STATISTICA MULTIVARIATA

Corso di laurea in matematica | Anno accademico 2004/05 | Numero 1

Difficoltà in alta dimensione

L'obiettivo della statistica multidimensionale geometrica è di scoprire relazioni tra dati rappresentati da punti in spazi $\mathbb{R}_m$ ad alta dimensione con ad esempio $40 \leq m \leq 100$. Una delle maggiori difficoltà in questo intento è la cosiddetta maledizione dell'alta dimensione (che nella letteratura inglese è nota sotto il termine di *curse of dimensionality*), dovuta soprattutto al fatto che i concetti metrici in spazi a così alte dimensioni perdono gran parte del loro significato perché il volume

$$\frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)}$$

della palla di raggio 1 in $\mathbb{R}_m$ converge rapidamente a zero; per il volume v_m della palla di raggio $\frac{1}{2}$ in $\mathbb{R}_m$, cioè della palla iscritta a un cubo unitario, si ha ad esempio, come vedremo, la formula di ricorsione

$$v_m = \frac{\pi}{2m} v_{m-2}$$

e ciò comporta che già nel $\mathbb{R}_{10}$ la palla iscritta occupa solo il 2.5 per mille del volume del cubo. In un cubo ad alta dimensione perciò il volume è concentrato vicino al bordo e ciò crea notevoli problemi per l'interpretazione statistica di considerazioni metriche e gli algoritmi che le utilizzano. Il primo numero del corso è dedicato alla discussione di queste difficoltà che proprio nella statistica medica, uno dei campi in cui attualmente sono prodotte grandi quantità di dati ad alta dimensione, vengono spesso trascurate.

In questo numero

Sfere in $\mathbb{R}_m$

Situazione 1.1. Siano $m \in \mathbb{N}$ ed $r \in \mathbb{R}$ con $r \geq 0$. Useremo in tutto il corso m per indicare la dimensione dello spazio in cui si trovano i nostri dati.

Per $\alpha \in \mathbb{R}$ denotiamo con $[\alpha]$ la parte intera di α.

Definizione 1.2. Per $m > 0$ sia $v_m(r)$ il volume di una palla di raggio r in $\mathbb{R}_m$. Useremo l'abbreviazione

$$v_m := v_m(\tfrac{1}{2})$$

v_m è quindi il volume di una palla *iscritta* a un cubo di lato 1 in $\mathbb{R}_m$. Il volume del cubo è naturalmente uguale a 1.

Poniamo $v_0(r) := 1$ e perciò anche $v_0 = 1$.

Osservazione 1.3. $v_1(r) = 2r$ *e quindi* $v_1 = 1$.

Nota 1.4. Denotiamo con Γ la *funzione gamma* che, come è noto dall'analisi, è in primo luogo un'interpolazione del fattoriale, che però appare in molti altri campi della matematica e deve essere considerata come la più importante funzione non elementare. Essa è definita e olomorfa su tutto il piano complesso tranne nei punti z della forma $z = -k$ con $k \in \mathbb{N}$ e soddisfa l'equazione funzionale

$$\Gamma(z+1) = z\Gamma(z)$$

per $z \in \mathbb{C} \setminus (-\mathbb{N})$. Vale la condizione iniziale $\Gamma(1) = 1$, da cui per induzione si ha come conseguenza immediata che

$$\Gamma(n+1) = n!$$

per ogni $n \in \mathbb{N}$. Si dimostra inoltre che

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

Purtroppo, per pure ragioni storiche, la funzione è definita in modo tale che al fattoriale $n!$ non corrisponde l'argomento n in Γ.

Una trattazione molto dettagliata della funzione Γ si trova nel testo di analisi complessa di Remmert.

Teorema 1.5. *Il volume della palla unitaria in* $\mathbb{R}_m$ *è dato da*

$$v_m(1) = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)}$$

Dimostrazione. Corsi di analisi. La formula è facile da ricordare nella forma

$$v_m(1) = \frac{\pi^{\frac{m}{2}}}{\frac{m}{2}!}$$

che è corretta per m pari e può essere considerata come abbreviazione simbolica nel caso che m sia dispari.

Proposizione 1.6. *Valgono le formule di ricorsione*

$$v_m(1) = \frac{2\pi}{m} v_{m-2}(1)$$
$$v_m(r) = \frac{2\pi r^2}{m} v_{m-2}(r)$$
$$v_m = \frac{\pi}{2m} v_{m-2}$$

per ogni $m \geq 2$.

Dimostrazione. (1) Abbiamo

$$v_m(1) = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)} = \frac{\pi\pi^{\frac{m-2}{2}}}{\frac{m}{2}\Gamma(\frac{m}{2})}$$
$$= \frac{\pi\pi^{\frac{m-2}{2}}}{\frac{m}{2}\Gamma(\frac{m-2}{2}+1)}$$
$$= \frac{2\pi}{m}\frac{\pi^{\frac{m-2}{2}}}{\Gamma(\frac{m-2}{2}+1)} = \frac{2\pi}{m} v_{m-2}(1)$$

Ciò mostra la prima formula da cui si ottengono facilmente le altre due perché è chiaro che aumentando la dimensione di due il volume deve essere moltiplicato con r^2.

Corollario 1.7. $\lim_{m \longrightarrow \infty} v_m(r) = 0$

per ogni r.

Dimostrazione. Per $m \geq 4\pi r^2$ si ha

$$v_m(r) = \frac{2\pi r^2}{m} v_{m-2}(r) \leq \frac{1}{2} v_{m-2}(r)$$

Calcolo di v_m

Possiamo così scrivere una funzione in R che ci permette di calcolare v_m.

```
M.volume = function (m)
{if (m<=1) 1
else M.volume(m-2)*pi/(m+m)}
```

Con essa possiamo stampare sullo schermo i valori di v_m per $m \leq 20$ con poche istruzioni. Siccome R calcola con una precisione di circa 15 cifre significative (tra cui non contano però gli zeri iniziali) i valori sono rappresentati con 14 cifre dietro il punto decimale. L'ultima cifra è arrotondata, le altre cifre sono corrette come abbiamo verificato usando il sofisticato pacchetto aritmetico *pari*, disponibile gratuitamente.

```
for (m in 1:20)
{x=sprintf('%2d   %.14f',m,M.volume(m))
print(x)}
```

m	v_m
1	1.00000000000000
2	0.78539816339745
3	0.52359877559830
4	0.30842513753404
5	0.16449340668482
6	0.08074551218828
7	0.03691223414321
8	0.01585434424382
9	0.00644240020066
10	0.00249039457019
11	0.00091997259736
12	0.00032599188693
13	0.00011116073667
14	0.00003657620418
15	0.00001164072512
16	0.00000359086045
17	0.00000107560049
18	0.00000031336169
19	0.00000008892365
20	0.00000002461137

Vediamo che per $m = 10$ la palla occupa solo il 2.49 per mille del volume del cubo a cui è iscritta!

Quale vicinanza?

Dalla tabella a pagina 1 si vede che $v_{20} = 0.0000000246\ldots < 10^{-7}$. Se abbiamo quindi raccolto le concentrazioni nel sangue di 20 molecole rappresentate da numeri in $[0,1]$ di un milione di pazienti (un numero difficilmente raggiungibile nella realtà) e se volessimo considerare i dati x e y di due pazienti simili se $|x-y| < 0.5$ nella metrica euclidea di $\mathbb{R}_{20}$, la probabilità che per un punto x ce ne sia uno distinto e vicino (in questo senso) è solo circa 0.1 e quindi spesso questo concetto di vicinanza risulta poco utilizzabile.

I prodotti p_m e i quozienti q_m

Definizione 2.1. Per $m \geq 1$ denotiamo con p_m il prodotto che si ottiene con i fattori $m, m-2, m-4, \ldots$, andando avanti finché i fattori rimangono tutti positivi. Poniamo $p_0 := 1$. Quindi

m	p_m		
0	1	=	1
1	1	=	1
2	2	=	2
3	$3 \cdot 1$	=	3
4	$4 \cdot 2$	=	8
5	$5 \cdot 3 \cdot 1$	=	15
6	$6 \cdot 4 \cdot 2$	=	24
7	$7 \cdot 5 \cdot 3 \cdot 1$	=	105
8	$8 \cdot 6 \cdot 4 \cdot 2$	=	384
9	$9 \cdot 7 \cdot 5 \cdot 3 \cdot 1$	=	945
10	$10 \cdot 8 \cdot 6 \cdot 4 \cdot 2$	=	3840

È chiaro che

$$p_m = m p_{m-2}$$

per $m \geq 2$. Per m pari il numero dei fattori è quindi uguale a $\frac{m}{2}$ (se come al solito consideriamo 1 come prodotto di 0 fattori), per m dispari abbiamo $\frac{m+1}{2}$ fattori di cui però possiamo tralasciare l'1 finale senza cambiare il valore del prodotto che quindi per m dispari può essere anche rappresentato mediante $\frac{m-1}{2}$ fattori della forma indicata.

Definizione 2.2. Per $m \geq 1$ definiamo

$$q_m := \frac{p_{m-1}}{p_m}$$

mentre poniamo $q_0 := 1$.

m	q_m
0	1
1	1
2	$\frac{1}{2}$
3	$\frac{2}{3}$
4	$\frac{3 \cdot 1}{4 \cdot 2}$
5	$\frac{4 \cdot 2}{5 \cdot 3}$
6	$\frac{5 \cdot 3 \cdot 1}{6 \cdot 4 \cdot 2}$
7	$\frac{6 \cdot 4 \cdot 2}{7 \cdot 5 \cdot 3}$
8	$\frac{7 \cdot 5 \cdot 3 \cdot 1}{8 \cdot 6 \cdot 4 \cdot 2}$
9	$\frac{8 \cdot 6 \cdot 4 \cdot 2}{9 \cdot 7 \cdot 5 \cdot 3}$

Per ogni $m \geq 2$ abbiamo evidentemente le relazioni

$$q_m = \frac{m-1}{m} q_{m-2} \quad (*)$$

che possiamo usare per definire la funzione

```
Q = function (m)
{if (m<=1) 1
else Q(m-2)*(m-1)/m}
```

Con

```
for (m in 1:20)
{x=sprintf("%2d   %.5f",m,Q(m))
print(x)}
```

otteniamo adesso i valori

m	q_m
1	1.00000
2	0.50000
3	0.66667
4	0.37500
5	0.53333
6	0.31250
7	0.45714
8	0.27344
9	0.40635
10	0.24609
11	0.36941
12	0.22559
13	0.34099
14	0.20947
15	0.31826
16	0.19638
17	0.29954
18	0.18547
19	0.28377
20	0.17620

Dalle relazioni (*) è chiaro che $q_m < q_{m-2}$ per ogni $m \geq 2$ e che quindi $q_m \leq \frac{1}{2}$ per m pari ≥ 2 e $q_m \leq \frac{2}{3}$ per m dispari ≥ 3.

Osservazione 2.3. $q_m = \dfrac{1}{m q_{m-1}}$

per $m \geq 1$.

Dimostrazione. Per $m = 1$ abbiamo $\frac{1}{1 q_0} = 1 = q_1$, per $m \geq 2$ invece

$$q_m = \frac{p_{m-1}}{p_m} = \frac{p_{m-1}}{m p_{m-2}} = \frac{1}{m} \frac{1}{q_{m-1}}$$

Osservazione 2.4. *Per* $m \geq 1$ *vale*

$$\frac{q_m}{q_{m-1}} = m q_m^2 = \frac{1}{m q_{m-1}^2}$$

Dimostrazione. Per l'osservazione 2.3

$$\frac{q_m}{q_{m-1}} = \frac{m q_m}{m q_{m-1}} = m q_m^2 = m \frac{1}{m^2 q_{m-1}^2} = \frac{1}{m q_{m-1}^2}$$

Proposizione 2.5. *Abbiamo*

$$v_m = \begin{cases} \dfrac{1}{p_m} \left(\dfrac{\pi}{2}\right)^{\frac{m}{2}} & \text{per } m \text{ pari} \\ \dfrac{1}{p_m} \left(\dfrac{\pi}{2}\right)^{\frac{m-1}{2}} & \text{per } m \text{ dispari} \end{cases}$$

Dimostrazione. Induzione su m.

$\underline{m = 0}$: $\dfrac{1}{p_0} \left(\dfrac{\pi}{2}\right)^0 = 1 = v_0$.

$\underline{m = 1}$: $\dfrac{1}{p_1} \left(\dfrac{\pi}{2}\right)^0 = 1 = v_1$.

Sia $m \geq 2$ pari: In questo caso abbiamo

$$v_m = \frac{\pi}{2m} v_{m-2} \overset{\text{IND}}{=} \frac{\pi}{2m} \frac{1}{p_{m-2}} \left(\frac{\pi}{2}\right)^{\frac{m-2}{2}} = \frac{1}{m p_{m-2}} \left(\frac{\pi}{2}\right)^{\frac{m}{2}} = \frac{1}{p_m} \left(\frac{\pi}{2}\right)^{\frac{m}{2}}$$

Sia $m \geq 3$ dispari: Abbiamo

$$v_m = \frac{\pi}{2m} v_{m-2} \overset{\text{IND}}{=} \frac{\pi}{2m} \frac{1}{p_{m-2}} \left(\frac{\pi}{2}\right)^{\frac{m-3}{2}} = \frac{1}{m p_{m-2}} \left(\frac{\pi}{2}\right)^{\frac{m-1}{2}} = \frac{1}{p_m} \left(\frac{\pi}{2}\right)^{\frac{m-1}{2}}$$

Nota 2.6. Otteniamo la seguente tabella:

m	v_m		
2	$\frac{\pi}{2 \cdot 2}$	=	$\frac{\pi}{4}$
3	$\frac{\pi}{3 \cdot 2}$	=	$\frac{\pi}{6}$
4	$\frac{\pi^2}{8 \cdot 4}$	=	$\frac{\pi^2}{32}$
5	$\frac{\pi^2}{15 \cdot 4}$	=	$\frac{\pi^2}{60}$
6	$\frac{\pi^3}{24 \cdot 8}$	=	$\frac{\pi^3}{192}$
7	$\frac{\pi^3}{105 \cdot 8}$	=	$\frac{\pi^3}{840}$
8	$\frac{\pi^4}{384 \cdot 16}$	=	$\frac{\pi^4}{6144}$
9	$\frac{\pi^4}{945 \cdot 16}$	=	$\frac{\pi^4}{15120}$

Corollario 2.7. $v_m = \dfrac{1}{p_m} \left(\dfrac{\pi}{2}\right)^{[\frac{m}{2}]}$

Il prodotto di Wallis

Proposizione 2.8 (prodotto di Wallis).

$$\lim_{k \longrightarrow \infty} \frac{2^2 \cdot 4^2 \cdots (2k)^2}{1^2 \cdot 3^2 \cdots (2k-1)^2} \frac{1}{2k+1} = \frac{\pi}{2}$$

Dimostrazione. Corsi di analisi.

Corollario 2.9.

$$\lim_{k \longrightarrow \infty} (2k+1) q_{2k}^2 = \lim_{k \longrightarrow \infty} 2k q_{2k}^2 = \frac{2}{\pi}$$

$$\lim_{k \longrightarrow \infty} (2k+1) q_{2k+1}^2 = \lim_{k \longrightarrow \infty} 2k q_{2k+1}^2 = \frac{\pi}{2}$$

Dimostrazione. I fattori del prodotto nella proposizione 2.8 possono essere scritti nella forma

$$\frac{p_{2k}^2}{p_{2k-1}^2} \frac{1}{2k+1} = \frac{1}{(2k+1) q_{2k}^2}$$

da cui segue il primo enunciato, osservando che

$$\lim_{k \longrightarrow \infty} \frac{2k}{2k+1} = 1$$

Ponendo $m = 2k+1$ nell'osservazione 2.4 abbiamo

$$(2k+1) q_{2k+1}^2 = \frac{1}{(2k+1) q_{2k}^2}$$

e il secondo enunciato segue dal primo.

Stime per q_m

Corollario 3.1. $\lim_{m \longrightarrow \infty} q_m = 0$.

Lemma 3.2. *Per ogni m abbiamo*

$$(m+1)q_m^2 < (m-1)q_{m-2}^2$$
$$mq_m^2 > (m-2)q_{m-2}^2$$

Dimostrazione. Per la relazione (*) nella definizione 2.2 abbiamo

$$(m+1)q_m^2 = (m+1)\left(\frac{m-1}{m}\right)^2 q_{m-2}^2 = \frac{(m+1)(m-1)}{m^2}(m-1)q_{m-2}^2$$

con

$$\frac{(m+1)(m-1)}{m^2} = \frac{m^2-1}{m^2} < 1$$

Similmente

$$mq_m^2 = m\left(\frac{m-1}{m}\right)^2 q_{m-2}^2 = \frac{(m-1)^2}{m}q_{m-2}^2 = \frac{(m-1)^2}{m(m-2)}(m-2)q_{m-2}^2$$

con

$$\frac{(m-1)^2}{m(m-2)} = \frac{m^2-2m+1}{m^2-2m} > 1$$

Corollario 3.3. *Le successioni*

$$\bigcirc_{m\in 2\mathbb{N}} (m+1)q_m^2 \quad e \quad \bigcirc_{m\in 2\mathbb{N}+1} (m+1)q_m^2$$

sono strettamente decrescenti, mentre le successioni

$$\bigcirc_{m\in 2\mathbb{N}} mq_m^2 \quad e \quad \bigcirc_{m\in 2\mathbb{N}+1} mq_m^2$$

sono strettamente crescenti.

Proposizione 3.4. *Valgono le inclusioni*

$$m \text{ pari} \implies mq_m^2 < \frac{2}{\pi} < (m+1)q_m^2$$

$$m \text{ dispari} \implies mq_m^2 < \frac{\pi}{2} < (m+1)q_m^2$$

Dimostrazione. Usiamo il corollario 2.9 e il corollario 3.3.

Le successioni

$$\bigcirc_{m\in 2\mathbb{N}} mq_m^2 \quad \text{e} \quad \bigcirc_{m\in 2\mathbb{N}} (m+1)q_m^2$$

convergono a $\frac{2}{\pi}$; la prima è strettamente crescente, la seconda strettamente decrescente. Ciò implica il primo enunciato.

Le successioni

$$\bigcirc_{m\in 2\mathbb{N}+1} mq_m^2 \quad \text{e} \quad \bigcirc_{m\in 2\mathbb{N}+1} (m+1)q_m^2$$

convergono a $\frac{\pi}{2}$; la prima è strettamente crescente, la seconda strettamente decrescente. Ciò implica il secondo enunciato.

Corollario 3.5. *Sia $m \geq 1$. Allora:*

$$m \text{ pari} \implies \frac{2}{\pi}\frac{1}{m+1} < q_m^2 < \frac{2}{\pi}\frac{1}{m}$$

$$m \text{ dispari} \implies \frac{\pi}{2}\frac{1}{m+1} < q_m^2 < \frac{\pi}{2}\frac{1}{m}$$

Nota 3.6. Nei corsi di Analisi spesso si deduce la proposizione 3.4 da rappresentazioni degli integrali $\int_0^{\frac{\pi}{2}} \sin^m x\,dx$ che vengono calcolati con integrazione per parti, ottenendo poi da essa la formula di Wallis (proposizione 2.8) che noi abbiamo invece assunto come nota.

Il quoziente $\frac{v_m}{v_{m-1}}$

Proposizione 3.7. *Per $m \geq 1$ si ha*

$$\frac{v_m}{v_{m-1}} = \begin{cases} \frac{\pi}{2}q_m & \text{per } m \text{ pari} \\ q_m & \text{per } m \text{ dispari} \end{cases}$$

Dimostrazione. Usiamo la proposizione 2.5. Per m pari abbiamo

$$v_m = \frac{1}{p_m}\left(\frac{\pi}{2}\right)^{\frac{m}{2}}$$

e

$$v_{m-1} = \frac{1}{p_{m-1}}\left(\frac{\pi}{2}\right)^{\frac{m-2}{2}}$$

quindi

$$\frac{v_m}{v_{m-1}} = \frac{p_{m-1}}{p_m}\frac{\pi}{2} = \frac{\pi}{2}q_m$$

Per m dispari abbiamo

$$v_m = \frac{1}{p_m}\left(\frac{\pi}{2}\right)^{\frac{m-1}{2}}$$

e

$$v_{m-1} = \frac{1}{p_{m-1}}\left(\frac{\pi}{2}\right)^{\frac{m-1}{2}}$$

perciò

$$\frac{v_m}{v_{m-1}} = \frac{p_{m-1}}{p_m} = q_m$$

Corollario 3.8. $\frac{v_m}{v_{m-1}} < 1$ *per ogni $m \geq 2$.*

Dimostrazione. Sia m pari e perciò $m \geq 2$. Come osservato alla fine della definizione 2.2 abbiamo allora $q_m \leq \frac{1}{2}$, per cui

$$\frac{v_m}{v_{m-1}} = \frac{\pi}{2}q_m \leq \frac{\pi}{4} < 1$$

Sia m dispari. Allora $m \geq 3$ e

$$\frac{v_m}{v_{m-1}} = q_m \leq \frac{2}{3} < 1$$

Calcoliamo i rapporti $\frac{v_m}{v_{m-1}}$ dalla proposizione 3.7 (invece che dalla proposizione 1.6) usando la funzione `Q` della definizione 2.2. Le istruzioni che seguono prducono un output che può essere facilmente inserito in un file Latex con cui otteniamo poi la tabella.

```
for (m in 1:9)
{x=Q(m)
if (m%%2==0) x=pi*x/2
a=paste('v_',m,' = ',sprintf('%.3f',x),
  ' v_{',m-1,'}\\',sep='')
print(a)}
```

Abbiamo dovuto chiamare in aiuto la funzione `paste` di R a causa di qualche piccolo difetto in `sprintf`. Entrambe le funzioni verranno spiegate nel corso di Fondamenti di informatica.

$$\begin{aligned} v_1 &= 1.000 v_0 \\ v_2 &= 0.785 v_1 \\ v_3 &= 0.667 v_2 \\ v_4 &= 0.589 v_3 \\ v_5 &= 0.533 v_4 \\ v_6 &= 0.491 v_5 \\ v_7 &= 0.457 v_6 \\ v_8 &= 0.430 v_7 \\ v_9 &= 0.406 v_8 \end{aligned}$$

Vogliamo adesso dimostrare che i fattori $\frac{v_m}{v_{m-1}}$ in questa tabella formano una successione strettamente decrescente che converge a zero.

Osservazione 3.9. *Sia $m \geq 2$. Allora*

$$\frac{\frac{v_m}{v_{m-1}}}{\frac{v_{m-1}}{v_{m-2}}} = \begin{cases} \frac{\pi}{2}mq_m^2 & \text{per } m \text{ pari} \\ \frac{2}{\pi}mq_m^2 & \text{per } m \text{ dispari} \end{cases}$$

Dimostrazione. (1) Sia m pari. Per la proposizione 3.7 abbiamo allora

$$\frac{v_m}{v_{m-1}} = \frac{\pi}{2}q_m$$

e

$$\frac{v_{m-1}}{v_{m-2}} = q_{m-1}$$

Il quoziente è quindi uguale a

$$\frac{\pi}{2}\frac{q_m}{q_{m-1}} = \frac{\pi}{2}mq_m^2$$

usando l'osservazione 2.4.

(2) Sia m dispari. In questo caso abbiamo

$$\frac{v_m}{v_{m-1}} = q_m$$

e

$$\frac{v_{m-1}}{v_{m-2}} = \frac{\pi}{2}q_{m-1}$$

Il quoziente è perciò uguale a

$$\frac{2}{\pi}\frac{q_m}{q_{m-1}} = \frac{2}{\pi}mq_m^2$$

Corollario 3.10. *La successione* $\bigcirc_m \frac{v_m}{v_{m-1}}$ *è strettamente decrescente e converge a zero.*

Dimostrazione. Siano $m \geq 2$ e

$$q'_m := \frac{\frac{v_m}{v_{m-1}}}{\frac{v_{m-1}}{v_{m-2}}}$$

Se m è pari, allora

$$q'_m = \frac{\pi}{2}mq_m^2 < \frac{\pi}{2}m\frac{2}{\pi}\frac{1}{m} = 1$$

se m è dispari, allora

$$q'_m = \frac{2}{\pi}mq_m^2 < \frac{2}{\pi}m\frac{\pi}{2}\frac{1}{m} = 1$$

Ciò mostra che la successione è strettamente decrescente.

Dalla proposizione 3.7 sappiamo che

$$\frac{v_m}{v_{m-1}} = \begin{cases} \frac{\pi}{2}q_m & \text{per } m \text{ pari} \\ q_m & \text{per } m \text{ dispari} \end{cases}$$

Per il corollario $\lim_{m\longrightarrow\infty} q_m = 0$ e quindi anche $\lim_{m\longrightarrow\infty} \frac{v_m}{v_{m-1}} = 0$.

Il volume angolare $\hat{v}_m$

Definizione 4.1. Il cubo m-dimensionale possiede 2^m vertici. Se togliamo la palla iscritta dal cubo, rimangono 2^m regioni che chiamiamo le *regioni angolari* del cubo, ciascuna delle quali ha un volume uguale a

$$\hat{v}_m := \frac{1 - v_m}{2^m}$$

Chiamiamo $\hat{v}_m$ il *volume angolare* m-dimensionale.

Proposizione 4.2. $\lim_{m \longrightarrow \infty} \hat{v}_m = 0$

e, più sorprendentemente,

$$\lim_{m \longrightarrow \infty} \frac{\hat{v}_m}{v_m} = \infty$$

Dimostrazione. Siccome $0 \leq 1 - v_m < 1$, è chiaro che

$$\hat{v}_m = \frac{1 - v_m}{2^m}$$

converge a zero. Invece

$$\frac{\hat{v}_m}{v_m} = \frac{1 - v_m}{2^m v_m} = \frac{1}{2^m v_m} - \frac{1}{2^m}$$

Siccome $\lim_{m \longrightarrow \infty} \frac{1}{2^m} = 0$, è sufficiente dimostrare che $\lim_{m \longrightarrow \infty} \frac{1}{2^m v_m} = \infty$.

Ma $2^m v_m = v_m(1)$ e sappiamo dal corollario 1.7 che $\lim_{m \longrightarrow \infty} v_m(1) = 0$.

Creiamo un programma in R per calcolare $\hat{v}_m$ e $\frac{\hat{v}_m}{v_m}$ per $1 \leq m \leq 30$.

```
for (m in 1:30)
{v=M.volume(m)
x=(1-v)/2^m
a=sprintf('%2d %.8f %12.3f',m,x,x/v)
print(a)}
```

Otteniamo la tabella

m	$\hat{v}_m$	$\hat{v}_m / v_m$
1	0.00000000	0.000
2	0.05365046	0.068
3	0.05955015	0.114
4	0.04322343	0.140
5	0.02610958	0.159
6	0.01436335	0.178
7	0.00752412	0.204
8	0.00384432	0.242
9	0.00194054	0.301
10	0.00097413	0.391
11	0.00048783	0.530
12	0.00024406	0.749
13	0.00012206	1.098
14	0.00006103	1.669
15	0.00003052	2.622
16	0.00001526	4.249
17	0.00000763	7.093
18	0.00000381	12.173
19	0.00000191	21.449
20	0.00000095	38.749
21	0.00000048	71.689
22	0.00000024	135.677
23	0.00000012	262.422
24	0.00000006	518.249
25	0.00000003	1044.144
26	0.00000001	2144.529
27	0.00000001	4486.878
28	0.00000000	9556.748
29	0.00000000	20709.155
30	0.00000000	45630.111

che mostra che per $m \leq 12$ la palla iscritta al cubo ha ancora un volume superiore a quello di ciascuna regione angolare, ma che per $m \geq 13$ il volume angolare supera il volume della palla iscritta. Per questa ragione spesso si paragona il cubo m-dimensionale a un riccio con 2^m aculei, ciascuno dei quali a partire da $m = 13$ ha un volume maggiore di quello del corpo centrale del riccio. Come la tabella mostra, per $m = 20$ ciascuno dei 2^{20} aculei (circa un milione) ha un volume più di 38 volte maggiore di quello del corpo centrale (la palla iscritta) e per $m = 30$ ciascuno dei 2^{30} aculei (più di un miliardo) ha un volume più di 45000 volte maggiore di quello del corpo centrale.

Questa geometria complicata rende molto difficile la statistica in alte dimensioni. Anche se molti algoritmi forniscono risultati in ogni dimensione, l'interpretazione di quanto il software ha calcolato deve essere fatta con molta prudenza e spesso l'unica strada veramente proponibile è la riduzione della dimensione tramite una ragionata proiezione $\mathbb{R}_m \longrightarrow \mathbb{R}_l$ con ad esempio $l \leq 5$.

Istogrammi multidimensionali

Assumiamo di nuovo che i nostri dati siano punti nel cubo unitario di $\mathbb{R}_m$. Data una decomposizione

$$[0,1]_m = A_1 \dot{\cup} \ldots \dot{\cup} A_k$$

possiamo definire, per $1 \leq j \leq k$, il numero n_j come il numero di quei punti tra i punti dati che si trovano in A_j. La successione finita $(n_1, \ldots, n_k)$ può essere considerata come un *istogramma* m-dimensionale. Ma anche qui in alta dimensione incontriamo delle difficoltà: Assumiamo che gli insiemi A_j siano ottenuti dividendo ogni lato di $[0,1]_m$ in 10 intervalli uguali e formando i sottocubi (semiaperti affinché siano disgiunti, ma non è importante) corrispondenti. Allora per $m = 2$ avremo 100 sottocubi (cioè $k = 100$), per $m = 10$ invece abbiamo $k = 10^{10}$ sottocubi. Da ciò segue però che 10^6 punti casuali in $[0,1]_{10}$ saranno molto meno densi che 4 punti in $[0,1]_2$. Infatti nel piano ogni sottocubo conterrà in media $4 \cdot 10^{-2} = 0.04$ punti, in $[0,1]_{10}$ invece in media $10^{6-10} = 10^{-4} = 0.0001$. In altre parole i 4 punti in $[0,1]_2$ sono 400 volte più densi del milione di punti in $[0,1]_{10}$ benché, come già osservato, ad esempio nella statistica medica, è molto difficile raccogliere i dati di un milione di pazienti che però, invece, contengono anche 20 o 40 parametri accentuando ancora di molto la difficoltà che abbiamo illustrato.

A ciò si aggiunge il problema di memorizzare un istogramma che consiste di 10^{10}, 10^{20} o 10^{40} numeri n_j.

Anche in questo caso si cercherà di ridurre la dimensione. Un altro approccio, matematicamente molto difficile e interessante, è quello di trovare decomposizioni con un numero k relativamente basso di insiemi A_j in qualche modo ben distribuiti nel cubo $[0,1]_m$.

La diagonale

Mentre il raggio della palla iscritta al cubo unitario in $\mathbb{R}_m$ è sempre uguale a $\frac{1}{2}$, il diametro del cubo, cioè la lunghezza della diagonale tra l'origine e il punto $(1, \ldots, 1)$, è uguale a $\sqrt{m}$.

Ciò implica che la palla, pur toccando il bordo del cubo (nei centri dei sottocubi di dimensione $m - 1$), dista invece dai vertici di $\frac{\sqrt{m} - 1}{2}$; siccome questa distanza diventa sempre più grande, ciò conferma l'impressione che il cubo m-dimensionale al crescere di m assomiglia sempre di più a un riccio con corpo sferico sempre più piccolo e aculei sempre più lunghi.

La statistica del futuro

„The coming century is surely the century of data. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently. Hyperspectral imagery, Internet portals, financial tick-by-tick data, and DNA microarrays are just a few of the better-known sources, feeding data in torrential streams into scientific and business databases ...

Classical methods are simply not designed to cope with this kind of explosive growth of dimensionality of the observation vector. We can say with complete confidence that in the coming century high-dimensional data analysis will be a very significant activity, and completely new methods of high-dimensional data analysis will be developed; we just don't know what they are.“ (David Donoho)

„È un momento particolarmente felice per la biostatistica in generale e per la statistica clinica in particolare. Gli sviluppi della biologia molecolare e della medicina stanno producendo enormi quantità di dati che devono essere ordinati e interpretati, creando così una domanda di competenze statistiche mai vista in precedenza.

Gli statistici clinici, cioè gli statistici che lavorano nella ricerca medica su umani, partecipano al clima di euforia grazie anche alla crescente disponibilità di risorse della ricerca medica, alla sua crescente matematizzazione e, con riferimento all'industria farmaceutica, grazie ad un ventennio di impressionante sviluppo. Per doveri legali e per tradizione culturale, l'industria farmaceutica è uno dei pochi settori produttivi che offre agli statistici la possibilità di una carriera non accademica di alto profilo scientifico. Insieme ai centri di cura e ricerca medica pubblici e privati, l'industria farmaceutica partecipa così attivamente alla richiesta e alla produzione di metodologia statistica.“ (Mauro Gasparini)

Basi di dati

La struttura complessa e sorprendente degli spazi ad alta dimensione crea difficoltà non solo in statistica, ma ad esempio anche negli *algoritmi di ricerca* in grandi insiemi di dati (basi di dati in medicina, nell'industria, in geografia, in biologia molecolare) che spesso vengono rappresentati (mediante tecniche sofisticate di trasformazione) come punti di qualche $\mathbb{R}_m$ ad alta e talvolta altissima dimensione. Gli algoritmi di ricerca classici spesso utilizzano concetti di *somiglianza* basati ad esempio sulla vicinanza nella metrica euclidea che però in questi spazi ad alta dimensione perde gran parte del suo significato. Superare questa difficiltà è uno dei compiti più attuali e più interessanti studiati dalla teoria delle basi di dati.

Il problema del guscio

X sia un sottoinsieme misurabile di misura $v(X) < \infty$ in $\mathbb{R}_m$ e $0 \leq \alpha < 1$. Allora $v(\alpha X) = \alpha^m v(X)$. X sia *stellato* rispetto all'origine e quindi $\alpha X \subset X$. Per il volume del guscio $X \setminus \alpha X$ si ha allora

$$v(X \setminus \alpha X = (1 - \alpha^m) v(X)$$

e quindi

$$\frac{v(X \setminus \alpha X)}{v(X)} = 1 - \alpha^m$$

Siccome $\alpha < 1$ questo rapporto tende a 1; ciò significa che il guscio occupa, con il crescere di m, un volume relativo sempre maggiore.

Questo fenomeno è importante in statistica perché implica che in alta dimensione la maggior parte di una popolazione casuale si troverà in posizioni marginali dello spazio dei dati venendo così meno quanto si osserva nella stastica univariata in cui i valori di una popolazione *normale* si concentrano nella vicinanza del valore medio.

Il paradosso delle pareti

Già in dimensione 5 si verifica un fenomeno molto sorprendente impossibile in dimensioni ≤ 3 e probabilmente anche in dimensione 4. Troviamo infatti adesso una sfera in $\mathbb{R}_5$ che interseca tutti i lati 4-dimensionali del cubo unitario, ma non contiene il centro del cubo! Procediamo in questo modo:

Il centro del cubo è

$$c := (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$$

Il punto

$$p := (\alpha, \alpha, \alpha, \alpha, \alpha)$$

con $\alpha = 0.08$ appartiene anch'esso al cubo. Inoltre

$$|p - c|^2 = 5 \cdot (\alpha - \frac{1}{2})^2 = 5 \cdot 0.42^2 = 0.882$$

Se scegliamo il raggio $\rho = 0.93$, allora $\rho^2 = 0.8649$, per cui il centro c non appartiene alla palla di raggio ρ attorno a p. Facciamo adesso vedere che la sfera di raggio ρ interseca ogni lato 4-dimensionale del cubo.

Un tale lato è dato dall'intersezione del cubo con un iperpiano dato da un'equazione della forma $x_j = 0$ oppure $x_j = 1$. Per simmetria possiamo assumere che $j = 1$.

È suffciente dimostrare che esistono $\beta, \gamma \in [0, 1]$ tali che i punti

$$q_1 := (0, \beta, \beta, \alpha, \alpha)$$

e

$$q_2 := (1, \gamma, \alpha, \alpha, \alpha)$$

hanno distanza ρ da p. Abbiamo

$$|q_1 - p|^2 = \alpha^2 + 2(\beta - \alpha)^2$$

e

$$|q_2 - p|^2 = (1 - \alpha)^2 + (\gamma - \alpha)^2$$

(1) Per q dobbiamo soddisfare l'equazione

$$\rho^2 = \alpha + 2(\beta - \alpha)^2$$

cioè

$$\frac{\rho^2 - \alpha^2}{2} = (\beta - \alpha)^2$$

Ma

$$\frac{\rho^2 - \alpha^2}{2} = 0.42925$$

quindi bisogna avere

$$|\beta - \alpha| = \sqrt{0.42925}$$

per cui possiamo porre

$$\beta = \sqrt{0.42925} + \alpha$$
$$= 0.65517 \ldots + 0.08 = 0.73517 \ldots$$

Abbiamo quindi $\beta \in [0, 1]$.

(2) Per q_2 dobbiamo avere

$$\rho^2 = (1 - \alpha)^2 + (\gamma + \alpha)^2$$

cioè

$$\rho^2 - (1 - \alpha)^2 = (\gamma - \alpha)^2$$

Ma

$$\rho^2 - (1 - \alpha)^2 = 0.8649 - 0.92^2$$
$$= 0.8649 - 0.8464 = 0.0185$$

e quindi bisogna avere

$$|\gamma - \alpha| = \sqrt{0.0185}$$

per cui possiamo porre

$$\gamma = \sqrt{0.0185} + \alpha = 0.21601 \ldots$$

Anche $\gamma \in [0, 1]$.

Modificato da Böhm/, pag. 6, in cui si dà un esempio per $m = 16$. Abbiamo chiamato questo esempio il paradosso delle pareti, perché per convincersi della stranezza dell'enunciato è sufficiente immaginare che una sfera possa intersecare tutte le pareti di una stanza cubica senza contenere il punto centrale della stanza.

Questi fenomeni creano molti problemi nell'interpretazione statistica e nello sviluppo degli algoritmi in alte dimensioni.

Il paradosso della sfera centrale

Consideriamo un cubo stavolta con centro nell'origine di $\mathbb{R}_m$ e di lato 4. Nei 2^m punti della forma $(\pm 1, \ldots, \pm 1)$ in cui i segni $+$ o $-$ vengono scelti in tutte i modi possibili, poniamo una sfera di raggio 1 con centro in quel punto. Consideriamo poi la sfera con centro nell'origine tangente a tutte quelle altre sfere. Il suo raggio sia ρ. La situazione è illustrata per $m = 2$ dalla figura in alto nella colonna accanto.

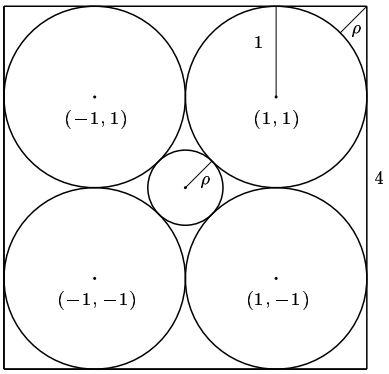
Il mezzo diametro del cubo è uguale a $2\sqrt{m}$ e anche uguale a $2\rho + 2$, abbiamo quindi

$$1 + \rho = \sqrt{m}$$

per cui

$$\rho = \sqrt{m} - 1$$

Ciò significa che per $m = 9$ la sfera interna tocca il bordo del cubo e per $m \geq 10$ esce addirittura da esse, benché le altre sfere rimangano naturalmente tutte contenute nel cubo. Da Gentle, pag. 297.



Proiezioni ottimali

La teoria delle proiezioni ottimali (che nella letteratura inglese appare sotto il nome di *projection pursuit*) è stata iniziata da Friedman e Tukey. Si cercano proiezioni ottimali rispetto a una funzione (*indice*) di rilevanza che può essere scelta in vari modi. Questo metodo interessante, piuttosto impegnativo nel calcolo, che, almeno nelle intenzioni, permette di superare le difficoltà delle alte dimensioni e che contiene come casi speciali e in un certo senso migliora molti metodi classici della statistica multivariata (come l'analisi delle componenti principali e l'analisi delle discriminanti) è esposto in un famoso articolo di Peter Huber e nella tesi di Guy Nason. Il pacchetto *XGobi* di R contiene funzioni per questa tecnica.

www.stats.bris.ac.uk/ guy/Research/PP/PP.html

www.ggobi.org

Bibliografia

15561 **C. Böhm/S. Berchtold/D. Keim:** Searching in high-dimensional spaces - index structures for improving the performance of multimedia databases.
Internet ca. 2001, 74p.

16486 **D. Donoho:** High-dimensional data analysis - the curses and blessings of dimensionality. Internet 2000, 32p.

15981 **M. Gasparini:** La statistica nelle prove cliniche. Boll. UMI Mat. Soc. Cult. 6/A (2003), 119-140.

16041 **J. Gentle:** Elements of computational statistics. Springer 2002.

17077 **P. Huber:** Projection pursuit. Ann. Statistics 13 (1985), 435-475.

17081 **G. Nason:** Design and choice of projection indices. PhD thesis Bath Univ. 1992.

(1300/2) **R. Remmert:** Classical topics in complex function theory. Springer 1998.