

Módulo 5: Aprendizaje no supervisado

Diplomatura Cs. de Datos - FaCENA-UNNE

Docentes: Magdalena Lucini, Luis Duarte, Griselda Bóbeda

Aprendizaje no supervisado

¿Qué se entiende por Aprendizaje No Supervisado (ANS) ?

- Tipo de aprendizaje automático en el que los modelos o técnicas:
 - ▶ aprenden a partir de un conjunto de datos SIN etiquetar
 - ▶ actúan sobre esos datos SIN supervisión
- Se dispone de los datos de entrada, pero no los de salida ⇒ encontrar estructuras ocultas en los datos, agruparlos según semejanzas, devolver una representación “útil” de ese conjunto de datos.

Introducción ANS

Objetivo: Analizar un conjunto de datos en crudo para convertirlo en información relevante. (estudiar la estructura de los datos)

- Se interpretan estos datos para encontrar relaciones o **patrones** ocultos (diferencias y semejanzas, reglas, clases) .
- los datos no necesariamente deben ser crudos (puede hacerse un EDA antes)
- ¿Qué nos dicen los resultados? Hay que interpretarlos, analizarlos, consultar con un experto de dominio.

Algunas reglas/métodos de ANS

Reducción de dimensionalidad

Objetivos:

- Reducir la dimensionalidad del conjunto de datos cuando el número de características (variables) es elevado, o son redundantes, preservando la mayor cantidad de información posible.
- proyectar los datos a otro espacio (quizás de menor dimensión) para poder separarlos

Veremos:

- Análisis de Componentes Principales (ACP, PCA)
- Análisis de Componentes Independientes (ICA)
- Embeddings: tSNE (t-distributed Stochastic Neighbour Embedding), UMAP

Algunas reglas/métodos de ANS

Agrupamiento (clustering)

Objetivo: encontrar grupos naturales en los datos sin que haya información previa de cómo agruparlos (qué elementos son similares entre sí)

Ejemplo: Agrupar células de acuerdo a la información genética que ellas poseen, etc.

Veremos

- K-means
- Mezcla de Gaussianas
- DBscan
- HDBscan

Algunas reglas/métodos de ANS

Estimación de densidades

Objetivo: dado un conjunto de datos, $\{\mathbf{X}\}$, encontrar la distribución de probabilidades que los genera: $p(x)$, $p(z/x)$, $p(x/z)$, donde Z es una variable latente que influye en la variable observada X

Veremos

- Técnicas de muestreo
- Naive Bayes

Ejemplo 1: Indicadores Sociodemográficos

- Datos
 - ▶ $n = 26$ individuos (países)
 - ▶ $p = 10$ variables (indicadores demográficos)
- cf. Population Reference Bureau (<http://www.prb.org/>)

País	Tasa de Nacimientos(%)	Tasa Mortalidad(%)	...	Población
Afganistán	47	21		6384000
Albania	13	21		1443000
Argentina	19	8		36324000
⋮				
Zimbabwe	31	21		5024000

Table: Ejemplo 1 de datos multivariados

Ejemplo 2: Indicadores corporales

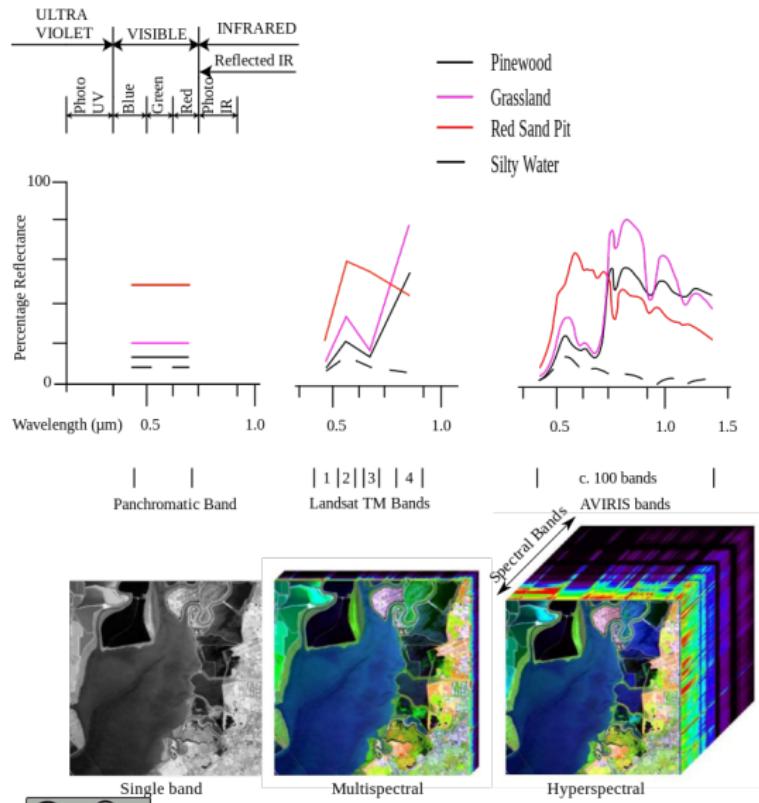
- Datos

- ▶ $n = 507$ individuos (personas)
- ▶ $p = 24$ variables (indicadores del cuerpo)

Persona	Prof. Pecho (cm)	Largo Pierna(cm)	...	Edad	Peso (kg)
1	17.7	106.2		21	65.6
2	16.9	110.5		23	71.8
:					
506	15.5	107.1		33	66.4
507	20.4	100.5		38	57.3

Table: Ejemplo 2 de datos multivariados

Ejemplo 3: Imágenes teledetección



This work is licensed under a Creative Commons Attribution 3.0 Unported License.
Author: <http://commons.wikimedia.org/wiki/User:Arbeck>

Preguntas

- Extraer y sintetizar variables pertinentes:
 - ▶ ej 1 → ¿hay indicadores similares?
 - ▶ ej 2 → ¿hay indicadores de cuerpos parecidos?
 - ▶ ej 3 → ¿se podrá reducir la dimensionalidad de los datos, preservando información?
- Formar grupos de individuos con mismas características:
 - ▶ ej 1 → ¿hay países que se comporten de manera similar?
 - ▶ ej 2 → ¿hay personas con características corporales similares?
 - ▶ ej 3 → ¿hay pixeles que tienen mismas características espectrales?
- Modelar una variable en función de otras:
 - ▶ ej 1 → ¿puede explicarse la tasa de mortalidad en función de las otras variables medidas?
 - ▶ ej 2 → ¿podemos explicar el peso de una persona?

Notación : datos $\{X\}$

individuos	Variable X_1	...	Variable X_j	...	Variable X_p
x_1				\vdots	
\vdots				\vdots	
x_i		$x_{i,j}$	
\vdots					
x_n					

Table: Representación esquemática de una tabla de datos multivariados

- n : número de individuos/objetos/elementos/observaciones x_1, \dots, x_n
- p : número de variables X_1, \dots, X_p
- $x_{i,j}$: respuesta de un individuo/objeto/elemento i a la variable j

Reducción de dimensionalidad

Objetivo

Reducir la dimensionalidad de los datos: describir un conjunto de datos con un número menor de variables preservando la mayor cantidad de información posible

Métodos:

- Reducir características (variables): por “intuición”, eliminar las que tienen poca varianza, eliminar características redundantes.
- **Análisis de Componentes Principales (ACP)** **variables cuantitativas**
- Análisis de Correspondencias **variables cualitativas**
- **Embeddings**

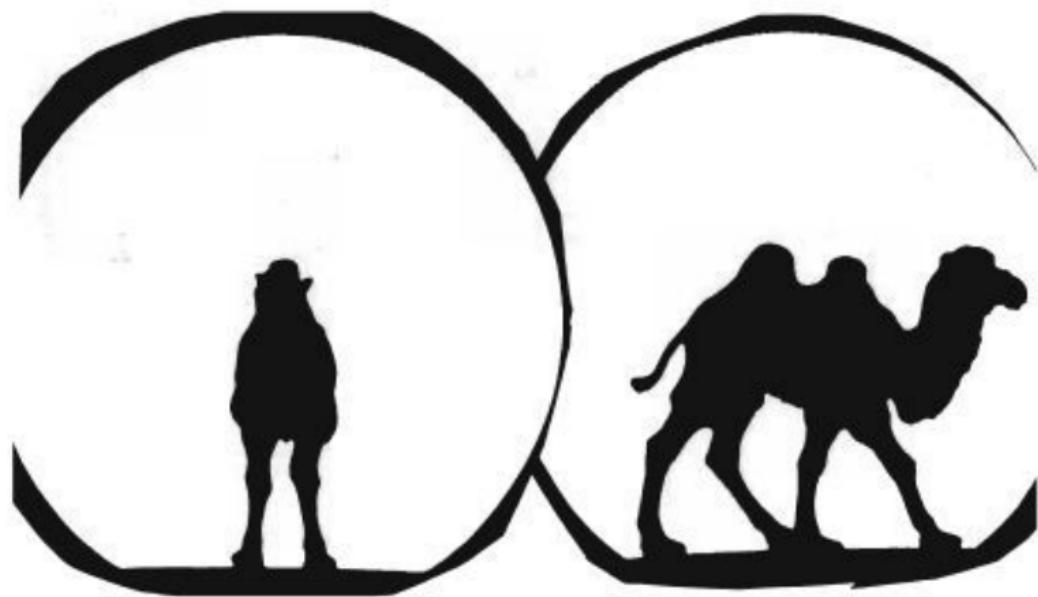
Aplicaciones

compresión de datos, reconstrucción de datos, preprocesamiento de datos (antes de agrupamiento), etc..

Análisis de Componentes Principales - ACP

Objetivo

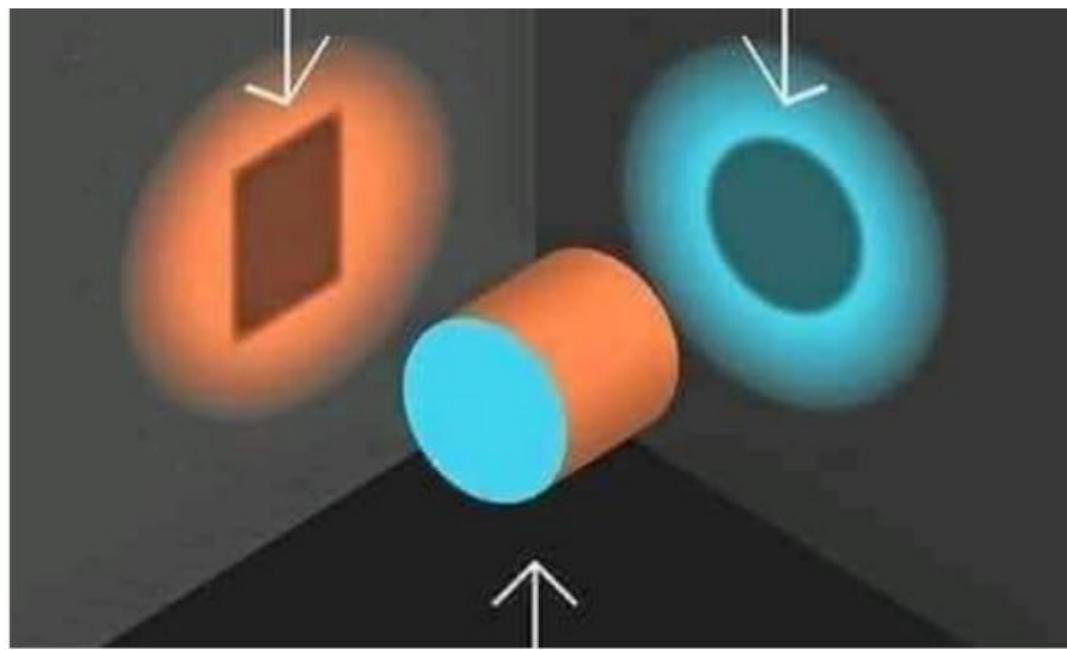
Encontrar el subespacio que **mejor** describa los datos: El más “cercano” por proyecciones.



Análisis de Componentes Principales - ACP

Objetivo

Encontrar el subespacio que mejor describa los datos, teniendo en cuenta el propósito del estudio



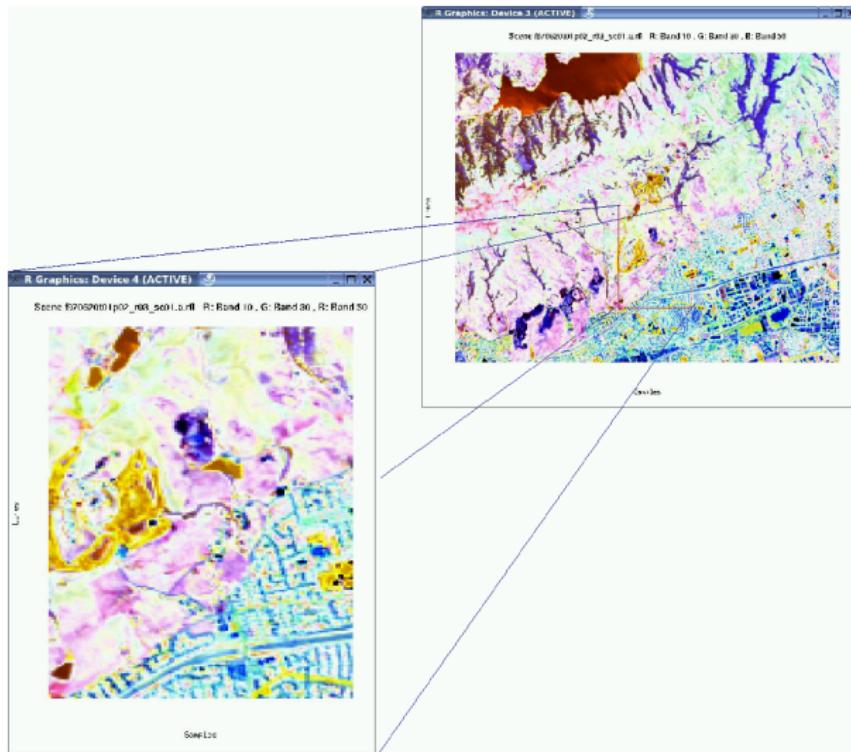
Análisis de Componentes Principales - ACP

Datos

Datos satelitales provistos por el sensor **AVIRIS** (Airborne Visible/Infrared Imaging Spectrometer)

- Identificación, medición y monitoreo de constituyentes de la superficie y la atmósfera terrestres basado en la absorción molecular y firma espectral de las partículas.
- 224 bands (0.4 - 2.5 μm)
- Ancho de cada banda aprox. $0.1\mu\text{m}$

Componentes Principales



Componentes Principales

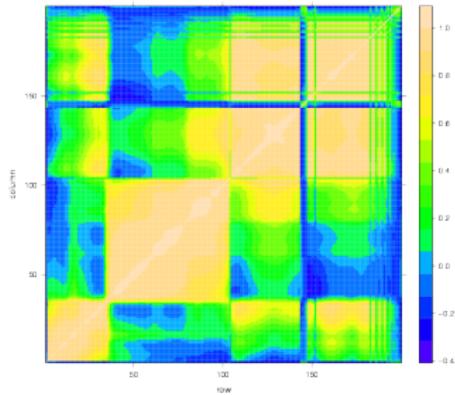
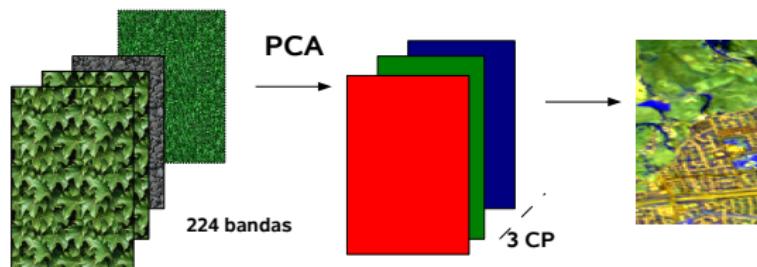


Figure: Matriz de Correlación de la imagen en estudio.

Dificultades:

- Excesiva correlación entre bandas (características) "vecinas"
- Gran volúmen de datos involucrados

Ejemplo-Datos AVIRIS



¿Cúantas “bandas” elegir para poder reconstruir la imagen original con poco error?

¿Qué bandas conviene elegir en función de lo que quiera analizar? (experto de dominio)

Análisis de Componentes Principales

Problema

Encontrar un espacio de dimensión más reducida que represente adecuadamente los datos y brinde la mejor representación de la variabilidad y diversidad de los mismos.

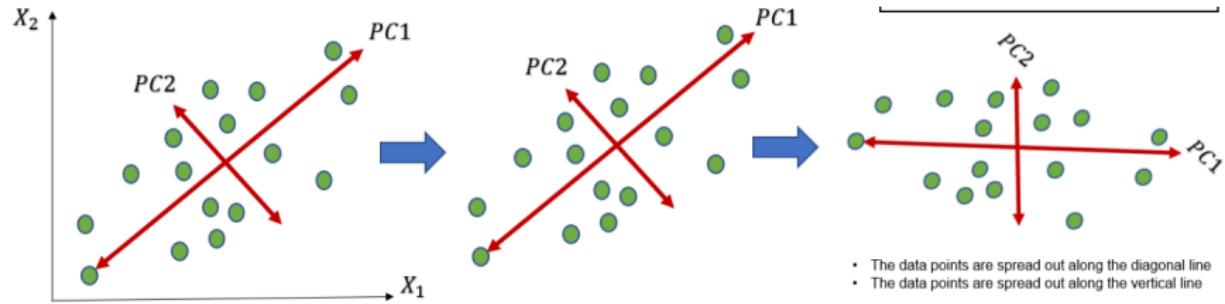
Objetivos

- Encontrar una proyección lineal de los datos de manera tal que la varianza en el espacio proyectado sea máxima.
- Reducir dimensionalidad describiendo las p variables de una matriz X por un subconjunto (pequeño) $r < p$ de combinaciones lineales de las variables originales.
- Describir patrones de correlación entre las variables involucradas.

PCA - Clásico

- Herramienta exploratoria: técnica basada en una muestra para facilitar descripción de los datos.
- Aplicaciones:
 - ▶ Descripción e interpretación de un conjunto de datos.
 - ▶ Utilizada como técnica de pre-procesamiento en diversas aplicaciones(agrupamiento, regresión, etc)
 - ▶ Utilizada en distintas disciplinas (economía, meteorología , procesamiento de imágenes de teledetección, psicología, etc).

Enfoque Geométrico



- The data points are spread out along the diagonal line
- The data points are spread out along the vertical line

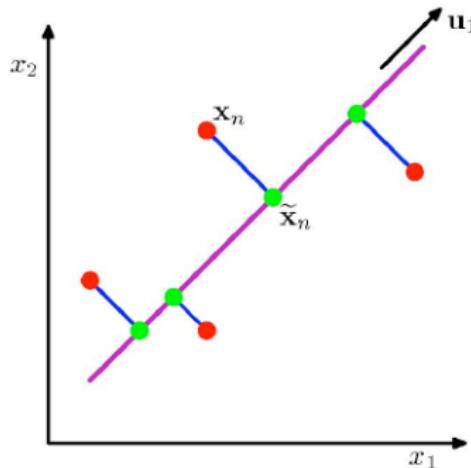
Enfoque Geométrico

- Si las variables X_i están correlacionadas entonces, en general, la nube de puntos forma un elipsoide con centro en \bar{x} cuyos ejes principales no son paralelos a los ejes cartesianos.
- La dirección del eje mayor del elipsoide y la proyección de los puntos sobre esta permiten describir la orientación de la nube de puntos. Este eje **minimiza** las distancias ortogonales de las observaciones a una recta que pase entre ellas, a la vez que **maximiza** la varianza de estas proyecciones.
- Encontrar los ejes del elipsoide es equivalente a encontrar la matriz ortogonal U que rota los ejes de manera tal que los alinea con los ejes del elipsoide.

PCA clásico

Datos : $X \in \mathbb{R}^{n \times p}$, observaciones $\{x_1, x_2, \dots, x_n\}$, $x_j \in \mathbb{R}^p$, o variables $\{X_1, X_2, \dots, X_p\}$, $X_j \in \mathbb{R}^n$

Objetivo: Proyectar los datos en un espacio r-dimensional ($r < p$) maximizando la varianza de los datos proyectados en este nuevo espacio.



Enfoques:

- **Maximizar Varianza:** dispersión de los puntos verdes
- **Minimizar error:** minimizar distancias entre puntos rojos y verdes.

Pasos de un PCA clásico

- Seleccionar las variables (descartar categóricas, etc.)
- Decidir si se van a estandarizar las variables o no. Si las variables tienen distintas unidades o magnitudes muy disímiles deben estandarizarse.
- Determinar el número de componentes que se desean retener.
- Si es necesario rotar componentes para mejorar interpretabilidad
- Interpretar resultados.

PCA clásico

- Primer componente principal es la dimensión en la cual las variables están más dispersas (varianza máxima).
- Segunda componente principal combinación lineal con máxima varianza con dirección ortogonal a la primer componente.
- ...
- Estas nuevas variables (PC) son no correlacionadas.

Si $X = [X_1, \dots, X_p]$, $n \times p$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \in \mathbb{R}^p \text{ vector de medias}$$

$$S = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^t \text{ es la matriz de covarianza de } X$$

PCA- Enfoque algebraico

- Encontrar un subespacio de dimensión $r < p$ tal que la proyección de los puntos sobre el mismo preserve la estructura (posiciones relativas) con la menor distorsión posible.
- Proyectar datos sobre la primer dimensión por un (vector latente) $u_1 \in \mathbb{R}^p$.
- Como sólo nos interesa la dirección de esta nueva dimensión la normalizamos de manera tal que $\|u_1\|^2 = u_1^t u_1 = 1$
- Se busca una combinación lineal $Z_1 = u_{11}X_1 + u_{12}X_2 + \dots + u_{1p}X_p$ de las variables originales que tenga varianza máxima.
Esta será la primer componente principal
- Los valores de la primer componente en los n individuos se representan por el vector

$$z_1 = [u_1^t(x_1 - \bar{x}), \dots, u_1^t(x_n - \bar{x})] \in \mathbb{R}^n$$

PCA- Enfoque algebraico

Haciendo algunas cuentas algebraicas puede verse que :

$$\text{var}(Z_1) = \dots = u_1^t S u_1$$

Para maximizar esa varianza, pidiendo además que $u_1' u_1 = 1$, se debe resolver:

$$S u_1 = \lambda_1 u_1$$

Luego u_1 y λ_1 son un autovector de S y su autovalor correspondiente.

$Z_1 = X u_1$ es una **componente principal**

Los elementos de u_1 son las **cargas (loadings)** de cada una de las variables originales sobre la CP.

Además $\lambda_1 = \text{var}(Z_1)$

Maximizar la varianza implica buscar el autovector con el mayor autovalor.

Pasos PCA

- Para calcular la segunda componente se busca una dirección que maximiza la varianza entre todas las direcciones ortogonales a la primera.
- Resto de las componentes se obtienen calculando los autovectores y autovalores de S .
- Si se quiere obtener un espacio de dimension $r < p$, entonces se repite el procedimiento hasta obtener r vectores ortogonales y se forma $U_r = [u_1, \dots, u_r] \in \mathbb{R}^{p \times r}$
- Los u_i son ortonormales, es decir: $u_i^t u_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

PCA clásico

- Se ordenan los autovalores de mayor a menor, $\lambda_1 \geq \dots \geq \lambda_r$, con correspondientes autovectores u_1, \dots, u_r
- La j -ésima Componente Principal es $Z_j = X_c u_j$, $j = 1, \dots, p$, donde X_c son los datos centrados en torno a la media.
- La proyección de los datos sobre el espacio generado por las primeras r componentes es

$$Z = U_r^t X_c \in \mathbb{R}^{n \times r}$$

- En algunos casos es conveniente usar la matriz de correlación R en lugar de S : si las varianzas difieren substancialmente o las unidades de medición son incommensurables las componentes de S serán dominadas por las variables con mayor varianza.

PCA: resumen (receta)

Dado un conjunto de datos $X \in \mathbb{R}^{n \times p}$,

- Calcular la media $\bar{x} \in \mathbb{R}^p$ y la matriz de covarianza S
- Encontrar autovalores λ_i y autovectores u_i de S
- Las CP son $Z_j = X_c u_j$, $j = 1, \dots, p$
- $Z_j = u_{1j}X_1 + \dots + u_{pj}X_p$ es **combinación lineal** de las p variables originales X_1, \dots, X_p .
Los u_{ij} son las **cargas** (loadings) de la variable X_i sobre la componente principal Z_j .
- La proyección de X sobre el nuevo espacio generado por los r primeros autovectores es

$$Z = U_r^t(X_c)$$

, con X_c datos centrados.

Propiedades

- $\sum_{i=1}^p \text{var}(Z_i) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \text{var}(X_i)$
- Proporción de variabilidad explicada por la componente k-ésima es $\frac{\lambda_k}{\sum_{i=1}^p \lambda_i}$
- La proporción de variabilidad explicada por las j primeras componentes es $\frac{\sum_{i=1}^j \lambda_i}{\sum_{i=1}^p \lambda_i}$
- $\text{cov}(Z_i, X_j) = \lambda_i u_{ij},$

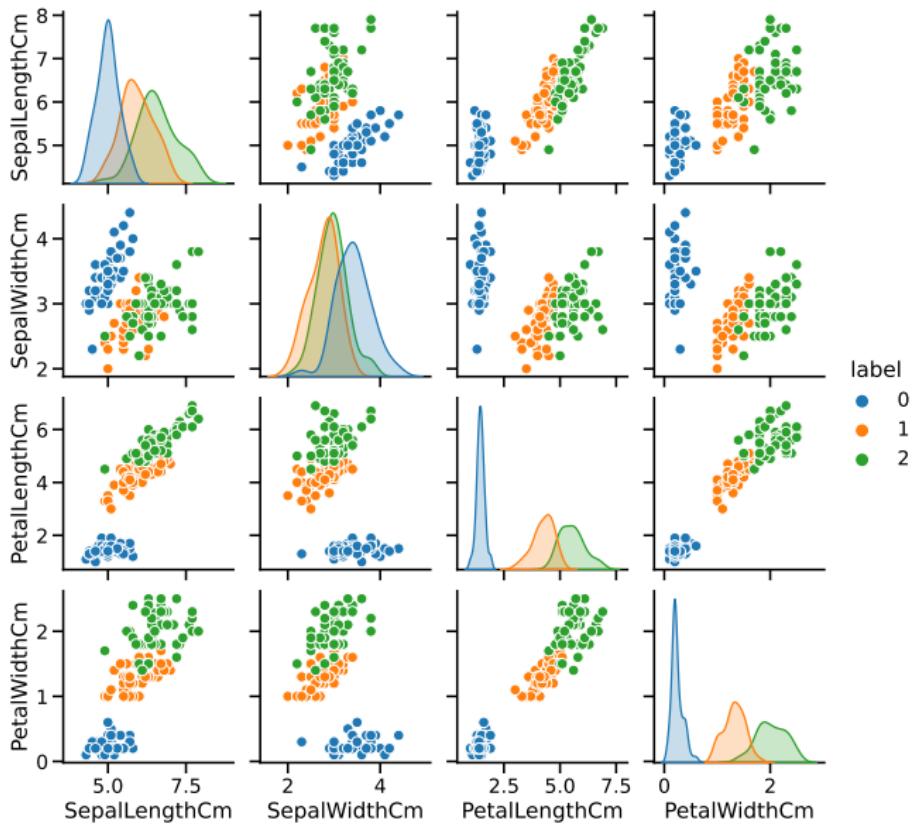
Caso 1

- Datos: Usaremos el conjunto *iris*
- $p = 5$ Variables: largo y ancho de sépalo (*Sepal.Length*, *Sepal.Width*), largo y ancho de pétalo (*Petal.Length*, *Petal.Width*) para flores de tres especies de *iris* (*Species*): *setosa*, *versicolor* y *virginica*.
- $n = 150$ individuos (50 por cada especie)

Estadística Descriptiva

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
:	:	:	:	:	:
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

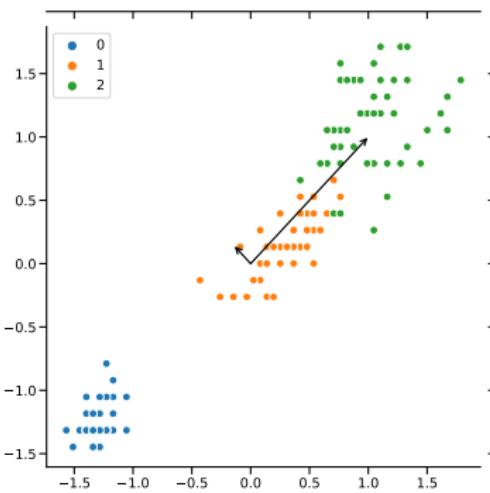
Ejemplo iris:



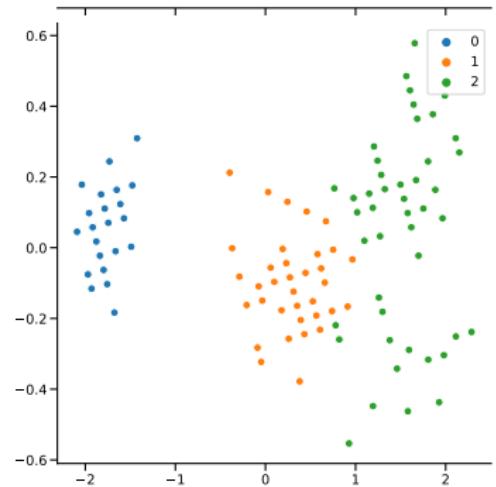
Ejemplo iris

Supongamos que el dataset consiste de solo dos variables, PetalLengths y PetalWidth

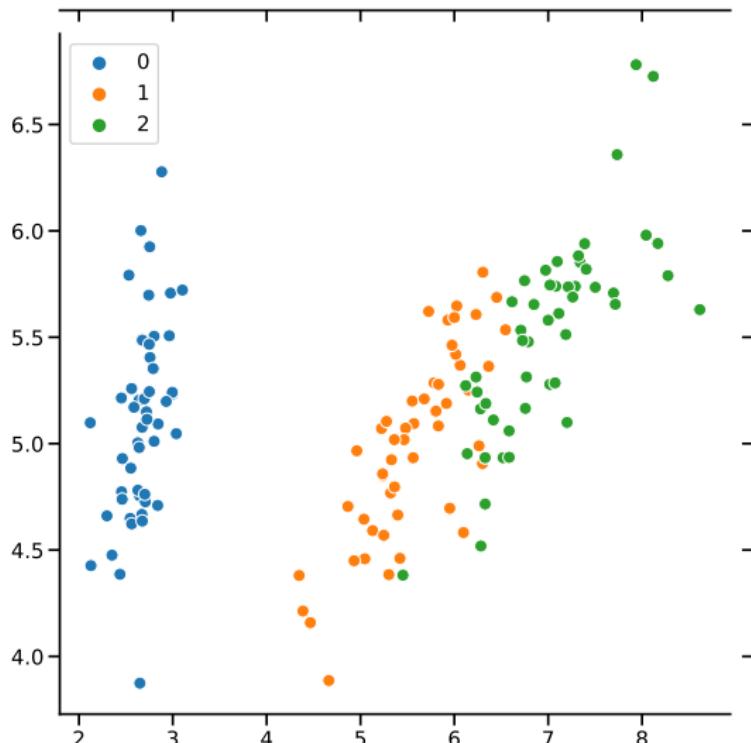
Datos originales



Proyección en Componentes principales



Ejemplo iris: ACP dataset completo



ACP: ¿Cuántas componentes seleccionar?

- Graficar λ_i vs i y buscar el corte (codo) entre autovalores “grandes” y “pequeños” (**scree plot**)
- Seleccionar las componentes necesarias hasta lograr una proporción determinada de la varianza (80%, 90%).
- Seleccionar las componentes cuyos autovalores sean mayores que el promedio de los mismos
$$\sum_{i=1}^p \lambda_i / p.$$

En el ejemplo,

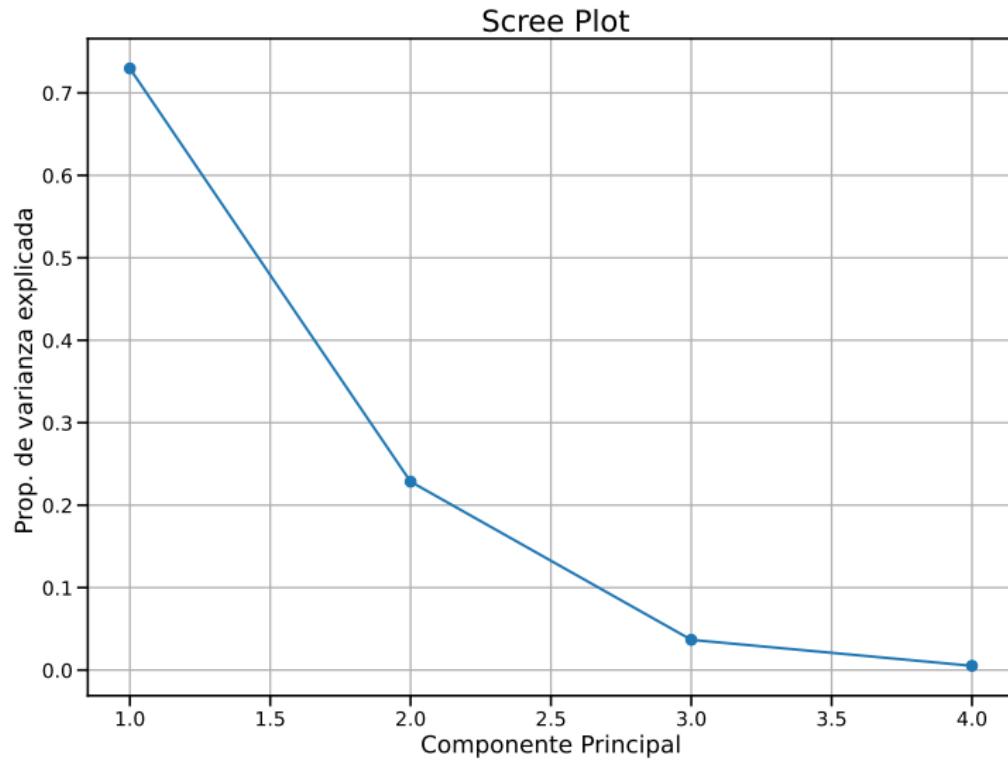
Componentes Principales

```
[[ 0.52106591 -0.26934744  0.5804131  0.56485654]
 [ 0.37741762  0.92329566  0.02449161  0.06694199]
 [-0.71956635  0.24438178  0.14212637  0.63427274]
 [-0.26128628  0.12350962  0.80144925 -0.52359713]]
```

Prop. de varianza explicada

```
[0.72962445 0.22850762 0.03668922 0.00517871]
```

Screeplot



Ejemplo iris: cargas de variables en nuevas componentes

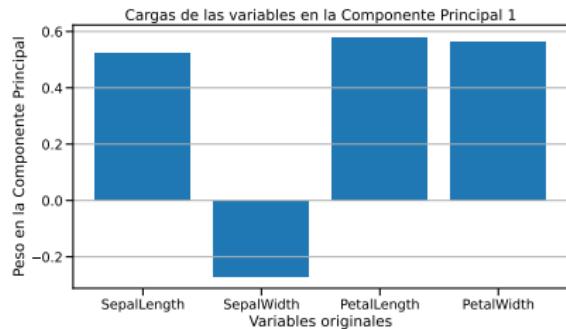


Figure: Cargas de cuatro variables originales en primer componente

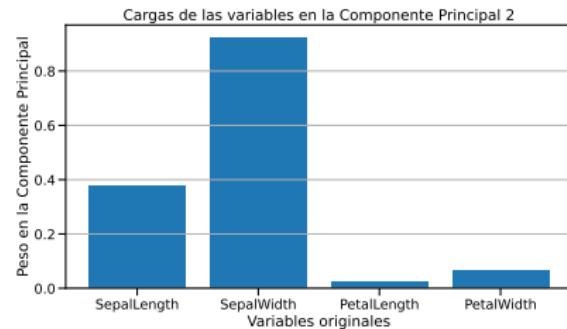


Figure: Cargas de cuatro variables originales en segunda componente

El Análisis de Componentes Principales (ACP) permite:

- **Analizar los individuos:** Hacer particiones entre individuos al detectar similaridades (distancia euclídea) entre ellos respecto a algunas variables o combinaciones de las mismas
- **Analizar las variables:** Se encuentran relaciones lineales entre las variables por medio de la descomposición de la matriz de correlación R (o bien S).
- Pueden describirse grupos de individuos por las variables

PCA vía error de reconstrucción mínimo

Supongamos hemos encontrado r PC que generan un subespacio de dimensión $r < p$

La proyección de la observación x_j en ese subespacio de dimensión r es

$$z_j = U_r^t(x_j - \bar{x}) \in \mathbb{R}^r$$

Luego, una **reconstrucción** de x_j será:

$$\tilde{x}_j \approx U_r z_j + \bar{x} \in \mathbb{R}^p$$

PCA vía error de reconstrucción mínimo

Objetivo:

Encontrar una transformación de los datos originales que minimize el error de reconstrucción:

$$\frac{1}{n} \sum_{i=1}^n \|x_i - \tilde{x}_i\|$$

donde \tilde{x}_i es la reconstrucción de la observación x_i usando solamente r CP,
 $\|\cdot\|$ es la norma euclídea.

PCA vía error de reconstrucción mínimo

A partir del planteo anterior, proponiendo una nueva base ortonormal, y luego de operaciones matriciales, se llega a que el problema a resolver es:

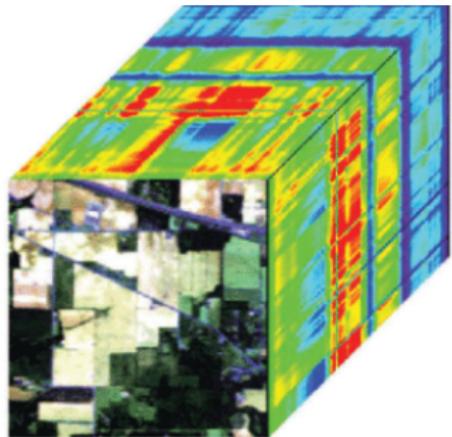
Encontrar los r autovalores más grandes (y los correspondientes autovectores) de la matriz de covarianza S

Es decir,

PCA por mínimo error de reconstrucción = PCA por maximización varianza

Caso 2: Datos Aviris

- Datos: Imagen Indian Pine (Aviris), obtenida de
<https://www.kaggle.com/datasets/pines-hyperspectral-dataset>
- Imagen hyperespectral (+200 bandas) con 145x145 pixeles, de una region de Indiana, EEUU.
- ¿Cómo se forma la matriz de datos?
¿Qué es observación y qué es variable?



Caso 2: Datos Aviris

- **Objetivo 1:** reducción de dimensionalidad: seleccionar 3 “bandas” para realizar una composición a color, maximizando información.
 - ▶ ¿son suficientes esas 3 bandas?
 - ▶ ¿cuánto explican de la varianza de los datos?
- **Objetivo 2:** reducción de dimensionalidad (compresión de datos), reconstrucción de la imagen:
 - ▶ ¿cuántas CP se necesitan para una buena reconstrucción?
 - ▶ Comparar dimensiones

Objetivo 1: reducción de dimensionalidad para visualización



Figure: Banda 30, escala de grises



Figure: Composición a color: r 30, g 100, b 199

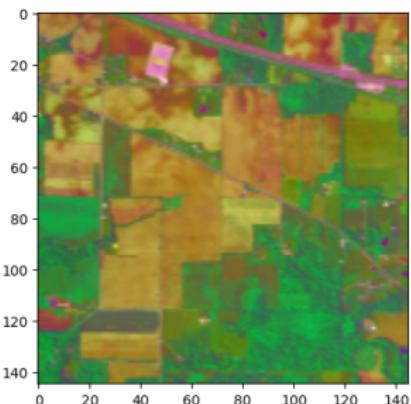
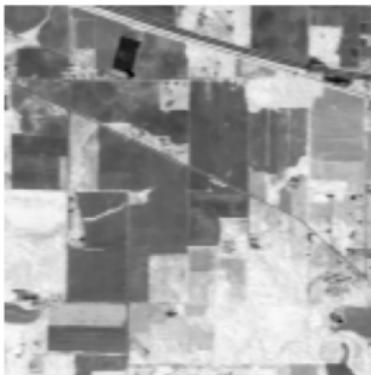


Figure: Composición a color 3 primeras PC

3 primeras PC - Indian Pine

PCA para Indian Pines

CP 1



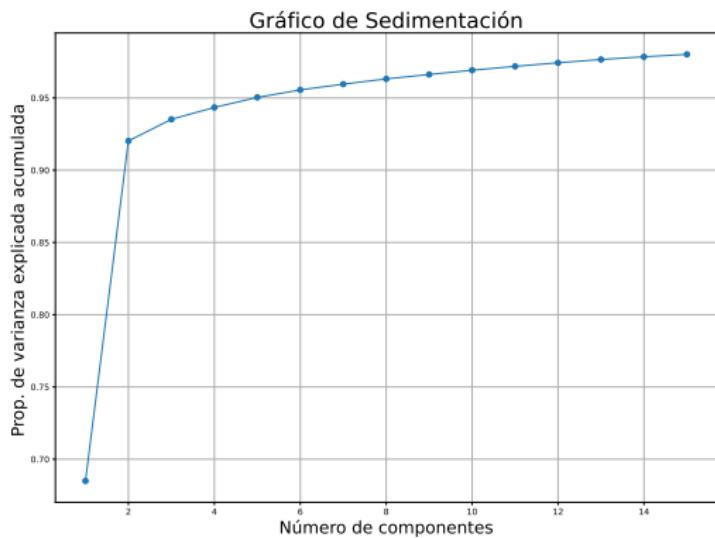
CP 2



CP 3



Objetivo 1: Proporción de varianza explicada



- Las 3 primeras PC explican más del 90% de la variabilidad de los datos
- Sin embargo, el objetivo final del estudio nos indicará qué sección del espectro electromagnético elegir previo al PCA: consultar con experto en dominio.

Objetivo 2: reducción de dimensionalidad + reconstrucción

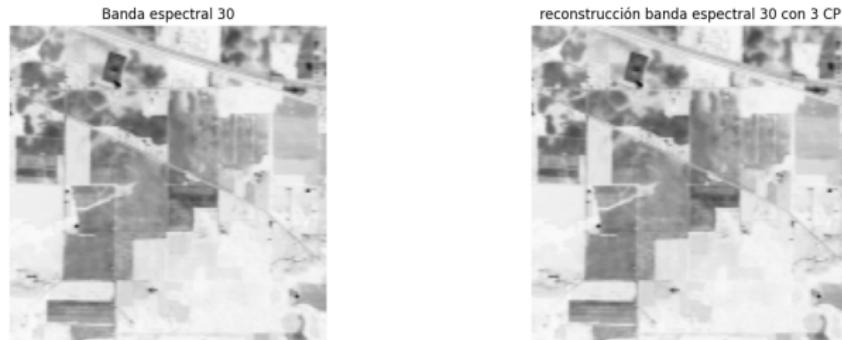


Figure: reconstrucción banda 30 usando solamente 3CP

Objetivo 2: reducción de dimensionalidad + reconstrucción

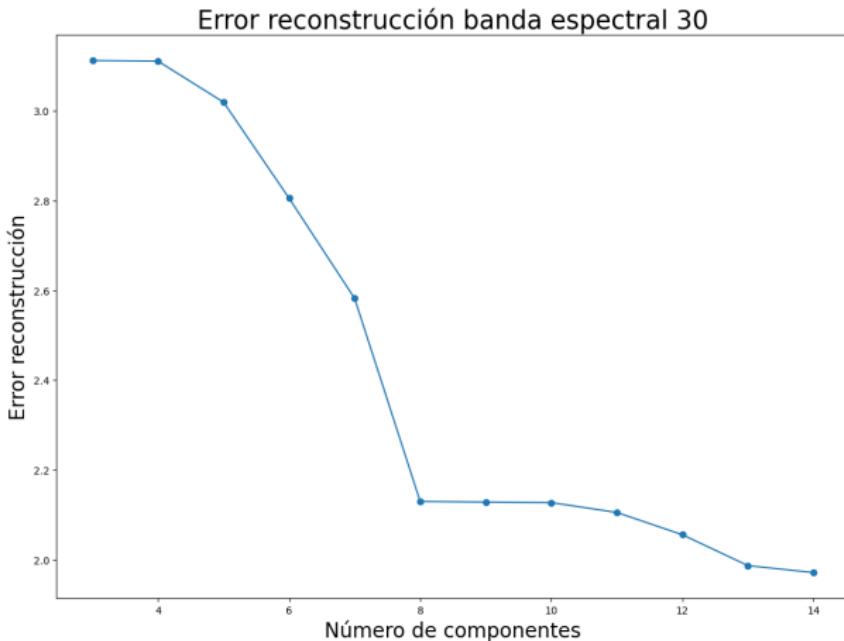


Figure: Error reconstrucción banda 30 según cantidad de CP usadas en la reconstrucción

Objetivo 2: reducción de dimensionalidad + reconstrucción

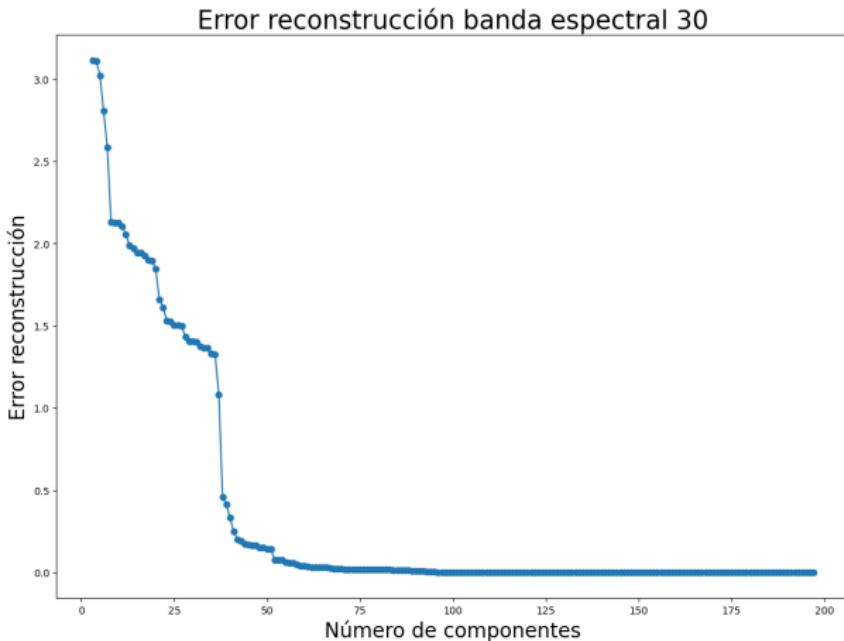


Figure: Error reconstrucción banda 30 según cantidad de CP usadas en la reconstrucción

Objetivo 2: dimensiones

- Tamaño Imagen Original = $145 \times 145 \times 200$
- Dim (X) = 21025×200
 $n = 21025, p = 200.$
Esto es: 4205000 datos.
- Luego de la reducción de dimensionalidad usando 3CP, para poder reconstruir la imagen deben almacenarse:
 - ▶ $\bar{x} \in \mathbb{R}^{200}$
 - ▶ $U_r \in \mathbb{R}^{200 \times 3}$ 63875 datos
 - ▶ Las proyecciones $z_i \in \mathbb{R}^3$,
 $i = 1, \dots, 21025$

Caso 3: Compresión de datos

Datos: Base de datos UTKFace

- Más de 20000 fotografías de rostros
- Personas de distintos géneros, razas, edades entre 0 y 116 años
- Distintas expresiones, iluminación, etc



Objetivo: reducir la dimensionalidad del dataset (compresión).

Preguntas:

- ¿Cómo se construye el conjunto de datos para poder realizar un PCA?
- ¿Cuáles son las observaciones? ¿Las variables?

Caso 3: Compresión de datos, base UTKFaces

Observaciones: cada una de las 20000 fotografías

Variables: cantidad de pixeles en cada fotografía

Método: expandir sobre las primeras componentes principales:



resultado:

Original



$$\approx \bar{\mathbf{X}} + \sum_{i=1}^M z_{ni} \mathbf{u}_i$$

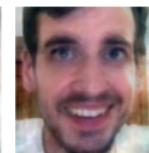
$M=1$



$M=10$



$M=50$



$M=150$



Bibliografía

Análisis de datos multivariantes, Daniel Peña, 2002, Mcgraw-hill Interamericana de España S.L.

Pattern recognition and machine learning, C.M. Bishop, 2006, Springer.