

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

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Métodos de visualización para datos multivariados

"Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent".

Carl Sagan (Contact)

Visualización de datos multivariados

Schmidt, 1954, "Handbook of Graphic Presentation":

- En comparación con otro tipo de presentación, las gráficas bien diseñadas son más efectivas en crear interés y llamar la atención de las personas.
- Las relaciones visuales dadas por las gráficas y figuras son entendidas y recordadas más fácilmente.
- El uso de gráficas y figuras ahorran tiempo, ya que el significado esencial del resultado de análisis estadísticos en los datos pueden visualizarse rápidamente.
- Proveen una representación adecuada para comprender el problema, mucho mejor que lo harían datos tabulados o escritos.
- Pueden resaltar características y relaciones ocultas, y pueden estimular y brindar una forma analítica de pensar e investigar.

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Visualización de datos multivariados

Algunos objetivos de las gráficas y figuras para datos multidimensionales:

- Dar un panorama general de los datos
- “Contar una historia”
- Sugerir hipótesis
- Criticar un modelo

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Visualización de datos multivariados



Visualizing the Broad Street Cholera Outbreak which helped find the root cause of the disease outbreak! (Source:
https://github.com/dipenjan/Sart_of_data_visualization)

Visualizing the Broad Street Cholera Outbreak

- A severe outbreak of cholera occurred in 1854 near Broad Street in the City of Westminster, London, England, unknowing to people causing over 600 deaths
- Physician Jon Snow identified the source of the outbreak as the public water pump on Broad Street
- Snow used a dot map to illustrate the cluster of cholera cases around the pump
- He also used statistics to illustrate the connection between the quality of the water source and cholera cases.

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Visualización de datos multivariados

En esta sección, veremos algunas técnicas de visualización como una forma de **representación** de datos multivariados.

No haremos énfasis en las técnicas de visualización actuales (y no tan actuales) de grandes volúmenes de datos, gráficos interactivos, o animaciones.

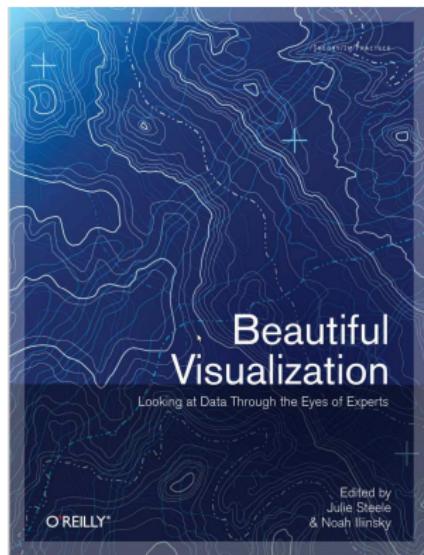
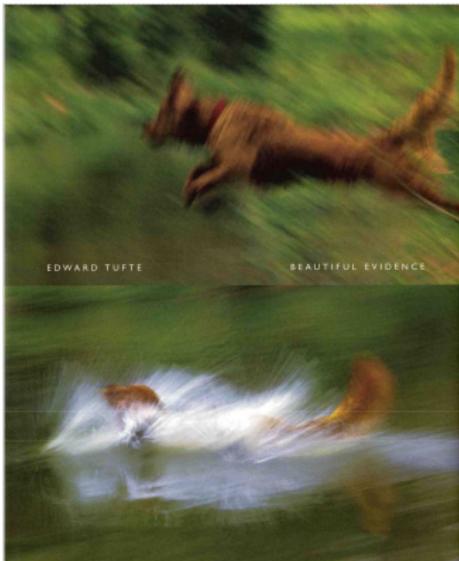
Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Visualización de datos multivariados

Para una referencia sobre éstas, y otras características de visualización de información, pueden recurrir a estos dos excelentes libros:



La maldición de la dimensionalidad

Generalidades

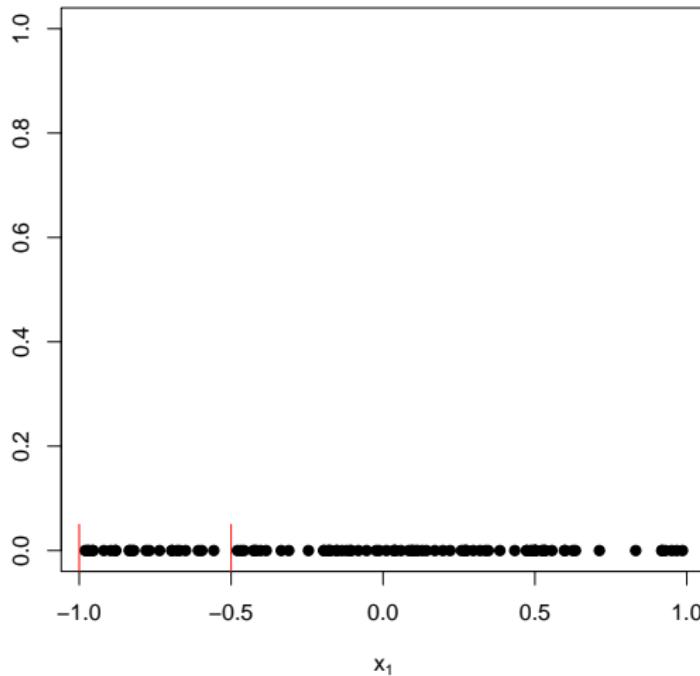
Introducción

Métodos de visualización y reducción de dimensión



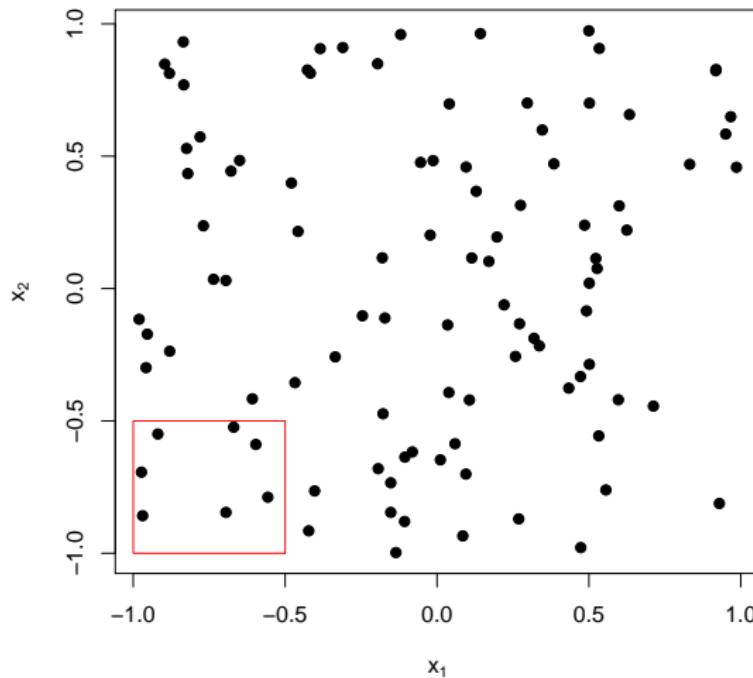
La maldición de la dimensionalidad

¿Porqué se complica la aplicación de los métodos de aprendizaje cuando pasamos a datos con mayores dimensiones?



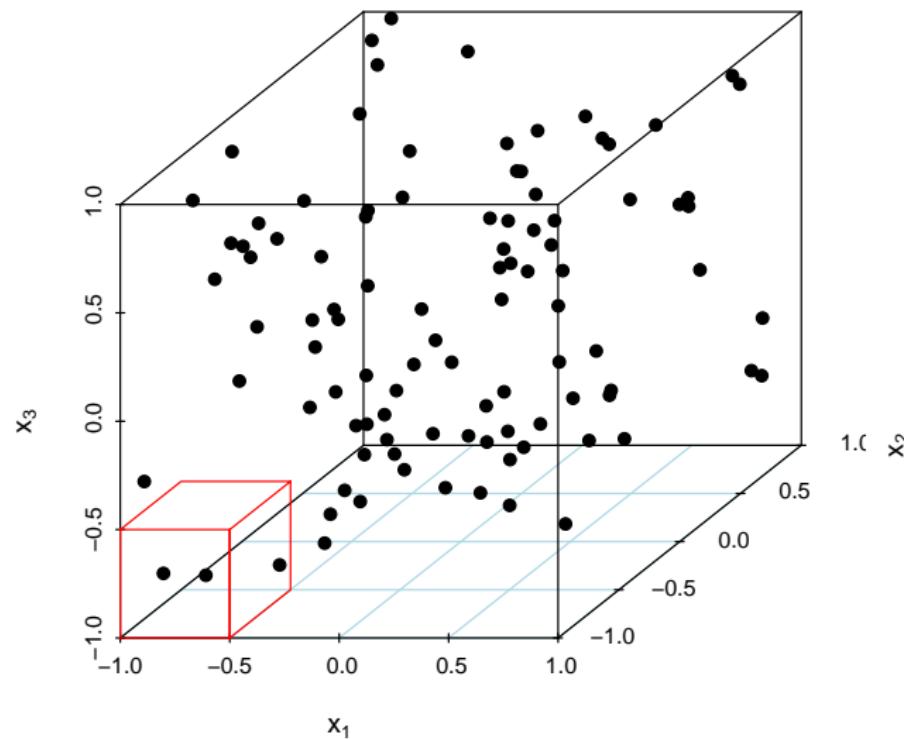
La maldición de la dimensionalidad

¿Porqué se complica la aplicación de los métodos de aprendizaje cuando pasamos a datos con mayores dimensiones?



La maldición de la dimensionalidad

¿Porqué se complica la aplicación de los métodos de aprendizaje cuando pasamos a datos con mayores dimensiones?



Generalidades

Introducción

Métodos de visualización y reducción de dimensión

La maldición de la dimensionalidad

La maldición de la dimensionalidad

- Nunca tendremos suficientes datos para cubrir cada parte de un espacio de entrada de alta dimensión, lo que hace extremadamente difícil determinar qué parte del espacio es importante para definir alguna relación entre los datos y cuál no lo es.
- Equivalentemente, la longitud necesaria en cada variable para cubrir cierto volumen de los datos, aumenta conforme aumenta la dimensión de los mismos.
- Es decir, en altas dimensiones, los datos tienden a ser “ralos” (sparse).

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

La maldición de la dimensionalidad

La maldición de la dimensionalidad

- Nunca tendremos suficientes datos para cubrir cada parte de un espacio de entrada de alta dimensión, lo que hace extremadamente difícil determinar qué parte del espacio es importante para definir alguna relación entre los datos y cuál no lo es.
- Equivalentemente, la longitud necesaria en cada variable para cubrir cierto volumen de los datos, aumenta conforme aumenta la dimensión de los mismos.
- Es decir, en altas dimensiones, los datos tienden a ser “ralos” (sparse).

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

La maldición de la dimensionalidad

La maldición de la dimensionalidad

- Nunca tendremos suficientes datos para cubrir cada parte de un espacio de entrada de alta dimensión, lo que hace extremadamente difícil determinar qué parte del espacio es importante para definir alguna relación entre los datos y cuál no lo es.
- Equivalentemente, la longitud necesaria en cada variable para cubrir cierto volumen de los datos, aumenta conforme aumenta la dimensión de los mismos.
- Es decir, en altas dimensiones, los datos tienden a ser “ralos” (sparse).

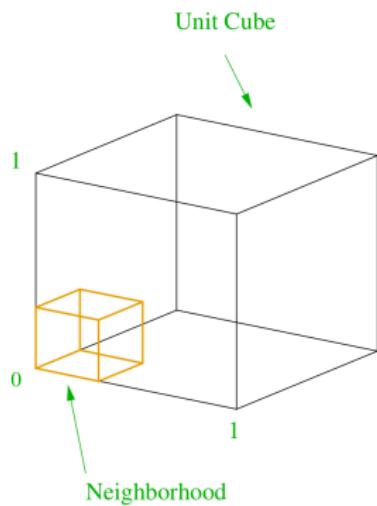
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

La maldición de la dimensionalidad

Considera datos uniformemente distribuidos en un hipercubo con longitud A en cada lado.



Si queremos abarcar una proporción p del hipercubo con un *subcubo* de longitud $A - \epsilon$ en cada lado, vemos que:

- La longitud proporcional de cada lado del subcubo es

$$A_d(p) = p^{1/d} \rightarrow 1 \text{ cuando } d \rightarrow \infty,$$

donde d es la dimensión de los datos.

Y de forma equivalente:

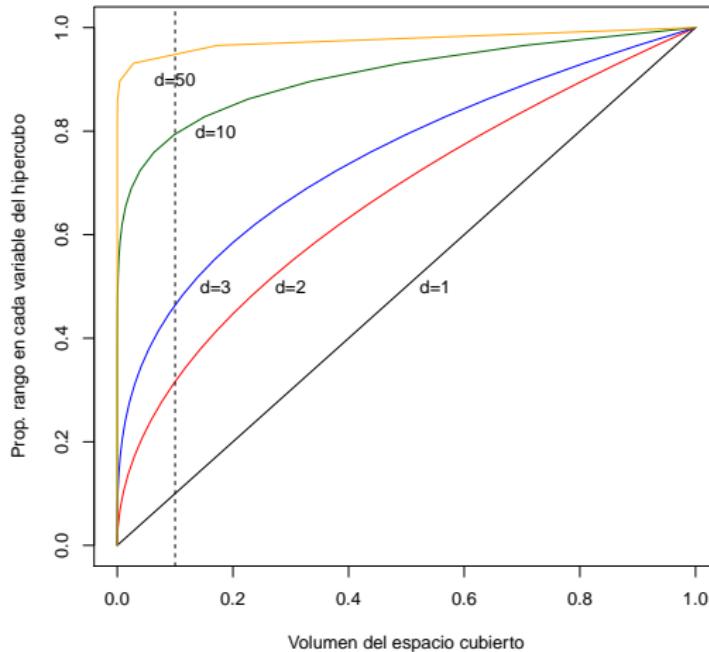
- La diferencia proporcional del volumen entre los dos cubos es

$$\frac{A^d - (A - \epsilon)^d}{A^d} = 1 - \left(1 - \frac{\epsilon}{A}\right)^d \rightarrow 1 \text{ cuando } d \rightarrow \infty.$$

La maldición de la dimensionalidad

Ejemplo para $A = 1$ (hipercubo unitario) y diferente dimensionalidad.

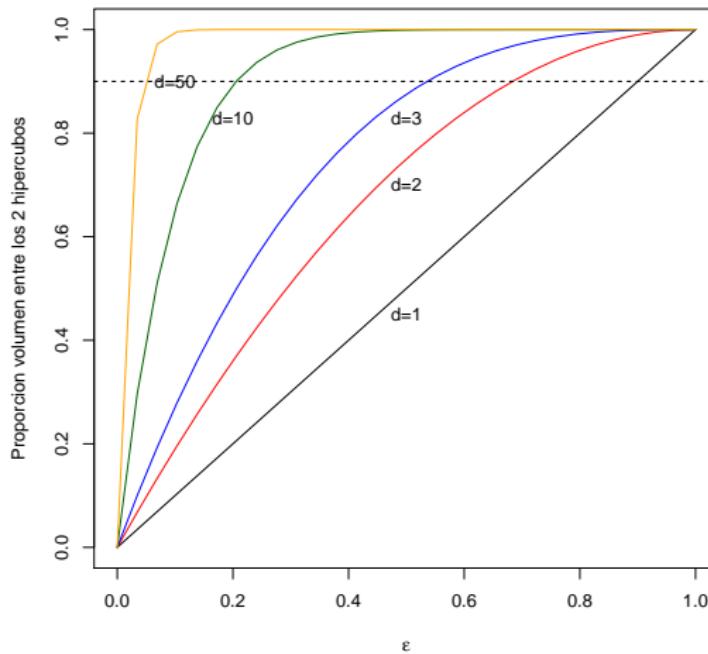
Supongamos que queremos capturar el 10 % de los datos...



La maldición de la dimensionalidad

Ejemplo para $A = 1$ (hipercubo unitario) y diferente dimensionalidad.

Supongamos que queremos capturar el 10 % de los datos...



Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

La maldición de la dimensionalidad

Entonces:

- Conforme aumenta el número de dimensiones, casi todo el volúmen dentro de una subregión del hipercubo definido por el espacio de entrada se aproxima a la superficie del hipercubo, en vez del centroide u otra región.
- El mismo fenómeno ocurre cuando consideramos regiones esféricas.

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Métodos básicos

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

- Gráficos de dispersión (scatterplot).
- Box plots y gráficos relacionados
- Bubbleplot
- Chernoff's faces
- Starplot
- entre otros...

[notebooks/2-visualizacion.ipynb](#)

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Gráficos dinámicos e interactivos para datos multivariados

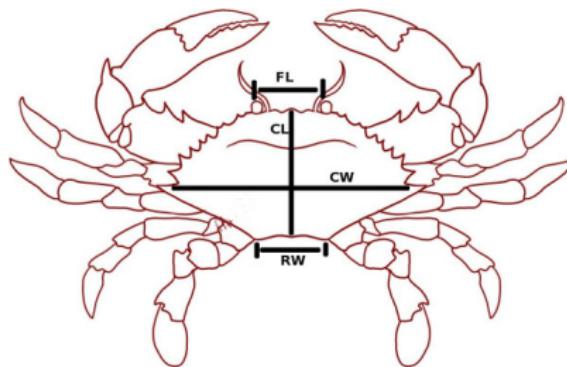
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos estáticos

Considera los siguientes datos de mediciones morfológicas de cangrejos australianos ¹



Un objetivo interesante sería indagar si podemos determinar el sexo o la especie de los cangrejos basados en sus mediciones.

¹ Campbell, N. A. and Mahon, R. J. (1974), A Multivariate Study of Variation in Two Species of Rock Crab of genus Leptograpsus, Australian Journal of Zoology 22, 417–425

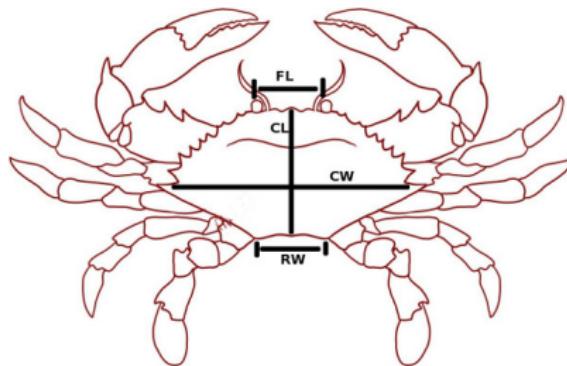
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos estáticos

Considera los siguientes datos de mediciones morfológicas de cangrejos australianos ¹



Un objetivo interesante sería indagar si podemos determinar el sexo o la especie de los cangrejos basados en sus mediciones.

¹ Campbell, N. A. and Mahon, R. J. (1974), A Multivariate Study of Variation in Two Species of Rock Crab of genus Leptograpsus, Australian Journal of Zoology 22, 417–425

Gráficos estáticos

Consideraremos 200 datos de mediciones morfológicas de cangrejos australianos:

	sp	sex	index	FL	RW	CL	CW	BD
1	B	M	1	8.1	6.7	16.1	19.0	7.0
2	B	M	2	8.8	7.7	18.1	20.8	7.4
3	B	M	3	9.2	7.8	19.0	22.4	7.7
4	B	M	4	9.6	7.9	20.1	23.1	8.2
5	B	M	5	9.8	8.0	20.3	23.0	8.2
6	B	M	6	10.8	9.0	23.0	26.5	9.8

sp: species, B or O for blue or orange.

sex: as it says.

index: index 1:50 within each of the four groups.

FL: frontal lobe size (mm).

RW: rear width (mm).

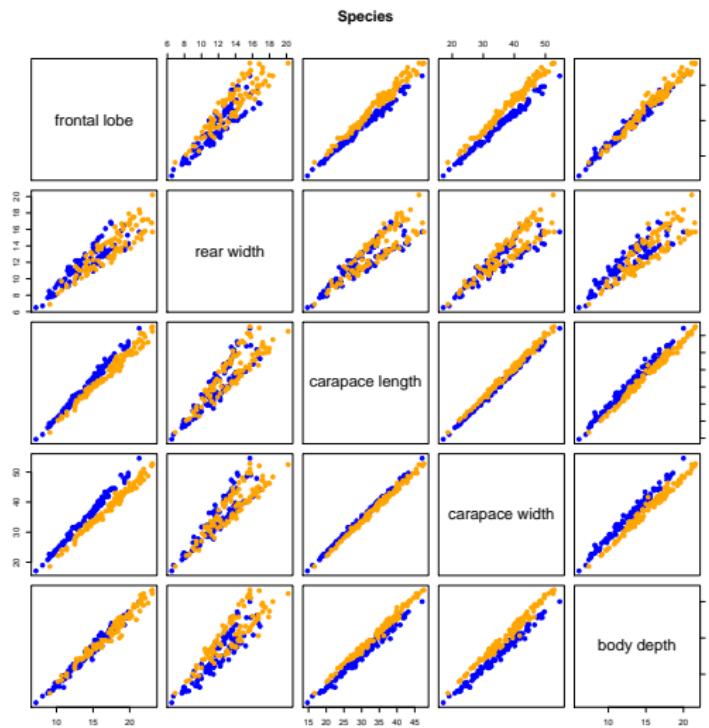
CL: carapace length (mm).

CW: carapace width (mm).

BD: body depth (mm).

Gráficos estáticos

Métodos de visualización y reducción de dimensión



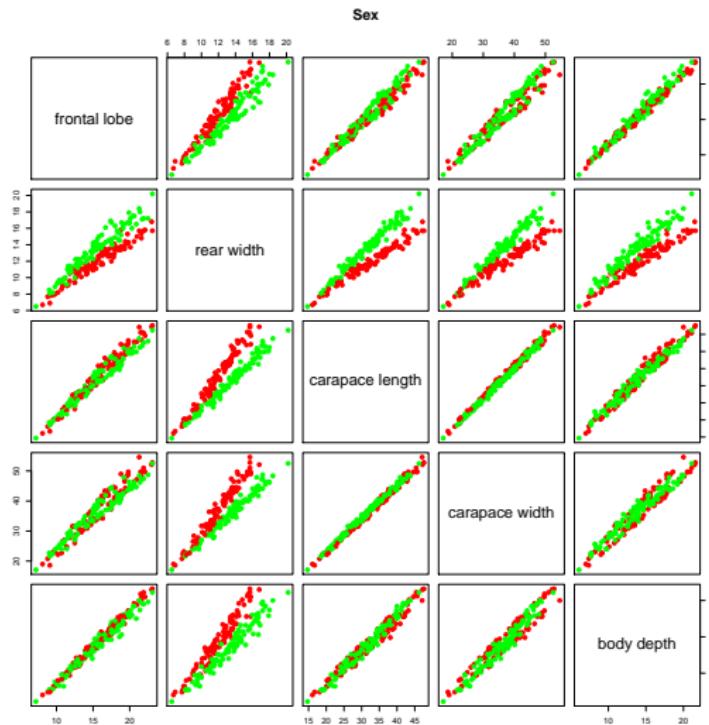
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos estáticos

Femenino, Masculino



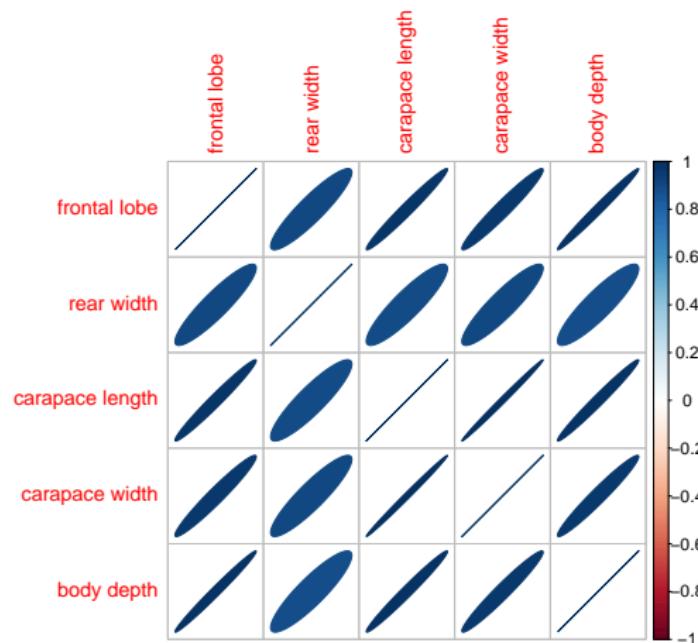
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos estáticos

Correlation plot



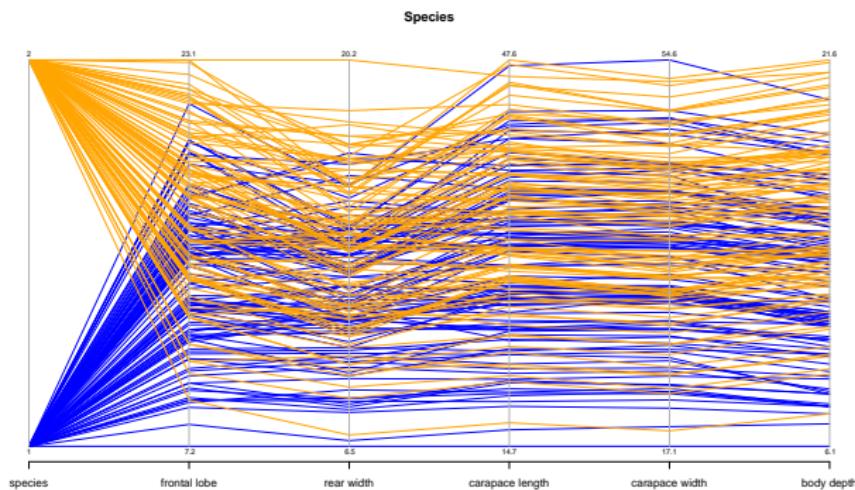
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos estáticos

Parallel coordinate plots (Inselberg 1985, Wegman 1990).

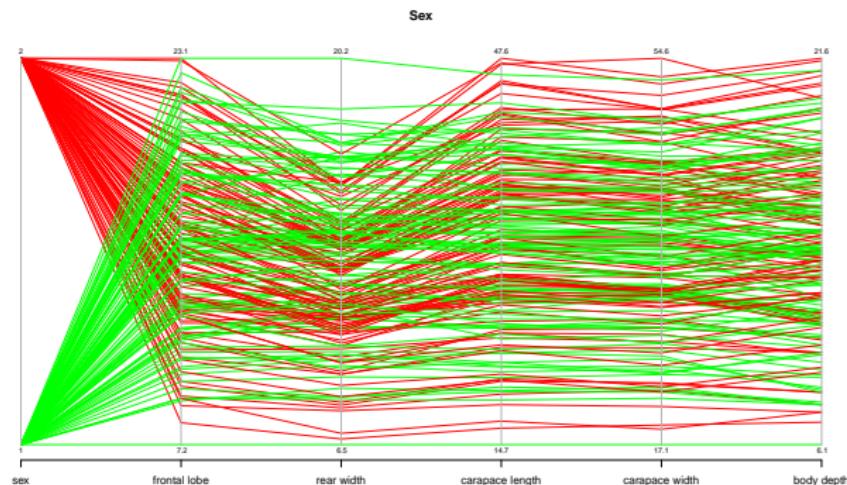


Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos estáticos



Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Métodos de reducción de dimensión

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Proyecciones en baja dimensión



Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Proyecciones en baja dimensión

Proyecciones en baja dimensión para variables continuas.

- Permite generar gráficas (estáticas o dinámicas), para estudiar la distribución conjunta de datos multivariados.
- Nos permite buscar relaciones que pueden involucrar varias variables.
- Se pueden construir gráficas animadas (“Tour”) generando una secuencia de proyecciones de baja dimensión (1D, 2D, 3D) que sean **interesantes**.

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Proyecciones en baja dimensión

Proyecciones en baja dimensión para variables continuas.

- Permite generar gráficas (estáticas o dinámicas), para estudiar la distribución conjunta de datos multivariados.
- Nos permite buscar relaciones que pueden involucrar varias variables.
- Se pueden construir gráficas animadas (“Tour”) generando una secuencia de proyecciones de baja dimensión (1D, 2D, 3D) que sean **interesantes**.

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Proyecciones en baja dimensión

Proyecciones en baja dimensión para variables continuas.

- Permite generar gráficas (estáticas o dinámicas), para estudiar la distribución conjunta de datos multivariados.
- Nos permite buscar relaciones que pueden involucrar varias variables.
- Se pueden construir gráficas animadas (“Tour”) generando una secuencia de proyecciones de baja dimensión (1D, 2D, 3D) que sean **interesantes**.

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Proyecciones en baja dimensión

Proyecciones en baja dimensión para variables continuas.

- Permite generar gráficas (estáticas o dinámicas), para estudiar la distribución conjunta de datos multivariados.
- Nos permite buscar relaciones que pueden involucrar varias variables.
- Se pueden construir gráficas animadas (“Tour”) generando una secuencia de proyecciones de baja dimensión (1D, 2D, 3D) que sean **interesantes**.

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Proyecciones en baja dimensión

Considera una matriz de n datos $\mathbf{X}_{n \times d}$, es decir, $\mathbf{x} \in \mathbb{R}^d$. Definimos una matriz de proyección ortonormal

$$\mathbf{A}_{d \times r},$$

donde d es la dimensión de los datos y r es la dimensión de la proyección (1, 2 o 3). La proyección está dada por

$$\mathbf{P} = (\mathbf{XA})_{n \times r}$$

Generalidades

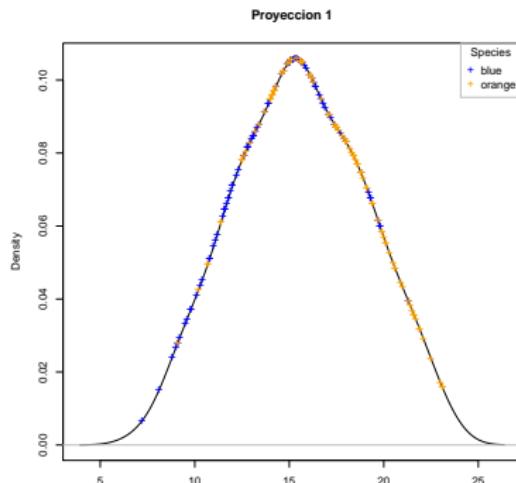
Introducción

Métodos de visualización y reducción de dimensión

Proyecciones en baja dimensión

Por ejemplo, para las 5 mediciones físicas de los cangrejos, una proyección puede ser:

$$\mathbf{A}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{P}_1 = \begin{pmatrix} 8.1 \\ 8.8 \\ 9.2 \\ 9.6 \\ 9.8 \\ \vdots \end{pmatrix}$$



Generalidades

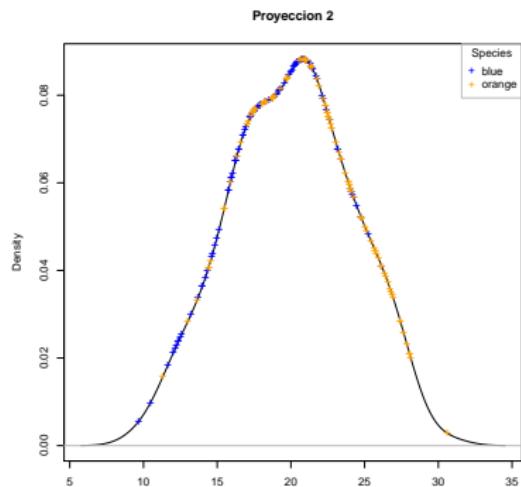
Introducción

Métodos de visualización y reducción de dimensión

Proyecciones en baja dimensión

O también:

$$\mathbf{A}_2 = \begin{pmatrix} 0.707 \\ 0.707 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{P}_2 = \begin{pmatrix} 8.1 \\ 8.8 \\ 9.2 \\ 9.6 \\ 9.8 \\ \vdots \end{pmatrix}$$



Generalidades

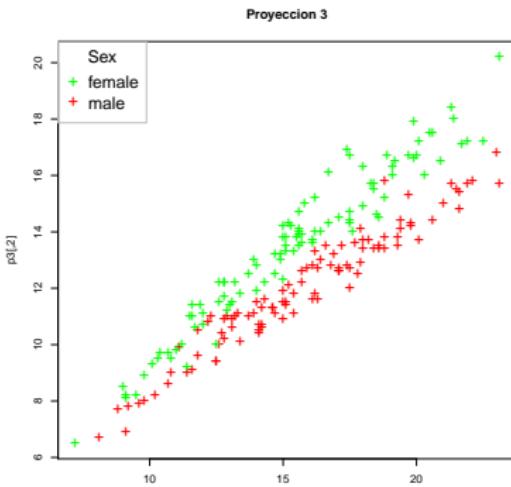
Introducción

Métodos de visualización y reducción de dimensión

Proyecciones en baja dimensión

En dos dimensiones, podemos obtener:

$$\mathbf{A}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{P}_3 = \begin{pmatrix} 8.1 & 6.7 \\ 8.8 & 7.7 \\ 9.2 & 7.8 \\ 9.6 & 7.9 \\ 9.8 & 8.0 \\ \vdots & \vdots \\ \vdots & \vdots \end{pmatrix}$$



Generalidades

Introducción

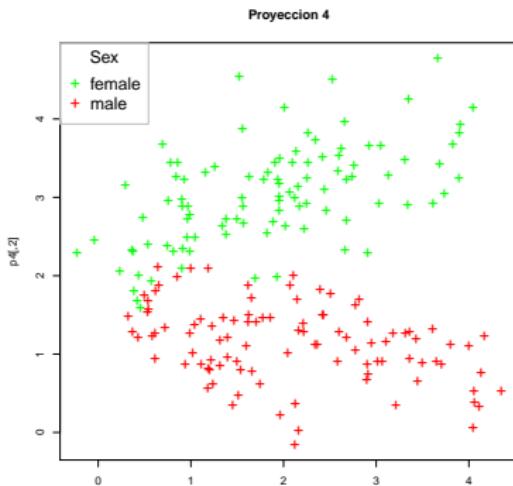
Métodos de visualización y reducción de dimensión

Proyecciones en baja dimensión

Y también:

$$A_4 = \begin{pmatrix} 0 & 0 \\ 0 & 0.950 \\ 0 & -0.312 \\ -0.312 & 0 \\ 0.950 & 0 \end{pmatrix}$$

$$P_4 = \begin{pmatrix} 8.1 & 6.7 \\ 8.8 & 7.7 \\ 9.2 & 7.8 \\ 9.6 & 7.9 \\ 9.8 & 8.0 \\ \vdots & \vdots \end{pmatrix}$$



Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

Projection Tours (D. Asimov, 1985. Buja et al., 2005, D. Cook & D. Swayne, 2007)

Una forma de generar proyecciones para explorar características *interesantes* en los datos y mostrando una *sensación de movimiento...*

- Genera una proyección inicial
- Genera una siguiente proyección
- Interpola entre ambas proyecciones ²

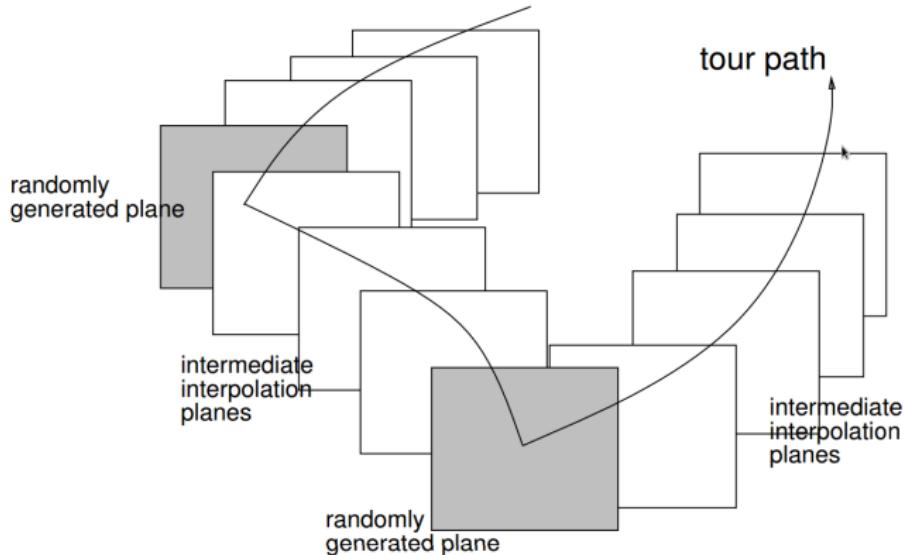
² La interpolación no es trivial. Puedes ver el procedimiento a detalle en: Buja, A., Cook, D., Asimov, D. & Hurley, C. (2005), Computational Methods for High-Dimensional Rotations in Data Visualization, in C. R. Rao, E. J. Wegman & J. L. Solka, eds, *Handbook of Statistics: Data Mining and Visualization*, Elsevier/North Holland, <http://www.elsevier.com>, pp. 391–413.

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

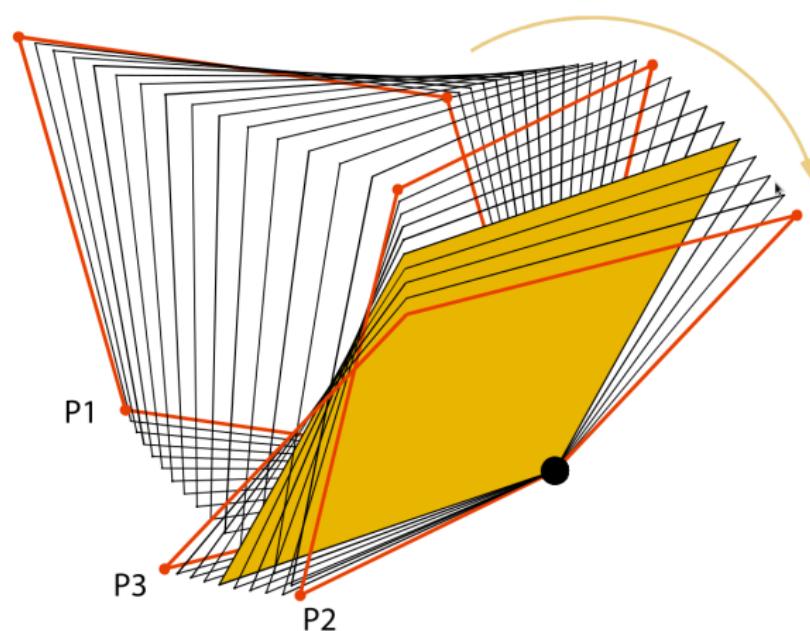


Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Gráficos interactivos para datos multivariados



Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Grand Tour: el recorrido es aleatorio,

- Las proyecciones son elegidas aleatoriamente, de tal forma que el espacio se cubre de manera eficiente.
- Puede considerarse como una caminata aleatoria en el espacio de todas las proyecciones.
- Las proyecciones aleatorias se generan muestreando de una distribución normal multivariada.
 - Muestra d valores de una distribución normal estándar univariada, y ese vector es entonces una muestra de una normal multivariada
 - Estandariza el vector para que tenga longitud 1, lo que nos da un punto aleatorio en una hiper-esfera
 - Realiza los mismos pasos para proyecciones en $r > 2$ dimensiones, pero ortonormalizando los vectores

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Grand Tour: el recorrido es aleatorio,
 - Las proyecciones son elegidas aleatoriamente, de tal forma que el espacio se cubre de manera eficiente.
 - Puede considerarse como una caminata aleatoria en el espacio de todas las proyecciones.
 - Las proyecciones aleatorias se generan muestreando de una distribución normal multivariada.
 - Muestra d valores de una distribución normal estándar univariada, y ese vector es entonces una muestra de una normal multivariada
 - Estandariza el vector para que tenga longitud 1, lo que nos da un punto aleatorio en una hiper-esfera
 - Realiza los mismos pasos para proyecciones en $r > 2$ dimensiones, pero ortonormalizando los vectores

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Grand Tour: el recorrido es aleatorio,
 - Las proyecciones son elegidas aleatoriamente, de tal forma que el espacio se cubre de manera eficiente.
 - Puede considerarse como una caminata aleatoria en el espacio de todas las proyecciones.
 - Las proyecciones aleatorias se generan muestreando de una distribución normal multivariada.
 - Muestra d valores de una distribución normal estándar univariada, y ese vector es entonces una muestra de una normal multivariada
 - Estandariza el vector para que tenga longitud 1, lo que nos da un punto aleatorio en una hiper-esfera
 - Realiza los mismos pasos para proyecciones en $r > 2$ dimensiones, pero ortonormalizando los vectores

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Grand Tour: el recorrido es aleatorio,
 - Las proyecciones son elegidas aleatoriamente, de tal forma que el espacio se cubre de manera eficiente.
 - Puede considerarse como una caminata aleatoria en el espacio de todas las proyecciones.
 - Las proyecciones aleatorias se generan muestreando de una distribución normal multivariada.
 - Muestra d valores de una distribución normal estándar univariada, y ese vector es entonces una muestra de una normal multivariada
 - Estandariza el vector para que tenga longitud 1, lo que nos da un punto aleatorio en una hiper-esfera
 - Realiza los mismos pasos para proyecciones en $r > 2$ dimensiones, pero ortonormalizando los vectores

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

- Grand Tour.

Ejemplo: Considera proyecciones en 1 dimensión a partir de $d = 3$ variables de los datos de cangrejos Australianos ($\mathbb{R}^3 \rightarrow \mathbb{R}$). Generamos aleatoriamente 100 proyecciones muestreando de una Normal multivariada. Las primeras 6 proyecciones son:

$$\mathbf{A}_1 = \left(\begin{array}{c|c} \text{FL} & 0.59 \\ \text{RW} & -0.76 \\ \text{CL} & -0.26 \end{array} \right) \quad \mathbf{A}_2 = \left(\begin{array}{c|c} \text{FL} & 0.79 \\ \text{RW} & -0.42 \\ \text{CL} & -0.46 \end{array} \right) \quad \mathbf{A}_3 = \left(\begin{array}{c|c} \text{FL} & -0.64 \\ \text{RW} & -0.07 \\ \text{CL} & -0.76 \end{array} \right)$$

$$\mathbf{A}_4 = \left(\begin{array}{c|c} \text{FL} & -0.17 \\ \text{RW} & 0.63 \\ \text{CL} & -0.76 \end{array} \right) \quad \mathbf{A}_5 = \left(\begin{array}{c|c} \text{FL} & 0.87 \\ \text{RW} & 0.15 \\ \text{CL} & 0.46 \end{array} \right) \quad \mathbf{A}_6 = \left(\begin{array}{c|c} \text{FL} & 0.64 \\ \text{RW} & 0.55 \\ \text{CL} & -0.54 \end{array} \right)$$

Generalidades

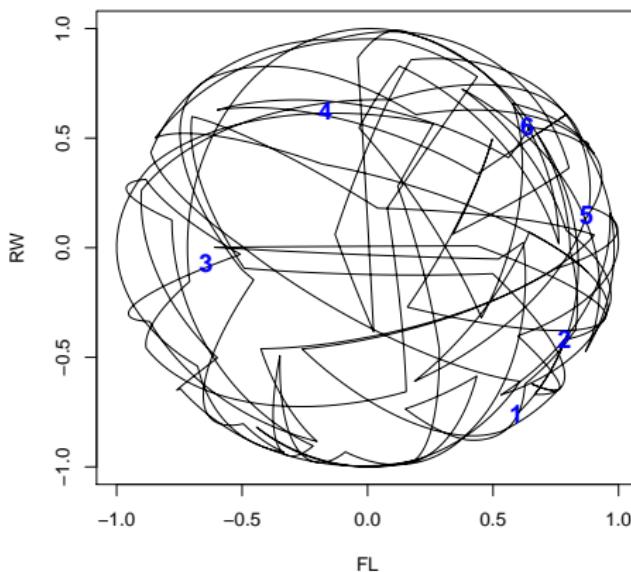
Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

- Grand Tour.

Ojo: el orden en que se generaron las proyecciones no indica el orden del recorrido. Este orden se elige con otro criterio (el de la interpolación).



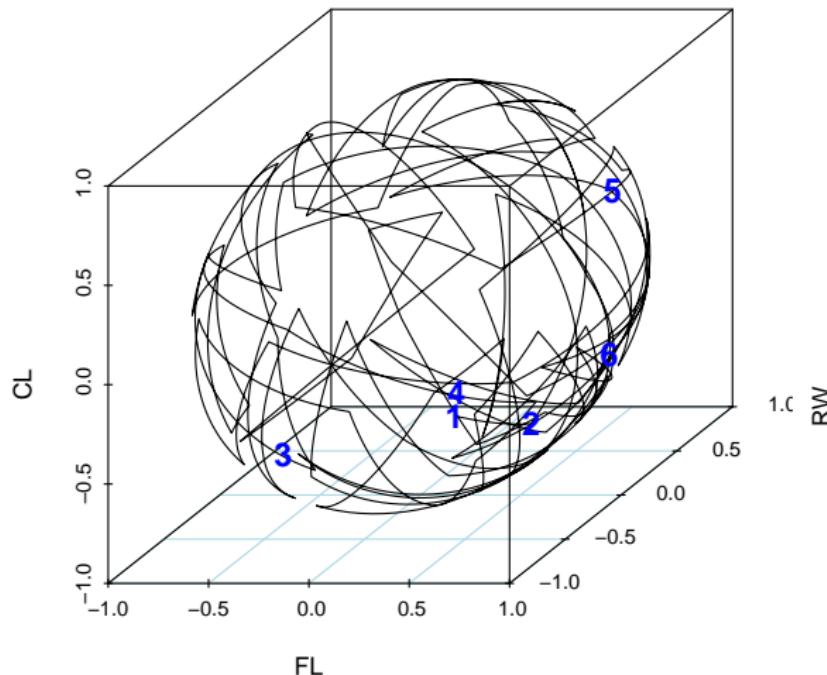
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

- Grand Tour.



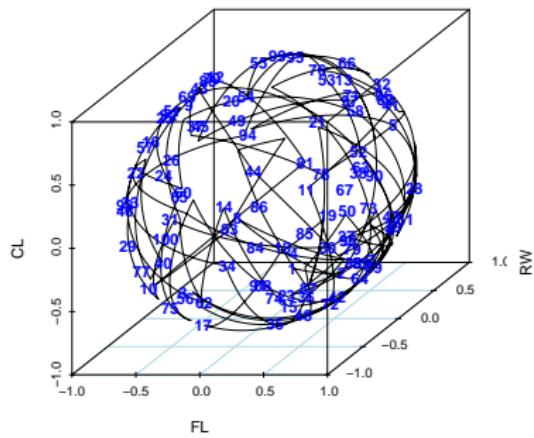
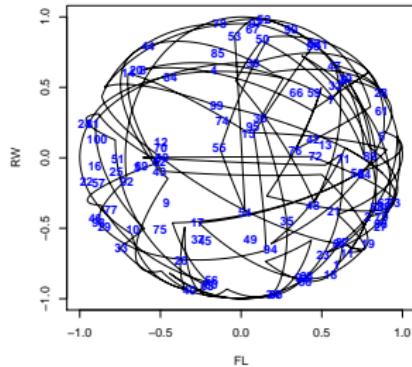
Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

● Grand Tour.



Gráficos interactivos para datos multivariados

Ahora, veamos las proyecciones 1D mediante un gráfico de densidad:

Gráficos interactivos para datos multivariados

Repetimos lo mismo pero ahora usando las $d = 5$ variables
 $(\mathbb{R}^5 \rightarrow \mathbb{R})$

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Manipulación manual: tu mismo eliges el recorrido, es decir, la secuencia de proyecciones

Ejercicio en ggobi.

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Ejemplo: olive oil dataset



Primary question: How do we distinguish the oils from different regions and areas in Italy based on their combinations of the fatty acids?

Olive oil samples from Italy

Description:

This data is from a paper by Forina, Armanino, Lanteri, Tiscornia (1983) Classification of Olive Oils from their Fatty Acid Composition, in Martens and Russwurm (ed) Food Research and Data Analysis.

Format:

A 572 x 10 numeric array

Details:

- region Three super-classes of Italy: North, South and the island of Sardinia
- area Nine collection areas: three from North, four from South and 2 from Sardinia
- palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic fatty acids percent x 100

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Ejemplo: olive oil dataset

	region		area	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic
1	1	North-Apulia		1075	75	226	7823	672	36
2	1	North-Apulia		1088	73	224	7709	781	31
3	1	North-Apulia		911	54	246	8113	549	31
4	1	North-Apulia		966	57	240	7952	619	50
5	1	North-Apulia		1051	67	259	7771	672	50
6	1	North-Apulia		911	49	268	7924	678	51
			arachidic	eicosenoic					
1			60	29					
2			61	29					
3			63	29					
4			78	35					
5			80	46					
6			70	44					

Veamos en ggobi

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Recorrido guiado: Projection Pursuit (Friedman and Tukey, *“A projection pursuit algorithm for exploratory data analysis”*, 1974).
Es un método desarrollado para descubrir proyecciones de baja dimensión “interesantes” en datos de alta dimensión.
Originalmente, fue diseñado para buscar proyecciones generalmente No-Gaussianas, es decir, para detectar características no Gaussianas (¿Porqué?).

Generalidades

Introducción

Métodos de
visualización y
reducción de
dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Recorrido guiado: Projection Pursuit (Friedman and Tukey, *“A projection pursuit algorithm for exploratory data analysis”*, 1974).
Es un método desarrollado para descubrir proyecciones de baja dimensión “interesantes” en datos de alta dimensión.
Originalmente, fue diseñado para buscar proyecciones generalmente No-Gaussianas, es decir, para detectar características no Gaussianas (¿Porqué?).

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

¿Cómo quieres tu recorrido?

- Recorrido guiado: Projection Pursuit.
La estrategia es la siguiente:

- ➊ Define un índice de proyección I para evaluar el desempeño de las proyecciones generadas.
- ➋ Usa un algoritmo de optimización para resolver

$$\max_{\mathbf{A}} I = f(\mathbf{XA}),$$

buscando óptimos locales y globales (si los hay) del índice utilizado, sobre todas las proyecciones r -dimensionales de los datos (generalmente, $r = 1, 2, 3$).

Gráficos interactivos para datos multivariados

- Recorrido guiado (projection pursuit). El método propuesto por D. Cook et al. e implementado en ggobi, es una combinación de **recocido simulado** e interpolación de las proyecciones para visualizar las etapas de la optimización.

Recocido simulado para PP

- 1: Define proyección inicial \mathbf{A}_0 e índice inicial $I_0 = f(\mathbf{X}\mathbf{A}_0)$.
- 2: Define valores de enfriamiento c_0 , T_0 , criterio de paro ϵ .
- 3: **for** $i = 1, 2, \dots$ **do**
- 4: Genera una proyección en la vecindad de \mathbf{A}_0 :

$$\mathbf{A}_i^* = \mathbf{A}_0 + c_0 \mathbf{A}_i$$

- 5: Calcula $I_i = f(\mathbf{X}\mathbf{A}_i^*)$, $\Delta I_i = I_i - I_0$, $T_i = \frac{T_0}{\log(i+1)}$
- 6: Escoge $\mathbf{A}_0 = \mathbf{A}_i^*$ e $I_0 = I_i$ con probabilidad
 $\rho = \min\left(e^{\Delta I_i / T_i}, 1\right)$
- 7: Interpola $\mathbf{A}_0 \rightarrow \mathbf{A}_i^*$ si hay un cambio
- 8: $i++$
- 9: Repite hasta que $\Delta I_i < \epsilon$
- 10: **end for**

Generalidades

Introducción

Métodos de visualización y reducción de dimensión

Gráficos interactivos para datos multivariados

- Recorrido guiado (projection pursuit).
PP indexes. Define $\mathbf{y} = (\mathbf{XA})$

- Holes

$$I = \frac{1 - \frac{1}{n} \sum_i \exp(-\frac{1}{2}\mathbf{y}_i \mathbf{y}'_i)}{1 - \exp(-d/2)}$$

- Central mass

$$I = \frac{\frac{1}{n} \sum_i \exp(-\frac{1}{2}\mathbf{y}_i \mathbf{y}'_i) - \exp(-d/2)}{1 - \exp(-d/2)}$$

- PCA (veremos más adelante...)
 - LDA

$$I = 1 - \frac{|\mathbf{A}' \mathbf{W} \mathbf{A}|}{|\mathbf{A}' (\mathbf{W} + \mathbf{B}) \mathbf{A}|},$$

también veremos más adelante.