極値理論における離散化の影響

志村　隆彰　数理・推論研究系　助教

【極値統計学の基本的な問題】

地震や洪水などの自然災害に代表される、めったに起こらないが、一旦起こると大変大きな影響を及ぼしたり、重要な意味を持つ現象は数多い。このような現象を統計的に扱う場合、日常を表す平均値のような指標ではなく、非日常を表わす最大値のような指標が重要になる。極値統計学はいわば非日常を研究対象にする分野であり、数学的にもっとも基本的な設定と関心事は$X_1, X_2, \ldots$を共通の確率分布Fに従う実数値独立確率変数列としたときのX_nまでの最大値$M_n = \max\{X_1, \ldots, X_n\}$の$n \to \infty$のときの挙動である。$M_n$が$F$の上端点$x_F = \sup\{x : F(x) < 1\}$（無限と有限の両方がありうる）に収束するのは明らかであるから、以下のように正規化したときの極限分布（極値分布という）と極限を持つ分布の集合（吸引領域という）を求める。適当な定数列$a_n > 0$と$b_n \in \mathbf{R}$により、

$$\mathcal{L}\left(\frac{M_n - b_n}{a_n}\right) \to G \quad (n \to \infty)$$

極値分布Gにはフレシェ分布、グンベル分布、（極値）ワイブル分布の３種類があり、それぞれの極値分布に収束する分布Fの全体はその（最大値）吸引領域と呼ばれ、裾（確率）

$$\bar{F}(x) = 1 - F(x)$$

の$x \to x_F$のときの漸近挙動で特徴付けされる。フレシェ分布$\Phi_\alpha(x) = \exp(-x^{-\alpha})(x \geq 0, \alpha > 0)$の吸引領域は任意の$\lambda > 0$に対して

$$\lim_{x \to \infty} \bar{F}(\lambda x)/\bar{F}(x) = \lambda^{-\alpha}$$

で特徴づけられる。これを裾$\bar{F}(x)$が指数$-\alpha$の正則変動性を持つという（$\bar{F}(x) \in \mathbf{RV}_{-\alpha}$と書く）。パレート分布や（非正規）安定分布などがこの性質を持つ。与えられたデータから裾が正則変動性を持つ分布の指数を推定することは基本的な問題であり、Hill推定量をはじめいくつかの推定量が提案されている。
更に詳細な裾挙動の性質である２次の正則変動性が以下のように定義される。

（２次正則変動性）

$$\lim_{x \to \infty} \frac{\bar{F}(\lambda x)/\bar{F}(x) - \lambda^{-\alpha}}{A(1/\bar{F}(x))} = \frac{\alpha}{\rho}\lambda^{-\alpha}(\lambda^{\alpha\rho} - 1),$$

ここで$\lim_{x \to \infty} A(x) = 0$, $|A(x)| \in \mathbf{RV}_\rho(\rho \leq 0)$. $\rho = 0$のときの右辺は$\alpha^2\lambda^{\alpha(\rho-1)}\log\lambda$である。

【分布の離散化と連続化】

分布Fの離散化を、nを整数とし、$(n-1, n]$の測度を$\{n\}$に集めて、離散分布（整数値分布）にする操作と定義し、離散分布の連続化を、連続分布でその離散化した分布がもとの離散分布と一致するような連続分布を対応させる操作とする（連続化は唯一には決まらない）。離散化は本来連続量であるものから得られた丸目誤差を含んだデータ、連続化は丸目誤差を含んだデータの元である連続量のデータを意味している。

【離散化の影響】

確率論（数学）を統計解析に応用する際、理論で仮定しているのは連続量であるのに実際に使うのは（精度の差はあるが）離散量になる。極値統計ではこの違いが無視できないことがある。第一に、吸引領域への属性と分布Fの連続性の間には密接な関係があり、不連続性は吸引領域への属性を損なうことがある。裾が軽いほど大きな影響を受け、パレート分布は離散化されてもフレシェ分布の吸引領域から外れないが、指数分布は離散化されて幾何分布になるとグンベル分布の吸引領域には入らなくなる。更にたとえ吸引領域から外れなくても２次正則変動性を失い、推定量の性質に悪い影響を与えることがある。たとえば、正則変動する裾をもつ分布の裾の代表的なHill推定量は離散データが激しく振動してしまうことが報告されている。このような事情から、離散データに対してより正確な極値解析が行われる手法が望まれる。

【これまでの成果と今後の課題】

離散化の吸引領域への属性への影響は、主に[1]で考察した。主な結果は、まず離散化されても吸引領域への属性が保たれるための必要十分条件を与えたこと。たとえば、フレシェ分布の吸引領域の分布と対数正規分布のようなグンベル分布の吸引領域に属する分布の一部は離散化されてもそこから外れない。次に、幾何分布と指数分布の関係を一般化した連続化して吸引領域に属するような離散分布の特徴付けを与えた。ポアソン分布はどの吸引領域にも属さない分布であるが、その適当な連続化はグンベル分布の吸引領域に属する。ポアソン分布だけでなく、かなり多くの分布が吸引領域にする適当な連続化をもつことが示され、連続化の多様性についても、連続化が漸近的な意味で唯一に決まるための必要十分条件やある程度自然な制限の元での唯一に決まる条件などを与えた。この意味で幾何分布と指数分布の対応は一対一である。

現在研究中の課題は正則変動性を持つ裾の指数の新しい推定方法や離散化の２次正則変動性への影響などである。このうち、後者について部分的な結果を挙げる。

例　次の分布は$\alpha = 1, \rho = -1$の２次正則変動性を持つが、離散化されるとそれを失う（一次の正則変動性は保たれる）。

$$\bar{F}(x) = \left(1 + \frac{1}{x}\right)\frac{1}{x}.$$

命題　$\bar{F}(x) \in \mathbf{RV}_{-\alpha}(\alpha > 0)$が正規化された緩慢変動関数$l(x)$により$\bar{F}(x) = x^{-\alpha}l(x)$と書けるとする：$l(x) = c\exp(\int_a^x \epsilon(t)t^{-1}dt)$, $c > 0$, $a \in \mathbf{R}$, $\epsilon(t) \to 0$ $(t \to \infty)$。ただし、$l(x)$が正定数へ収束する場合は除く。このとき、２次正則変動性をもてば、$\rho = 0$であり、離散化で２次正則変動性を失わない。

【極値統計関連情報】

極値理論・統計関連の研究会を紹介する。

統数研では、共同研究集会として「極値理論の工学への応用」を1994年以来毎年開催している。

昨年度は、文部科学省委託事業　数学協働プログラム「甚大災害の外力想定に必要となる極値統計解析法の背景と活用」(H26年12月8日(月)於：京都大学防災研究所）が開催された。

丁度現在極値理論の国際会議 Extreme Value Analysis(June 15-19, 2015 in Ann Arbor)がミシガン大学で開催中である。

http://sites.lsa.umich.edu/eva2015/

前回2013年は上海で開催され、2017年はオランダでの開催が計画されている。

参考資料

[1] Discretization of distributions in the maximum domain of attraction, Extremes, 15 (2012) 299-317.

[2] Limit distribution of a roundoff error, Statistics and Probability Letters 82 (2012), 713-719.

[3] A numerical characteristic of extreme values, Statistics and its Interface 7 (2014), 375-379.

http://www.ism.ac.jp/ shimura/