データ解析のツールとしての チェビシェフの不等式の多次元化の試み

新谷　明雲
山口県立大学　共通教育機構

Multi-Dimensional Chebychev's Inequality as a Tool in Data Analysis

Meiun SHINTANI
The General Education Division of Yamaguchi Prefectural University

Abstract

We provide three types of multi-dimensional Chebychev's inequqlity corresponding to the choice of domain surrounding the origin in the data space. The first choice of the domain is the hyper-cube , the second one is the hyper-sphere, and the last one is the hyper-ellipse characterized by the correlation coefficient in the L-dimensional data space. The corresponding inequality is respectively given. That each of them has merits and demerits is illustrated

Key words: multi-dimensional Chebychev's inequality, hyper-cube, hyper-sphere, hyper-ellipse, set theory, de Morgan's theorem, correlation coefficient, covariance matrix

キーワード：多次元チェビシェフの不等式、超立方体、超球体、超楕円体、集合論、ド・モルガンの定理、相関係数、共分散行列

§１　はじめに

近年、チェビシェフの不等式［１］は大量のデータを解析する分野においてにきわめて有効なツールとなっていることはつとに知られている［２］。

本稿のねらいは、多変量データを取り扱うために１次元データ列に対するチェビシェフの不等式を多次元化することにある。

まず「ｎ個の２次元データ列とは何か」ということから説明しよう。１つの対象に属する２組の変量（例えばある人の身長ｘと体重ｙ）に注目し、その値を２次元ユークリッド空間の１点に対応させ、これを直交座標（デカルト座標）（ｘ，ｙ）で記す。対象がｎ個（ｎ人）有れば２次元ユークリッド空間のｎ個の点として表示できる。これがｎ個の２次元データ列の意味である。本稿では便宜のため生のデータ列（ｘ，ｙ）よりも無次元化した標準化変数（基準化変数ともいう）（z_X，z_Y）を扱うこととする。ｘの標準化変数 z_X は（ｘ－$\bar{x}$）／ｓ（ｘ）で定義される。ここで　は$\bar{x}$平均値、ｓ（ｘ）はｘの標準偏差であり、ともにデータ列（統計量ともいう）ｘから算出される。

「２次元チェビシェフの不等式」とは、２次元空間内のｎ個のデータ点のうち原点を含む決められた閉領域Ｄの中にどれだけの割合（相対度数）で存在するかを不等式の形で評価することを可能ならしめるものである。同じくＬ次元データ列にこれを拡張したものが「Ｌ次元チェビシェフの不等式」となる。ここで「原点を含む決められた閉領域Ｄ」とは、いったい何を指すのかあいまいかもしれない。確かにあいまいである。チェビシェフの不等式はこの閉領域Ｄに応じた数だけ不等式は存在しうるわけである。あなたが好む領域を選びなさい。が好きなようにＤを選んだからと言って、不等式の導出が可能であるかはわからない。またそれが可能としても直観的に領域の図形を想起できるか、描画し易いか、意味のある結論が引き出せるか、などといった実用面からの不等式の真価が問われよう。

本稿では、閉領域ＤとしてＬ次元立方体（Ｌ＝１は

線分、Ｌ＝２は正方形）をまず考え、それに内接するＬ次元球体、Ｌ次元楕円体を扱うこととする。不等式の下限値が同じであれば体積（２次元では面積）が最も小さくなるような領域選択が最適と言えるだろう。

第２節ではチェビシェフの不等式の高次元への拡張を試みる。領域Ｄの選び方として、最初に正方形の拡張である超立方体を扱う。次に、円の拡張である超球体を、最後に共分散行列から導かれる超楕円体を扱う。それぞれの場合に、２次元データからはじめ３次元、４次元、…、Ｌ次元へと議論を進める。

第３節は、今後の展望と議論にあてられる。

§２　チェビシェフの不等式の多次元への拡張

２．１　１次元チェビシェフの不等式（復習）

チェビシェフの不等式の多次元化への拡張作業は基本的に一次元のチェビシェフの不等式が土台となるのでその導出過程を復習を兼ね概観してみよう。

次のｎ個の１次元（標準化）データ列

$$z_1, z_2, \cdots, z_n \quad , \left(z_i \equiv \frac{(x_i - \bar{x})}{s(x)} \right) \tag{1}$$

を考える。ｚの平均値　および分散ｓ^2（ｚ）は

$$\bar{z} = 0, \quad s(z)^2 \equiv \frac{1}{n} \sum_{i=1}^{n} z_i^2 = 1. \tag{2}$$

今、１次元ｚ空間を、$|z| \leq k$，および$|z| > k$，（$k > 1$）に分ける。分散１の式から、

$$1 = \frac{1}{n} \sum_{i=1}^{n} z_i^2 = \frac{1}{n} \sum_{|z|>k} z_i^2 + \frac{1}{n} \sum_{|z| \leq k} z_i^2 \geq \frac{1}{n} \sum_{|z|>k} z_i^2 \geq \frac{k^2}{n} \sum_{|z|>k} 1$$

が成り立つので

$$\frac{1}{n} \sum_{|z|>k} 1 \leq \frac{1}{k^2} \tag{3}$$

ないしは、

$$\frac{1}{n} \sum_{|z| \leq k} 1 \geq 1 - \frac{1}{k^2} \tag{4}$$

をうる。(3)、(4)の不等式が成り立つためには

$$k > 1 \tag{5}$$

であることが必要とされる。これが１次元のチェビシェフの不等式である。

冒頭で述べたようにＬ次元へのチェビシェフの不等式の拡張といっても、Ｌ次元空間の原点を含む閉領域Ｄの選び方は無数にありうるがその中でも最も単純な以下の３つの場合、立方体（§２．２）、球体（§２．３）、楕円体（§２．４）、に限定しそれぞれについて不等式の一般化をおこなう。

２．２　１辺の長さが$2\sqrt{Q}$のＬ次元立方体の場合

〈Ｌ＝２の場合〉

今、全データの集合をＵとし、その部分集合として

$$A = \{ (z_{x_i}, z_{y_i}) \mid |z_x| \leq \sqrt{Q}, |z_y| \leq \infty \} \subseteq U, \tag{6}$$

$$B = \{ (z_{x_i}, z_{y_i}) \mid |z_x| \leq \infty, |z_y| \leq \sqrt{Q} \} \subseteq U \tag{7}$$

を導入する。ここで正の数$\sqrt{Q}$は１より大であるが、相対割合に対する不等式が意味を持つように改めて定義される。閉領域ＤはＡとＢの共通集合、すなわちＤ＝Ａ∩Ｂとして定義する。Ｄは１辺が$2\sqrt{Q}$の正方形に入るデータの集合を表し、１次元チェビシェフの不等式からＵの真部分集合（Ｄ⊂Ｕ）となることが分かる。集合Ｓの要素の数をｎ（Ｓ）とあらわせば、ｎ（Ｕ）＝ｎとなる。各部分集合の要素数の間には次の関係が成り立つ。

$$n(U) = n(A) + n(B) - n(A \cap B) \tag{8}$$

また、

$(A \cup B) \cup (\overline{A \cup B}) = U$，　$(A \cup B) \cap (\overline{A \cup B}) = \phi$

であるので、

$$n(U) = n = n(A \cup B) + n(\overline{A \cup B}) \tag{9}$$

が成り立つ。ここで$\overline{A}$はＡの補集合を表す。(8)と(9)より、

$$n(\overline{A \cup B}) = n - n(A) - n(B) + n(A \cap B) \tag{10}$$

を得る。この式よりｎ（Ａ∩Ｂ）の下限値は、$n(\overline{A \cup B})$の下限値により達成される。

$$n(\overline{A \cup B}) \geq 0 \tag{11}$$

が、成り立つので、

$$n(A \cap B) \geq -n + n(A) + n(B) \tag{12}$$

をうる。一方、１次元チェビシェフの不等式により、

$$\frac{n(A)}{n} \geq 1 - \frac{1}{\sqrt{Q}}, \quad \frac{n(B)}{n} \geq 1 - \frac{1}{\sqrt{Q}}, \text{ for } Q > 1 \tag{13}$$

が成り立つので、(12)式は

$$\frac{n(D)}{n} \geq 1 - \frac{2}{\sqrt{Q}} \tag{14}$$

を得る。これが正方形領域Ｄ対する２次元チェビシェフの不等式にほかならない。ただし、(14)が意味を持つためには、正方形の一辺の半分（$\sqrt{Q}$）が

$$Q > 4 \tag{15}$$

とならねばならない。

〈Ｌ＝３の場合〉

次に一辺の長さが２ｋの立方体の場合を考えよう。全集合Ｕの部分集合Ａ、Ｂ、Ｃを定義する。

$$A = \{ (z_{x_i}, z_{y_i}, z_{z_i}) \mid |z_x| \leq \sqrt{Q}, |z_y| \leq \infty, |z_z| \leq \infty \} \subseteq U, \tag{16}$$

$$B = \{ (z_{x_i}, z_{y_i}, z_{z_i}) \mid |z_x| \leq \infty, |z_y| \leq \sqrt{Q}, |z_z| \leq \infty \} \subseteq U \tag{17}$$

$$C = \{ (z_{x_i}, z_{y_i}, z_{z_i}) \mid |z_x| \leq \infty, |z_y| \leq \infty, |z_z| \leq \sqrt{Q} \} \subseteq U \tag{18}$$

閉領域Ｄは、共通集合Ｄ＝Ａ∩Ｂ∩Ｃで定義される一辺が２ｋの立方体である。要素間の関係式として次の関係が成立する。

$$n(A \cup B \cup C) = n(A) + n(B) + n(C) - n(A \cap B) - (B \cap C) - n(C \cap A) + n(A \cap B \cap C) \tag{19}$$

また、１次元の場合の(4)式と２次元の場合の(14)式を用いることにより以下の式が成立する。

$$\frac{n(A \cap B \cap C)}{n} + \frac{n(\overline{A \cup B \cup C})}{n} \geq 1 - \frac{3}{\sqrt{Q}} \tag{20}$$

この式よりｎ（Ａ∩Ｂ∩Ｃ）の下限値は$n(\overline{A\cup B\cup C})$の上限値によって達成される。したがって、$n(\overline{A\cup B\cup C})$の上限値を求めよう。ド・モルガンの定理により$\overline{A\cup B\cup C}=\overline{A}\cap\overline{B}\cap\overline{C}$であるので包含関係

$$\overline{A\cup B\cup C}\subset\overline{A},\ \overline{B},\ \overline{C} \tag{21}$$

が成立するため要素数間に不等式

$$n(\overline{A\cup B\cup C})\le n(\overline{A}),n(\overline{B}),n(\overline{C}) \tag{22}$$

が成り立つ。したがって、１次元の１次元チェビシェフの不等式により

$$\frac{n(\overline{A\cup B\cup C})}{n}\le\frac{1}{\sqrt{Q}} \tag{23}$$

をうる。この式を⑳式に代入して

$$\frac{n(D)}{n}\ge 1-\frac{4}{\sqrt{Q}} \tag{24}$$

をうる。ここで

$$Q>16 \tag{25}$$

である。これが立方体Ｄ領域に対する３次元チェビシェフの不等式に他ならない。

一般Ｌ次元立方体に対しても集合論の定理と一次元チェビシェフの不等式を繰り返し用いることによりＬ次元チェビシェフの不等式を求めることが可能であると予想される。この点についての詳細報告は稿を改めて紹介することとする。

２．３　Ｌ次元立方体に内接する半径$\sqrt{Q}$の超球体の場合

〈Ｌ＝２の場合〉

ｎ個の２次元データ列（z_{x_i},z_{y_i}）（ｉ＝１，２，…，ｎ）は、各成分が独立に⑵式を満足する。

$$\bar{z}_x=0,\bar{z}_y=0 \tag{26}$$

$$s^2(z_x)=\frac{1}{n}\Sigma_{i=1}^{n}z_{x_i}^2=1,\quad s^2(z_y)=\frac{1}{n}\Sigma_{i=1}^{n}z_{y_i}^2=1 \tag{27}$$

いま、２次元極座標（ｒ，θ）を導入し、

$$z_{x_i}=r_i\cos\theta_i,\ z_{y_i}=r_i\sin\theta_i,\ (0<r_i<\infty,\ 0\le\theta_i<2\pi) \tag{28}$$

⑵式を書き換えると

$$1=\frac{1}{n}\Sigma_{i=1}^{n}r_i^2\cos^2\theta_i,\quad 1=\frac{1}{n}\Sigma_{i=1}^{n}r_i^2\sin^2\theta_i, \tag{29}$$

となり、この２つの式の辺々を足し合わせると

$$2=\frac{1}{n}\sum_{i=1}^{n}r_i^2(=\bar{r}^2) \tag{30}$$

をうる。この式はｒの二乗平均値が２であることを意味する。この値はＬ次元ではＬとなる。この等式を、半径$\sqrt{Q}$の内側と外側に分けることを考える。すなわち,

$$2=\frac{1}{n}\sum_{i=1}^{n}r_i^2=\frac{1}{n}\sum_{r>\sqrt{Q}}r_i^2+\frac{1}{n}\sum_{r\le\sqrt{Q}}r_i^2\ge\frac{1}{n}\sum_{r>\sqrt{Q}}r_i^2\ge\frac{Q}{n}\sum_{r>\sqrt{Q}}1$$

が成り立つので

$$\frac{1}{n}\sum_{r>\sqrt{Q}}1\le\frac{2}{Q} \tag{31}$$

ないしは、

$$\frac{1}{n}\sum_{r\le\sqrt{Q}}1\ge 1-\frac{2}{Q} \tag{32}$$

をうる。ただし

$$Q>2 \tag{33}$$

を満たす必要がある。これが２次元球面（円）を閉領域Ｄとするチェビシェフの不等式に他ならない。⑶式の右辺は、正方形に対するチェビシェフの不等式のそれに等しい。したがって、正方形とそれに内接する円が、同じ評価式を与えるので、面積の小さな円の方が評価式としてはより強力と言える。

〈Ｌ≧３の場合〉

⑶式のＬ次元超球体に対する拡張は容易で、

$$\frac{1}{n}\sum_{r\le\sqrt{Q}}1\ge 1-\frac{L}{Q} \tag{34}$$

として与えられる。ただしＱに対する制限は⑶式を

$$Q>L \tag{35}$$

で置き換えればよい。３次元においては、立方体の評価式は㉔式で与えられるので、球体の方がより大きな下限値を与える。したがって、球体の方が不等式としては強力と言える。４次元以上でも同様の結論となることが予想されるが、検証は今後の研究に委ねることとしよう。

２．４　１辺２$\sqrt{Q}$のＬ次元立方体に内接する超楕円体の場合

これまでは閉領域Ｄの選び方は変量間の相関の強さとは無関係な選択であった。この節では相関の強さを考慮に入れた領域Ｄに対するチェビシェフの不等式を導こう。

〈Ｌ＝２の場合〉

ｎ個の２次元データ列（z_{x_i},z_{y_i}）（ｉ＝１，２，…，ｎ）に対し㉖、㉗式がなりたつ。ｘとｙの間の相関の強さを表す統計量として相関係数r_{xy}を導入する。

$$r_{xy}=r(x,y)=r(z_x,z_y)\equiv\frac{1}{n}\sum_{i=1}^{n}z_{x_i}z_{y_i},\qquad(-1\le r_{xy}\le 1). \tag{36}$$

いま、次の変量ｑを考えよう。

$$q_i\equiv\frac{1}{1-r_{xy}^2}\left\{z_{x_i}^2-2r_{xy}z_{x_i}z_{y_i}+z_{y_i}^2\right\},\quad(i=1,2,\cdots,n). \tag{37}$$

この量は、判別式が負となることからz_{xi}、z_{yi}の値によらず常に正値を取る。平均値$\bar{q}$は

$$\bar{q}=\frac{1}{n}\Sigma_{i=1}^{n}q_i=2 \tag{38}$$

で、円の場合の㉚式に酷似する。とくに無相関のとき、すなわち$r_{xy}=0$の場合

$$q_i=r_i^2 \tag{39}$$

となり円の半径の２乗値に一致する。いまこの２次元空間に次の楕円の方程式を考える。

$$Q=\frac{1}{1-r_{xy}^2}\left\{z_x^2-2r_{xy}z_xz_y+z_y^2\right\},Q>0 \tag{40}$$

Ｑ値は常に正であり楕円の大きさを表す量である。式（３８）を、Ｑ値より大きい領域と小さい領域に分けて考えよう。

$$2=\frac{1}{n}\sum_{i=1}^{n}q_i=\frac{1}{n}\sum_{q>Q}q_i+\frac{1}{n}\sum_{q\le Q}q_i\ge\frac{1}{n}\sum_{q>Q}q_i\ge\frac{Q}{n}\sum_{q>Q}1 \tag{41}$$

となるので、次の不等式をうる

$$\frac{1}{n}\sum_{q>Q}1 \le \frac{2}{Q} \tag{42}$$

ないしは、

$$\frac{1}{n}\sum_{q\le Q}1 \le 1-\frac{2}{Q} \tag{43}$$

ただし、Qは

$$Q>2 \tag{44}$$

を満たさなければならない。これが楕円領域に対する２次元チェビシェフの定理に他ならない。(40)式であたえられる楕円が１辺$2\sqrt{Q}$の正方形に内接する楕円であることは、座標軸を45°回転することにより理解できる。 $z_x=\frac{\sqrt{2}}{2}(z_{x'}+z_{y'})$ ，

$$z_y=\frac{\sqrt{2}}{2}(z_{x'}+z_{y'})$$

として座標変換すると、(40)式は楕円の標準形

$$\frac{z_{x'}^{\,2}}{1+r_{xy}}+\frac{z_{y'}^{\,2}}{1-r_{xy}}=Q \tag{45}$$

となる。Fig.１にみられるように正の相関（$r_{xy}>0$）の場合には長軸は$\sqrt{(1+r_{xy})Q}$で、短軸が$\sqrt{(1-r_{xy})Q}$となる。原点と正方形と内接する点を結べば楕円長軸を挟んだ２本の直線

$$z_y=r_{xy}z_x \ ,\ z_y=\left(1/r_{xy}\right)z_x$$

が得られるが、傾きがr_{xy}のものを「z_xのz_yへの回帰直線」、傾きが１／r_{xy}のものを「z_yのz_xへの回帰直線」とそれぞれ呼ばれる。楕円の面積は

$$\pi Q\sqrt{1-r_{xy}^{\,2}} \tag{46}$$

となり、一方１辺が$2\sqrt{Q}$で面積が４Qの正方形に内接する円の面積は

$$\pi Q\ (\approx 3.14\ Q) \tag{47}$$

であるので、同じ下限値に対し面積は楕円の方が$\sqrt{1-r_{xy}^{\,2}}$だけ小さくなる。とくに、r_{xy}の絶対値が１に近づくにつれ面積は急激に減少する。したがって、等面積であるならば、楕円領域のチェビシェフの不等式はより大きな下限値を与え、より強力な評価式を我々に提供することになる。

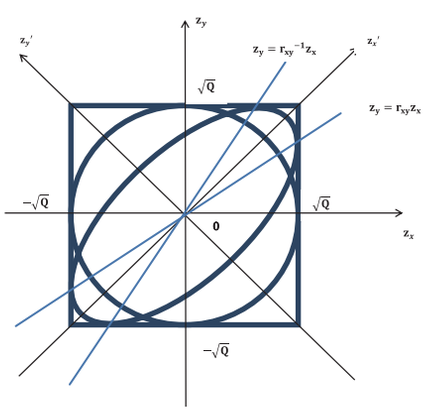

Fig.1　１辺が$2\sqrt{Q}$の正方形に内接する円と楕円($r_{xy}>0$の場合)

〈L≧３の場合〉

２次元の場合と同様に、行うことが可能である。Q値の定義式は多次元正規分布の指数ｅの肩が（$-\frac{Q}{2}$）であたえられる。すなわちQは共分散行列Vとデータを表す標準化された列ベクトルｚから作られる２次形式に他ならない。

$$Q=\mathbb{z}^{T}\ V^{-1}\mathbb{z} \tag{48}$$

ここで、V^{-1}は、共分散行列Vの逆行列を表し、$\mathbb{z}$を列ベクトル表記としたら$\mathbb{z}^{T}$は行ベクトルに対応する。ちなみに３次元の共分散行列Vは以下のように与えられる。

$$V=\begin{pmatrix}1 & r_{xy} & r_{xz}\\ r_{xy} & 1 & r_{yz}\\ r_{xz} & r_{yz} & 1\end{pmatrix} \tag{49}$$

L次元超楕円体に対するチェビシェフの不等式は、

$$\frac{1}{n}\sum_{q\le Q}1 \le 1-\frac{L}{Q} \tag{50}$$

となる。ただし、

$$Q>L \tag{51}$$

であたえられる。

§３　おわりに

前節で、多次元チェビシェフの不等式を、超立方体、超球体、超楕円体について求めることができた。それぞれに長短がある。相関係数を考慮に入れたチェビシェフの不等式は、大量の多変量データを数値解析する研究者向けには最適言える。特に相関係数の絶対値が１に近ければ近いほどその威力は絶大となる。一方、領域内の数のカウントという意味では、立方体が最も手軽であり、市販の表計算ソフトなどで簡単にカウントが可能である。球体はその中間に位置し、高度のプログラムは必要としないがそれなりの覚悟が必要となる。

L次元の直方体に対する不等式の一般化もこの論文に基づき容易に行うことができる。読者への練習問題として残しておく。

４次元以上の超立方体の場合のチェビシェフの不等式については、別の稿に残すこととする。

参考文献

［１］例えば、西田敬義ほか著、大学演習「数理統計学（第８版）」、裳華房（昭和49年）

［２］北原知就ほか、「一般化チェビシェフ不等式とその最適化への応用」、数理解析研究所考究録、第1534巻（2008）21－24。この論文の参考文献にチェビシェフの不等式の一般化に関する文献が見られる。