcciones principales), como combinaciones de las variables observables (las X_i).

Existe varias familias de modelos matemáticos que consisten de construir variables latentes.
En particular, hay toda una familia de métodos basados en descomposiciones matriciales.

Whitening

Definición

*Sea $X \in \mathbb{R}^d$ un vector aleatorio con una matriz de covarianza no singular Σ y media $\mu = \mathbf{0}$. La transformación $Y = WX$, se llama una **matriz de blanqueamiento** (whitening matrix) W si satisface la condición $W^T W = \Sigma^{-1}$. Al vector aleatorio Y se le llama el vector **blanqueado** de X.*

¿Cuál es la media de Y? $\mathbb{E}(Y) = \mathbb{E}(WX) = W\mathbb{E}(X) = \mathbf{0}$.

¿Cuál es la covarianza de Y? $Cov(Y) = Cov(WX) = WCov(X)W^T = W\Sigma W^T = I$.

Existen muchas matrices de blanqueamiento. Las más comunes

- $W = \Sigma^{-1/2}$ (Mahalanobis whitening o ZCA),
- $W = Cholesky(\Sigma^{-1})$, (Cholesky whitening),
- $\Sigma^{-1} = USU^T \Rightarrow W = US^{-1/2}U^T$ (PCA whitening),

PCA

Recordemos el caso de PCA.

Si $\mathbb{X} = USV^T \in \mathbb{R}^{n \times d}$, es la matriz de datos, recordemos que las direcciones principales vienen dadas por

$$T = \mathbb{X}V,$$

donde las columnas de $V \in \mathbb{R}^{d \times d}$ son los autovectores de $\mathbb{X}^T\mathbb{X}$. A la matriz V^T usualmente se le llama la matriz de blanquemiento.

En este caso, también podemos escribir

$$T = \mathbb{X}V = (USV^T)V = US(V^TV) = US,$$

y la proyección a dimensión $1 \leq r \leq d$ se obtiene como $T_r = \mathbb{X}V_r = U_rS_r$.

ICA

El análisis de componente independientes es una técnica estadística para separa un vector o señal multivariada en sus subcomponentes aditivos.

A diferencia del PCA, ICA asume que los subcomponnentes no son gaussianos, pero que son estadísticamente independientes uno del otro.

Se utiliza mucho en procesamiento de señales, para separar las componentes que conforman una señal.

ICA

Dado un vector aleatorio $X = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ formado por d variables aleatorias simples, el ICA realiza una transformación lineal del tipo:

$$S = WX,$$

donde los componentes de S son máximamente independientes entre sí, según algún criterio dado por una función $F(S)$.

Las componentes de X (observables) se construyen como combinaciones llineales de las componentes de S (variables latentes):

$$X = AS, \quad x_i = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{id}s_d,$$

ó

$$X = AS + \varepsilon, \quad x_i = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{id}s_d + \varepsilon_i,$$

con $\varepsilon \sim \mathcal{N}_d(0, \sigma^2 I_d)$.

ICA

Lo que se desea en este caso, es que las variables latentes $S = (s_1, s_2, \ldots, s_d)$, seán lo mas independientes unas de otras.

Para ello, utilizamos algún criterio funcional $F(S) = F(s_1, s_2, \ldots, s_d)$ que mida la independencia estadística de las s_i.

Existen muchos criterios para medir esta independencia:

- negentropía,
- información mutua,
- divergencia de Kullback-Leibler,
- kurtosis,
- ...

ICA

En teoría de la información y estadística, la negentropía se utiliza como una medida de la distancia a la normalidad.

Recordemos que de todas las distribuciones con una media μ y varianza σ^2 dadas, la distribución normal es la que tiene la mayor entropía.

La negentropía mide la diferencia de entropía entre una distribución dada y la distribución gaussiana con parámetros μ y σ^2.

Definición

*La **negentropía** se define como*

$$J(f_X) = S(\varphi_X) - S(f_X),$$

donde $S(\varphi_X)$ es la entropía diferencial de la densidad gaussiana, y $S(f_X)$ es la entropía diferencial de la distribución $f_X S$.

Recordemos que

$$S(f_X) = -\int_{\mathbb{R}^d} f_X(\mathbf{u}) \log(f_X(\mathbf{u}))\, d\mathbf{u}.$$

$$\Rightarrow J(f_X) = \int_{\mathbb{R}^d} \Big(f_X(\mathbf{u}) \log(f_X(\mathbf{u}) - \varphi_X(\mathbf{u}) \log(\varphi_X(\mathbf{u})) \Big)\, d\mathbf{u}.$$

Fast ICA

En la práctica usamos otras versiones de ICA. La más popular se llama *Fast ICA* (Hyvärinen, 2000).

La técnica requiere hacer *whitening* de los datos. Si $\mathbb{X} \in \mathbb{R}^{n \times d}$ es la matriz de datos, hacemos

$$\mathbb{X} = \mathbb{X} - \mu,$$

y luego una descomposición en valores singulares de la matriz de covarianza

$$\mathbb{X}\mathbb{X}^T = USV^T.$$

Luego, consideramos la tranformación $W = S^{-1/2}V^T$, y hacemos

$$\mathbb{S} = W\mathbb{X} = S^{-1/2}V^T\mathbb{X}.$$

Fast ICA

Observe que la covarianza de los nuevos datos es

$$Cov(\mathbb{S}) = \mathbb{S}\mathbb{S}^T = W\mathbb{X}\mathbb{X}^T W^T = S^{-1/2}V^T VSV^T VS^{-1/2} = S^{-1/2}SS^{-1/2} = I.$$

Extracción de una sola componente:

El algoritmo iterativo encuentra la dirección del vector de peso $\mathbf{w} \in \mathbb{R}^n$ que maximiza la medida de no-gaussianidad de la proyección $\mathbf{w}^T\mathbb{X}$.
Para medir la no gaussianidad, FastICA usa una función no lineal no cuadrática $f(u)$, y sus derivadas $f'(u)$, $f''(u)$.
Hyvärinen propone

$$f(u) = \log\cosh(u), \quad f'(u) = \tanh(u), \quad f''(u) = 1 - \tanh^2(u),$$

son útiles para fines generales, mientras que las siguientes son robustas

$$f(u) = -e^{-u^2/2}, \quad f'(u) = ue^{-u^2/2}, \quad f''(u) = (1-u^2)e^{-u^2/2}.$$

Fast ICA

Algoritmo (*Fast ICA* single extraction):

1. Elegir un vector aleatorio $\mathbf{w}_0 \in \mathbb{R}^d$.
2. Repetir $i = 0, 1, 2, 3, \ldots$ hasta cumplir un criterio de paro:
   - Hacer $\mathbf{w}_i^+ = \mathbb{E}\big(\mathbb{X} f'(\mathbf{w}_i^T \mathbb{X})^T\big) - \mathbb{E}\big(f''(\mathbf{w}_i^T \mathbb{X})\big)\mathbf{w}_i$
     Aquí $\mathbb{E}$ significa centrar o restar la media por columnas.
   - Hacer $w_{i+1} = \dfrac{\mathbf{w}_i^+}{||\mathbf{w}_i^+||}$.

Fast ICA

Extracción de múltiples componentes:

El algoritmo iterativo anterior estima sólo un vector de peso que extrae un único componente. La estimación de componentes adicionales que son mutuamente “independientes”.

Esto requiere la repetición del algoritmo para obtener vectores de proyección linealmente independientes. Observe que la noción de independencia aquí se refiere a maximizar la no gaussianidad en los componentes estimados.

Hyvärinen proporciona varias formas de extraer múltiples componentes, siendo la más simple la siguiente.

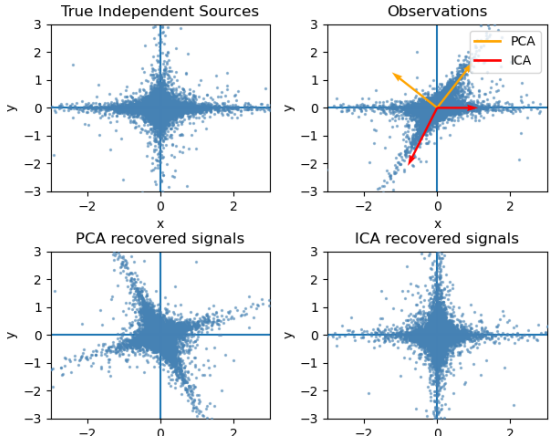
Fast ICA

Algoritmo (*Fast ICA* multiple extraction):
Inputs: $\mathbb{X}\mathbb{R}^{n\times d} =$ matriz de datos blanqueada,
Inputs: $c =$ número de componentes a extraer, $c << n$,
Outputs: $W \in \mathbb{R}^{c\times n} =$ matriz de proyección,
Outputs: $\mathbb{S} \in \mathbb{R}^{c\times d} =$ matriz de componentes independientes,
Para $k = 1, 2, \ldots, c$:

1. Elegir un vector aleatorio $\mathbf{w}_k \in \mathbb{R}^n$.
2. Repetir $i = 0, 1, 2, 3, \ldots$ hasta cumplir un criterio de paro:
   - Hacer $\mathbf{w}_k = \mathbb{E}\left(\mathbb{X}f'(\mathbf{w}_k^T\mathbb{X})^T\right) - \mathbb{E}\left(f''(\mathbf{w}_k^T\mathbb{X})\right)\mathbf{w}_k$
   - Hacer $\mathbf{w}_k = \mathbf{w}_k - \sum_{j=1}^{k-1}(\mathbf{w}_k^T\mathbf{w}_j)\,\mathbf{w}_j$
   - Hacer $w_k = \dfrac{\mathbf{w}_k}{||\mathbf{w}_k||}$.
3. Return $W = [\mathbf{w}_1, \ldots, \mathbf{w}_c]$, $\mathbb{S} = W\mathbb{X}$.

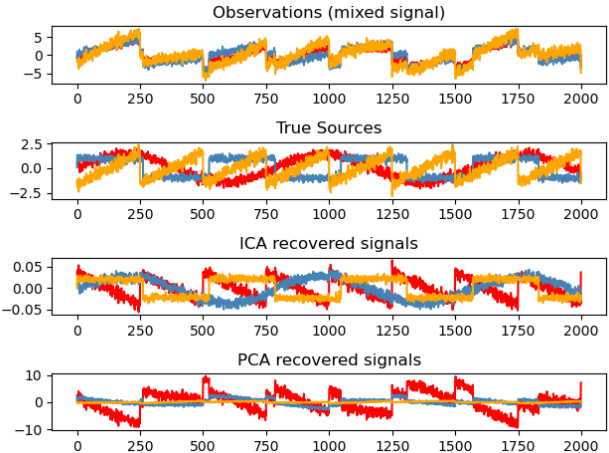
Ejemplos

1. Nube de puntos en 2D



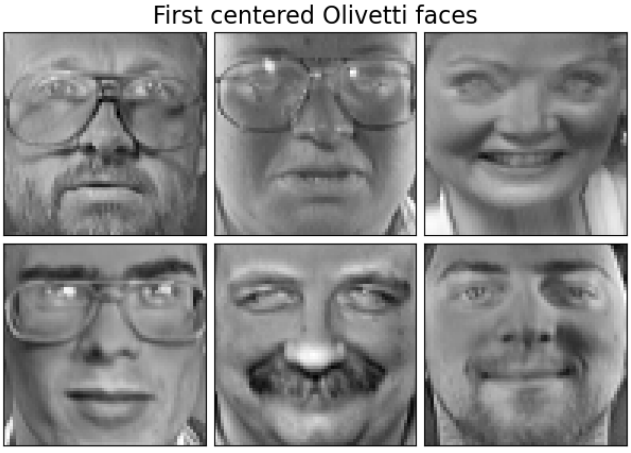
Ejemplos

2. Descomposición de una señal de audio



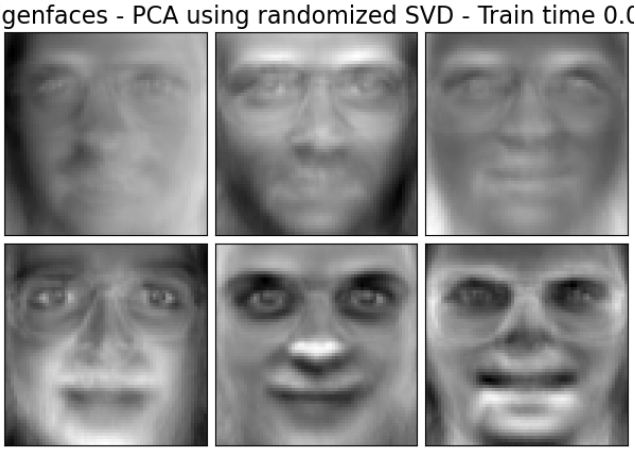
Ejemplos

3. Descomposición *eigenfaces*



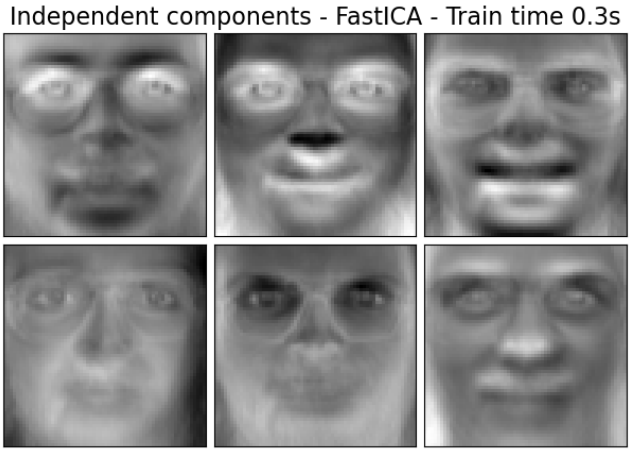
Ejemplos

3. Descomposición *eigenfaces*



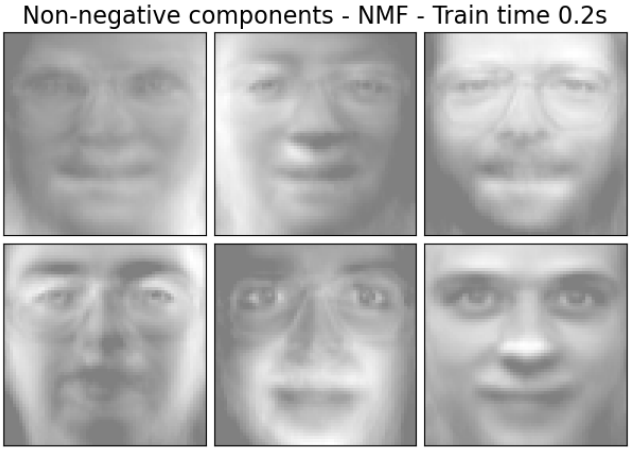
Ejemplos

3. Descomposición *eigenfaces*



Ejemplos

3. Descomposición *eigenfaces*



NNMF

La Factoración de Matrices No-Negativas (NNMF) es otra estrategia para obtener variables lantentes.

Sea $\mathbb{X} \in \mathbb{R}^{n \times d}$ una matriz de datos. La idea es descomponer $\mathbb{X}$ como el producto de dos matrices con entradas no-negativas $W \in \mathbb{R}^{n \times r}$ y $H \in \mathbb{R}^{r \times d}$

$$\mathbb{X} = WH,$$