Teoria ed applicazioni degli intervalli di tolleranza multivariati(*)

Agostino Tarsitano
Università degli studi della Calabria
Dipartimento di economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Riassunto
Il lavoro discute gli intervalli di tolleranza multivariati proponendoli come tecnica di classificazione in tre gruppi: le unità che possiedono un dato carattere multidimensionale in misura -carente", quelli che ne sono “eccedenti" e quelle "normali". La tecnica, sviluppata in ambito medico, è suggerita a riguardo della solvibilità degli affidati di banche e intermediari finanziari. In particolare si studia la possibilità di ridurre le sofferenze di tali società orientandole a procedure sbrigative di esclusione (di concessione) per le unità carenti (eccedenti) ed a concentrare gli sforzi deqli uffici istruttori sui clienti "normali".

Keywords:
classificazione, credit scoring

(*) Lavoro comparso negli Atti delle giornate di studio “Classificazione e analisi dei dati: Metodi-Software-Applicazioni, Pescara 11-12 ottobre, 1990. pp. 253-264. Marino Solfanelli editore, Chieti

1. Introduzione

Il lavoro affronta il seguente problema: le unità di una popolazione statistica avente funzione di distribuzione appartenente alla famiglia parametrica $F(\theta):\theta \in \Theta \supseteq R^m$ posseggono, in varia misura, un certo carattere n-dimensionale *X*. Quelle che ne posseggono una quantità compresa entro dati limiti sono considerate "normali"; quelle che ne possiedono quantità più piccole del limite inferiore sono "carenti" e quelle che ne possiedono quantità più grandi del limite superiore sono "eccedenti". Occorre determinare, sulla base delle osservazioni campionarie, un intervallo multivariato (regione di tolleranza) che contenga non meno di una frazione *p* ritenuta "normale" della popolazione con un livello di confidenza pari ad una soglia prefissata *g*. In breve, si vogliono determinare dei livelli di guardia per il carattere in modo da poter prevedere (*prediction interval* è infatti un altro termine con cui sono noti tali limiti) con quale probabilità le prossime osservazioni estratte dalla popolazione ricadranno nell'intervallo di normalità ovvero risulteranno eccedenti o carenti.

A questa formulazione possono essere ricondotti diversi problemi. Quello su cui si è centrato il presente lavoro riguarda la possibilità di ridurre il volume delle sofferenze di banche ed altri intermediari finanziari che quotidianamente esaminano numerose pratiche di concessione fidi da parte di aziende o di finanziamento di consumi da parte di privati, Sarebbe certamente utile una tecnica automatica efficace e tempestiva che almeno distingua i clienti quasi sicuramente insolventi o quasi sicuramente solvibili dai clienti normali sui quali concentrare sforzi ed attenzioni del personale degli uffici istruttori. Nel lavoro si mostrerà come la teoria degli intervalli di tolleranza con particolare riguardo alla distribuzione multinormale possa portare ad una definizione di gruppi accurata ed operativa che, senza intaccare le capacità di valutazione qualitativa di chi decide l'erogazione del finanziamento, si offre come utile sintesi delle informazioni quantitative sul potenziale creditore.

Il lavoro è così organizzato. Nel prossimo paragrafo si discuteranno gli intervalli di tolleranza nel caso unidimensionale ed in quello successivo le regioni di tolleranza multivariata. Nel quarto paragrafo si approfondirà l'applicazione di questa tecnica alla previsione delle insolvenze.

2. Gli intervalli di tolleranza unidimensionali

In questo paragrafo discuteremo la teoria deqli intervalli di tolleranza e dei principali problemi che si incontrano nella loro applicazione per il caso univariato, In particolare si vedrà la distinzione tra il caso *dlstribution-free* in cui le assunzioni concernenti la funzione di distribuzione sono minimali ed il caso parametrico dove la funzione di distribuzione è specificata fatta eccezione per un numero finito di parametri in essa coinvolti. In entrambi i casi si è alla ricerca di valori campionari che racchiudano, con una prefissata probabilità, almeno una data porzione dei valori della popolazione di origine.

Approccio distribution-free

E' la situazione in cui sulla distribuzione congiunta delle (Xi) non si fa alcuna assunzione tranne forse la continuità. Fraser (1953) e Kemperman (1956) fanno una eccellente esposizione di questo approccio che qui si riprende brevemente.

Si supponga che in base ad un campione di n osservazioni da una funzione di densità $f(x|\theta)$ si vogliano determinare i quantili L_i ed L_u in modo da asserire che, con probabilità del 99%, almeno il 90% della popolazione d'origine appartiene all'intervallo $L(\gamma,\beta)=(L_i, L_u)$. Allora

$$\Pr\left[\int_{L_i}^{L_o} f(x|\theta)dx \geq 0.90\right] = 0.99 \tag{1}$$

(il segno di uguaglianza va sostituito dal segno di ≥ nel caso in cui la $f(x|\theta)$ sia discreta).

Questo tipo di esigenza si differenzia dall'idea degli intervalli di confidenza in quanto non si tratta di includere, con probabilità prefissata, il valore teorico di un qualche parametro, ma di delimitare la regione, in questo caso dell'asse reale, che contiene una frazione β o maggiore di unità della popolazione.

Siano ora $X_{(l)}, X_{(2)}, ..., X_{(n)}$ le osservazioni campionarie disposte in ordine di grandezza e si ponga $L_i=X_{(i)}$, $L_u=X_{(u)}$. La (1) può essere scritta come segue

$$\Pr\left\{F\left(X_{(u)}\right) - F\left(X_{(i)}\right) \geq \beta\right\} = \gamma \tag{2}$$

E' possibile dimostrare (Kendall e Stuart, 1979, pp. 547-550) che la probabilità (2) può essere espressa in termini della funzione Beta incompleta:

$$\gamma = IB(1-\beta, n-o+i+1, u-i) = \sum_{j=0}^{u-i+1} \binom{n}{j} (1-\beta)^j \beta^{n-j} \tag{3}$$

Inoltre, l'intervallo risultante è indipendente dalla particolare funzione di densità considerata. Fissata la frazione β e la probabilità γ la (3) ha ancora tre valori indeterminati: l'ampiezza campionaria e le posizioni *u* ed *i*; tale indeterminatezza si elimina in genere scegliendo *u=n, i=l* e risolvendo rispetto ad *n* la risultante relazione. Nella tabella l sono stati riportati, per alcuni valori di γ e di β le soluzioni della (3) calcolate adoperando l' algoritmo proposto da Brooker e Selby (1975).

Tabella 1: ampiezze campionarie

	Valori di β			
Valori di γ	90%	95%	99%	99.9%
90.0%	38	46	64	89
95.0%	77	93	130	181
99.0%	388	473	662	920
99.9%	3889	4746	6336	9230

Ad esempio l'entrata *n=473* significa che è necessario un campione di 473 estrazioni indipendenti perché gli estremi campionari X_{min} ed X_{max} comprendano il 99% della popolazione (e non necessariamente del campione) con probabilità del 95%.

Approccio parametrico

E' naturale che, nel caso si abbiano più dettagliate informazioni sulla funzione di distribuzione d'origine le ampiezza della tabella l si possano ridurre oppure si possa pervenire alla definizione più efficiente e precisa degli estremi dell'intervallo. Sia ad esempio *X* una variabile casuale gaussiana con media μ e varianza σ^2 entrambe incognite e siano *m* e s^2 le loro usuali stime basate su di un campione di *n* osservazioni indipendenti. Per determinare i limiti dell'intervallo di tolleranza che con probabilità γ includa almeno il β% della popolazione si pone

$$\Pr\left\{\Phi(L_u) - \Phi(L_i) \geq \beta\right\} = \gamma \tag{4}$$

dove $\Phi(x)$ è la funzione che associa all'argomento le aree sottese alla funzione di densità gaussiana nell'intervallo $]-\infty,x]$. La soluzione per gli estremi L_u ed L_i è stata data da Wald e Wolfowitz (1946)

$$\begin{cases} L_1(\gamma,\beta) = m - zs \\ L_2(\gamma,\beta) = m + zs \end{cases} \qquad con \qquad z = \tau_{n,\beta} = \left[\sqrt{\frac{n-1}{\chi^2_{n-1,\gamma}}}\right] \tag{5}$$

dove $\chi^2_{n-1,\gamma}$ è il quantile di ordine $(1-\gamma)$ della chi quadro con $(n-1)$ gradi di libertà e $\tau_{n,\beta}$ scaturisce dalla soluzione dell'equazione:

$$\Phi\left(\frac{1}{\sqrt{n}} + \tau_{n,\beta}\right) - \Phi\left(\frac{1}{\sqrt{n}} - \tau_{n,\beta}\right) = \beta \tag{6}$$

Nella tabella 2 vengono forniti i valori di soglia per β=95% combinazione di *n* e di γ.
e per varie combinazioni di *n* e γ.

Tabella 2: soglia *z* per intervalli gaussiani di contenuto β=95%

	Valori di β			
ampiezza n	90%	95%	99%	99.9%
25	2.4738	0.6313	2.9759	3.1213
50	2.3834	2.3787	2.5766	2.656
100	2.1716	2.2328	2.3557	2.4036
200	2.1019	2.1425	2.2225	2.253

Ad esempio, per γ=95%,n=25 si ottiene il valore *z=2.9759* per cui l'intervallo *L(0.99,0.95)=m± t2.9759s* contiene almeno il 95% dei valori della popolazione con probabilità del 99%. Da osservare che, come era prevedibile, il coefficiente *z* aumenta all'aumentare del livello di confidenza e diminuisce all'aumentare di *n*. In aggiunta, *z* non dipende né da *m* né da *s*.

3. Le regioni di tolleranza

In questo paragrafo gli intervalli di tolleranza verranno estesi ad un carattere multidimensionale $X=\{X_1, X_2, ..., X_k\}$. Svilupperemo però solo l'approccio parametrico e segnatamente si ipotizzerà per X una distribuzione gaussiana multivariata. Vedremo che non sarà possibile, come nel caso univariato, fornire soluzioni esatte e che la qualità dei risultati dipenderà dall'ampiezza campionaria e dalla struttura di varianze-covarianze cui danno luogo le osservazioni su X.

Il problema di determinare i livelli di guardia per un carattere multivariato può essere proposto nei seguenti termini: si dispone di un campione di n osservazioni indipendenti $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$ con $n>k$ provenienti da una distribuzione congiunta $f(\boldsymbol{x}|\theta)$. Si cercano le $L_j^i(\boldsymbol{x})$ e $L_j^u(\boldsymbol{x})$ tali che

$$Pr\left[\int_{L_1^i}^{L_1^u}\int_{L_2^i}^{L_2^u}\cdots\int_{L_k^i}^{L_k^u} f(\boldsymbol{x}|\theta)dX_1dX_2\ldots dX_k\right]=\gamma \qquad (7)$$

ovvero la probabilità inclusa nel volume delimitato dalle statistiche $L_i(\boldsymbol{x})$ ed $L_u(\boldsymbol{x})$ è almeno b e questa asserzione ha probabilità γ di essere vera.

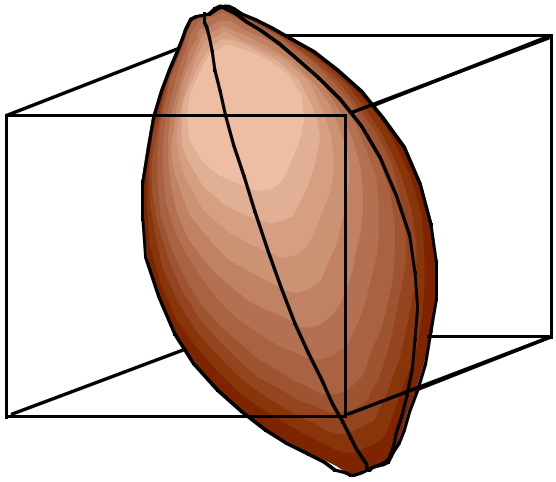

Fig.1: esempio di regioni di tolleranza

La espressione (7) non consente di definire regioni di tolleranza di forma conica: le quantità $L_i(\boldsymbol{x})$ ed $L_u(\boldsymbol{x})$ determinano comunque dei parallelepipedi che sezionano in modo regolare il volume di probabilità. Infatti, la prima generalizzazione multivariata degli intervalli di tolleranza che viene in mente (un'idea che del resto si ritrova nelle prime proposte fatte da Wilks (1942) e da Wald 1943c) è di determinare un intervallo unidimensionale $L_j(\gamma;\beta)$ per ciascuna variabile componente. L'intersezione delle varie bande dovrebbe delimitare la regione di tolleranza in cui ricadono le unita normali. In altre parole si dirà normale l'osservazione X che si colloca nell'ipercubo definito dalle disuguaglianze

$$L_j^i \le X \le L_j^u \quad per \quad j=1,2,\ldots,k \qquad (8)$$

Nella tabella 3 sono riportati i dati pubblicati da Aitchison e Dunsmore (1'75, p.74-77) e relativi al tasso di eliminazione per via urinaria di due steroidi metaboliti.

Tab.3: pazienti affetti dalla sindrome di Cushing.

Patient	cortisol	cortisone	Patient	cortisol	cortisone
1	0.41	0.38	20	0.42	1.48
2	0.16	0.18	21	0.16	2.50
3	0.26	0.15	22	0.48	2.94
4	0.34	0.33	23	0.26	0.24
5	1.12	0.60	24	0.16	0.53
6	0.15	0.14	25	0.19	0.19
7	0.20	0.16	26	0.18	1.92
8	0.26	0.18	27	0.35	1.03
9	0.56	0.32	a	0.32	0.18
10	0.26	0.20	b	1.12	0.32
11	0.16	0.13	c	1.12	0.48
12	0.56	0.33	d	0.48	0.31
13	0.33	0.08	e	0.96	0.32
14	0.26	0.22	f	1.04	0.40
15	0.48	0.36	g	11.20	0.75
16	0.80	0.39	h	0.04	0.03
17	0.40	0.24	i	0.07	0.10
18	0.10	3.46	j	1.60	0.40
19	0.24	1.44			

Alle primo 27 osservazioni è stata applicata la trasformazione Box-Cox bivariata e le stime di massima verosimiglianza dei parametri $\lambda=(\lambda_1,\lambda_2)$ sono state ottenute con la procedura Newton-Raphson suggerita da Rode e Chinchilli 1988.

Tab.4: elaborazione dati tab.3

Valori	X_1	X_2
Parametro λ	-0.4023	-0.4197
Media	-1.698	-1.3886
Mediana	-1.788	-1.4117
Dev.Std.	0.9159	1.4396
$\gamma 1$	-0.1203	0.0529
$\gamma 2$	-0.4461	-0.7483
Shapiro-Wilk	0.0182	0.2185
$L_i(0.9487,0.9)$	-3.6887	-4.5178
$L_u(0.9487,0.9)$	0.2927	1.7406

Come si vede la normalità marginale è accettabile (forse più per la X_1 che per la X_2) come confermano i valori piccoli di γ_1 e γ_2 e la non significatività del test di Shapiro-Wilk (Royston,1982). La normalità bivariata non é poi è contraddetta dalle misure di asimmetria e curtosi multivariata date da Mardia (1974):$\beta_{1,2}$=3.7883 e $\beta_{2,2}$=3.1765 entrambe poco significative. Nella Fig. 2 sono rappresentati i dati della tabella 1, opportunamente trasformati; il cerchio pieno indica le osservazioni di controllo (a-h).

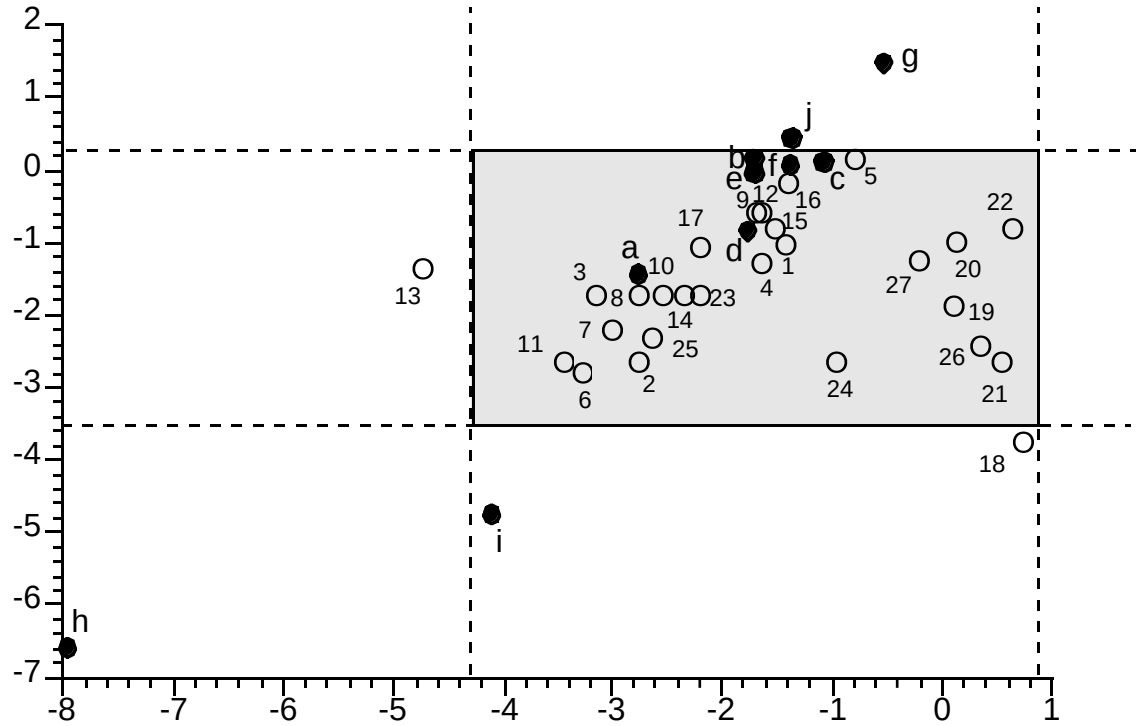

Figura 2: regioni di tolleranza come prodotto cartesiano di intervalli unidimensionali

Nell'esempio, la regione rettangolare è stata ottenuta assegnando lo stesso contenuto $\sqrt{\gamma}$=0.9487 agli intervalli marginali. Come si nota essa lascia fuori due osservazioni tra quelle usate per la stima e quattro tra quelle aggiuntive,

Per gli scopi di questo lavoro supporremo che sia possibile orientare ciascuna variabile separatamente dalle altre in modo che sull'asse che la rappresenta ci si sposti dalla carenza all'eccedenza e, a livello univariato, riterremo carente (eccedente) quella unità, esterna alla regione di tolleranza, che per quel carattere ha valore inferiore (superiore) alla mediana. Da un punto di vista aggregato l'unità sarà definita carente o eccedente se tale risulta in tutte le dimensioni. Resta. irrisolto il problema di come considerare quelle unità che si collocano nelle zone "miste", quelle cioè che risultano carenti rispetto ad una componente ed eccedenti o normali rispetto ad un'altra. In questo senso sarebbe necessario un ordinamento che permetta di graduare la distanza dalla normalità delle varie unità anche in funzione delle variabili su cui si realizza la carenza o l'eccesso (una linea molto prudenziale porterebbe a considerare carente una unità anche se la carenza si realizzasse in una sola dimensione). Tuttavia, come osserva Barnett 11976): *No reasonable basis exists for fully ordering a set of multivariate observations*. Si veda comunque Eddy (1985) per un tentativo in questa direzione). Per i dati dell'esempio risulta carente la osservazione *h*. Miste o, se si vuole essere prudenti, carenti le altre.

La tecnica del semplice prodotto cartesiano di intervalli unidimensionali lascia molto a desiderare sia dal punto di vista formale che da quello operativo: è valida solo nel caso di indipendenza delle *X*) e d'altra parte può succedere che una osservazione sia normale per ogni singola variabile, ma risultare poi anomala se si considerano tutti i caratteri congiuntamente. Parte dei problemi insiti nel caso *distribution-free* possono essere superati se si fanno assunzioni più esplicite sulla funzione di distribuzione. Nella fattispecie si ipotizze-

rà che la *f(X)* sia una gaussiana k-variata

$$f(\boldsymbol{x}|\Sigma,\mu)=\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}e^{-\frac{1}{2}\left\{(\boldsymbol{x}\text{-}\mu)^t\,\Sigma^{-1}(\boldsymbol{x}\text{-}\mu)\right\}} \tag{9}$$

con Σ di rango pieno. L'assunzione di multigaussianità implica che ognuno degli aspetti X_i ha un "polo" intorno a cui gravitano le uniti e che con frequenza approssimativamente uguale si riscontrano valori ad esso superiori o inferiori (per esempio sono escluse le variabili dicotome). In realtà, le deviazioni dalla multinormalità sono piuttosto la regola che l'eccezione almeno per le applicazioni nel contesto finanziario (Eisenbes, 1977) nell'ambito del quale vogliamo applicare la tecnica degli intervalli di tolleranza. Se i test sulla normalità danno esito negativo si renderanno necessarie sia una fase di ripulitura dei dati dai valori remoti adottando una opportuna tecnica di winsorizzazione (si vedano Gnanadesikan e Kettenring, 1972) nonché di trasformazione degli indicatori.

Se la (9) descrive opportunamente la densità congiunta degli indicatori la regione $R_\beta \supseteq R^k$ con volume minore che contiene il β% della popolazione à costituita da tutti i vettori **x** che soddisfano la relazione:

$$R_\beta : (\boldsymbol{x}\text{ - }\mu)^t\,\Sigma^{-1}(\boldsymbol{x}\text{ - }\mu)\le\chi^2_{k,\beta} \tag{10}$$

dove, $\chi^2_{k,\beta}$ è il quantile di ordine β della distribuzione chi quadro con k gradi di libertà. Il problema. allora di stimare R_β sulla base delle sole osservazioni campionarie assumendo, come è di solito, che il vettore delle medie μ e la matrice di varianze-covarianze Σ siano incogniti.

La stima naturale della (10) è

$$R_{\boldsymbol{m},\boldsymbol{S}} : (\boldsymbol{x}\text{ - }\boldsymbol{m})^t\,\boldsymbol{S}^{-1}(\boldsymbol{x}\text{ - }\boldsymbol{m})\le w_\beta,\ \ \boldsymbol{m}=(m_1,m_2,\ldots,m_k);\boldsymbol{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x_i}\text{ - }\boldsymbol{m})(\boldsymbol{x_i}\text{ - }\boldsymbol{m})^t \tag{11}$$

dove w_β è costante positiva scelta in modo che la (11) assicuri che

$$Pr\left[Pr\left(R_{\boldsymbol{m},\boldsymbol{S}}\right)\ge\beta\right]=\gamma \tag{12}$$

Il valore del parametro di soglia w_β dipende dal numero di variabili *k*, dal numero *n* delle osservazioni e dalle due probabilità: γ, e π; la sua determinazione esatta è molto complessa. Fraser e Guttman (1956) e Guttman (1970) hanno studiato la possibilità di controllare il grado di copertura medio della regione (11). In tal senso, se si pone

$$w_\beta=\frac{(n^2-1)k}{(n-k)n}F_{k;n-k,\gamma} \tag{13}$$

con $F_{k,n-k,\gamma}$quantile di ordine *(l-γ)* della distribuzione F di Fisher, si ottiene

$$E\left[Pr\left(R_{\boldsymbol{m},\boldsymbol{S}}\right) \geq \beta\right] = \gamma \tag{14}$$

In una direzione diversa e forse più operativa si è mosso John (1963) che ha studiato il problema di determinare una regione, basata esclusivamente su dati campionari che includesse la regione R_β della popolazione con probabilità prefissata. L'autore ha proposto due formule approssimate per w_β e la più operativa è (Johnson e Kotz,1972, p.35):

$$w_\beta = \left[\sqrt{\frac{(n-1)}{(n-k-2)}\chi^2_{\beta,k}} + \sqrt{\frac{(n-1)k}{n(n-k-2)}F_{k,n-k,\gamma}}\right]^2 \tag{15}$$

Se *n* e *k* non sono troppo piccoli la regione R_β è inclusa nell'insieme seguente:

$$(\boldsymbol{x} - \boldsymbol{m})^t \boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{m}) \leq w_\beta \tag{16}$$

con probabilità *γ*. Per i dati della tabella 3, trasformati come in precedenza, si ha

$$m=(-1.6982, -1.3886); \quad \boldsymbol{S} = \begin{bmatrix} 0.8389 & 0.0534 \\ 0.0534 & 2.0724 \end{bmatrix}$$

Inoltre, posto *β=90%*, *γ=95%* si ha $\chi^2_{0.9,2} = 4.559, F_{2,25,0.95} = 3.3582,\ w_\beta = 7.8544$. Nella Fig. 3 è rappresentato graficamente il problema.

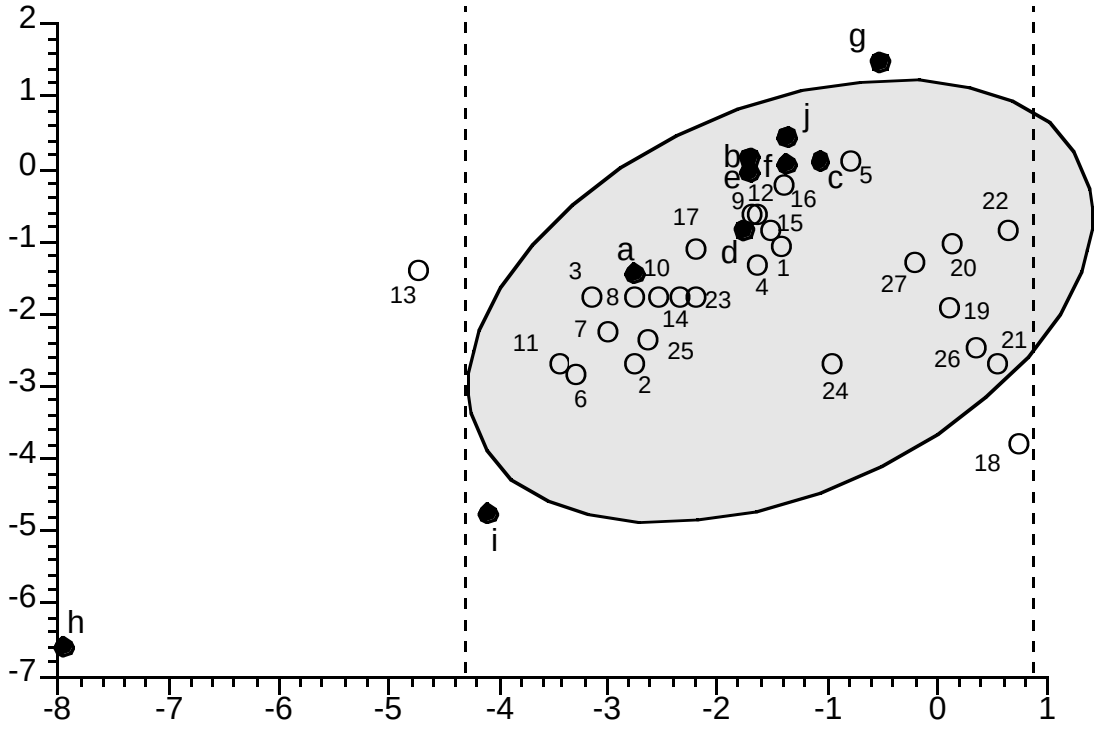

Figura 3: regioni di tolleranza di contenuto 90% confidenza al 95%

poiché la correlazione tra le due variabili è piccola le valutazioni no differiscono di molto dal caso precedente: solo l'osservazione j che prima risultava mista viene ora considerata normale.

4. Uso delle regioni di tolleranza per la previsione delle insolvenze

La valutazione dell'opportunità di affidare o meno una determinata azienda, di aprire o chiudere una linea di credito al consumo è un processo decisionale che si deve basare su di una visione generale, cicloramica dell'azienda o del singolo richiedente il credito. La banca o le società finanziarie sono costrette perciò ad esaminare una mole considerevole di informazioni contabili e personali che debbono essere sintetizzate in un giudizio sicuro sulla situazione economica, patrimoniale e finanziaria del potenziale creditore, soprattutto a fini prospettici. Si pensi poi che le informazioni più sicure su fallimenti, decreti ingiuntivi, pignoramenti, espropriazioni immobiliari raramente sono disponibili in tempo reale (solo in tempi recentissima è entrata in funzione una banca dati mirata sul problema rischio/ affidati). D'altra parte il numero di richieste che gli istituti finanziari (anche di ampiezza medio-piccola) ricevono è ormai tale da non consentire analisi di bilancio approfondite: gli uffici che effettuano le verifiche sono oberati dalle richieste e non hanno il tempo materiale di predisporre, per ogni pratica, una istruttoria esauriente con la conseguenza facilmente prevedibile di percentuali di sofferenza preoccupanti. Il ricorso alle tecniche quantitative -essenzialmente l'analisi discriminante tra due gruppi- per limitare il rischio connesso con procedure sommarie di avvio del prestito ha avuto una progressiva diffusione e consolidamento al punto che sono ormai operative tecniche automatizzate per l'istruzione delle pratiche di finanziamento.

Ci sono tuttavia diversi problemi con le tecniche di tipo discriminatorio. Innanzitutto la classificazione dicotoma dei clienti necessariamente in uno dei due gruppi "solvibili" ed "insolventi" si rivela troppo drastica in quanto tra le due è naturale ipotizzare una d*oubtful region*, cioè un'area di transvariazione in cui sono indistinguibili "gli apparentemente solvibili" dagli "apparentemente insolventi" e perciò l'analista dovrebbe limitarsi alla presentazione delle probabilità di appartenenza a posteriori ad uno dei due gruppi senza procedere alle assegnazioni che inficierebbero la capacità di corretta classificazione del modello di analisi discriminante.

Le regioni di tolleranza permettono di superare questo problema portando al rifiuto senza indugi del credito ai clienti che ricadono nella zona inferiore del carattere multidimensionale solvibilità ed alla concessione immediata del fido per quelli ricadenti nella zona superiore. Per quelli che si collocano nella zona di normalità si suggerisce di avviare maggiori indagini, di acquisire ulteriori informazioni o, se si vuole, di procedere con valutazioni intuitive basate sulla esperienza degli uffici istruttori delle società concedenti.

Applicazione

La tecnica degli intervalli di tolleranza multivariati è stata applicata a 162 aziende operanti nel Lazio ed in Toscana ed affidate da una importante banca romana e di cui era nota la posizione contabile.

Un primo passo di questa indagine avrebbe dovuto essere lo studio e la documentazione sul contesto economico-settoriale in cui si esplica l'attività delle imprese in modo da poter stabilire con ragionevole precisione la proporzione β che deve trovarsi nella regione

di tolleranza, ovvero quella di potenziale sofferenza. In mancanza di informazioni complete si è scelta la proporzione del 20% ovvero si è posto $\beta=0.60$ ritenendo, per simmetria, che ci sia un 20% di imprese certamente solvibile. Per il livello di confidenza γ è stato prefissato al 95%. Si è passati poi alla scelta delle variabili. In tal senso si candidavano gli indicatori con maggiori capacità di differenziazione tra clienti solvibili ed insolventi e che comunque fossero tali da rendere plausibile, magari con opportuni trattamenti, l'ipotesi di multigaussianità Dopo molte operazioni di cernita su 47 indicatori di bilancio (si veda Appetiti, 1984) la scelta è caduta sui cinque rapporti indicati nella tabella 5. Agli stessi è stata applicata la trasformazione Box-Cox multivariata

$$y_i = \frac{\left(x_i + a_i\right)^{\lambda_i} - 1}{\lambda_i} \quad i = 1,2,\ldots,5 \tag{17}$$

dove a_i è tale che $Min\{y_i\} \geq 0.01$. Tutti gli indicatori sono strutturati in modo che i valori superiori delle $\{y_i\}$ siano prevalentemente riscontrati nelle imprese solvibili. Nella tabella 5 si riportano, sommariamente, gli estremi dell'elaborazione. In particolare, le misure di asimmetria e curtosi multivariata sono risultate: $\beta_{1,5}=56.8534$; $\beta_{2,5}=7.1593$.

Tabella 5: elaborazioni sui dati finanziari

Quozienti finanziari	*Media*	*St.dev.*	γ_1	γ_2
$R1: \frac{Risultato\ netto}{margine\ operativo\ netto}$	0.91	0.40	0.14	0.73
$R2: \frac{oneri\ finanziari}{margine\ operativo\ lordo}$	−10.64	3.94	0.07	0.12
$R3: \frac{attività\ correnti}{ricavi\ netti\ d'esercizio}$	−12.59	3.50	−0.48	0.06
$R4: \frac{Circolante}{Ricavi\ netti\ d'esercizio}$	−255.08	77.98	−0.58	2.62
$R5$; *Variazione % attivo netto*	−10.19	3.72	−0.28	1.98

La regione (11) con valore di soglia $w_{o.60}= 6.6494$ ha determinato 118 unità normali e 44 unità anormali avvicinandosi quindi alla frazione del 60% stabilita per la popolazione. Le unità anormali, in base al criterio di ordinamento dato nel paragrafo 3, sono risultate: 6 eccedenti, 38 carenti. La concessione del fido avrebbe quindi dovuto essere attivata per le 6 eccedenti, esclusa per le 38 carenti, e discussa per le 118 normali. Nella realtà le eccedenti sono risultate tutte solvibili e quindi la rapidità d'esame non avrebbe prodotto costi per la banca. Delle 38 carenti però ben 26 erano solvibili e la loro esclusione avrebbe prodotto alla banca il costo del mancato guadagno della concessione.

I risultati non sono brillantissimi ed in effetti il numero di indicatori considerato è decisamente piccolo rispetto alla mole cospicua di variabili necessaria per cogliere tutti gli aspetti del fattore solvibilità. D'altra parte l'aumento delle dimensioni ha però come *trade-off* la difficoltà di avvicinare le variabili all'Ipotesi di multigaussianità che peraltro è già contestabile, rispetto alla curtosi, per questa applicazione a causa degli indicatori R4 ed R5. Poi, il criterio di ordinamento adottato eleva molto il rischio di mancata concessione a clienti solvibili. In questo senso sarebbe necessario intervenire con una graduazione più efficace che tenga conto in modo diverso degli indicato sui quali la carenza si verifica.

Bibliografia

Appetiti S. (1984). Identifying unsound firms in Ita1y. *Journal of Banking and Finance*, 8, 269-279.

Barnett V. (1976). The ordering of multivariate data. *Journal of the Rojal Statistical Society, A*, 139, 577-580.

Brooker P., Selby H.J.P. (1975). Algorithm AS92. The sample size for a distribution-free tolerance interva1. *Applied Statistics*, 24, 388-390.

Eddy W.F. (1985). Ordering of multivariate data. In Computer Science and Statistics. L. Billiard (ed.) Elsevier Publishers, S.V. (North-Ho1land), 25-30.

Eisenbes R.A. (1977). Pitfalls in the application of discriminant analysis in business, Finance, and economics. *Journal or Finance*, 32, 875-900.

Fraser D.A.S. (1953). Nonparametric tolerance regions. *Annals of Mathematical Statistics*, 24, 44-55.

Gnadesikan R., Kettenring J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 21, 27-58.

Guttman I. (1970). Construction of β-content tolerance regions at confidence level γ for large samples from the k-variate normal distribution. *Annals of Mathematical Statistics*, 41, 376-400.

Johnson N.L. Kotz S. (1972). Distributions in statistics. Continuous multivariate distributions. John Wiley, Sons, New York.

Kemperman J.H.B. (1956). Generalized to1erance limits. *Annals of Mathematical Statistics*, 27, 180-186.

Kendall M., Stuart A. (1979).The advanced theory of statistics.Vol.II, Fourth edition. Macmillan Publishing Co. New York.

Mardia K.V.(1974). Applications of some measures of multivariate skewness and kurtosys in testing normality and robustness studies. *Sankhya, B*, 36, 115-128.

Rode R.A., Chinchilli V. H. (1988). The use of Box-Cox transformations in the development of multivariate to1erance regions with applications to clinical chemistry. *The American Statistician*, 42, 23-30.

Royston J.P. (1982). An extension of Shapiro-Wi1k's test for norma1ity to large samples. *App1ied Statistics*, 31, 115-124.

Tiff1er R. J. (1982). Forecasting company failure in the UK using discriminant analysis and financial ratio data. *Journal or the Royal Statistical Society, A*, 145, 342-358.

Wald A. (1943). An extensions of Wilk's method for setting tolerance limits. *Annals of Mathematical Statistics*, 14, 45-55.

Wa1d A.,Wo1fowitz J. (1946). To1erance limits for a norma1 distribution. *Annals of Mathematical Statistics*, 17, 208-215.

Wilks S.S. (1942). Statistical prediction with specia1 reference to the problem of tolerance limits. *Annals ol Mathematical Statistics*, 13,400-409.