離散性の極値解析への影響

志村　隆彰　数理・推論研究系　准教授

【極値】

数多くの数値データを大きさの順に並べたものを順序統計量といい、その中で特に大きい（或は小さい）ものを極値(extreme value)と呼ぶ。たとえば、その数値がある日の時間ごとの気温のデータであれば、最高気温や最低気温は極値であり、ネット上の同じ商品の店ごとの価格であれば、最安値は極値である。少し考えれば、スポーツ大会の優勝記録や最高齢者なども極値であることに気づき、極値が身近で、かつ特別な意味を持つものであることがわかる。

【極値統計・極値理論】

この極値を扱う統計学が極値統計であり、歴史的には災害対策と強く結びついて発展してきた。数値データが毎日の降水量を表すものであれば、長い期間での極値は水害をもたらすような大雨や豪雨を意味し、その起こり方は防災上極めて重要な関心事になる。極値統計では、数値がランダムなものとみなし、その起こり方を考える。降水量の例で、一日の降水量をランダムとみなし、その分布 F であらわすことにする。極値の視点からは、分布 F が大きな値を取る場合が重要であるから、この確率を表す裾（確率）(tail (probbility))

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

の $x \to x_F$ のときの挙動(0へ行く程度（速さ）)がキーとなる（x_F は 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）)。$x_F = \infty$ の場合であれば、$x \to \infty$　のときに $\bar{F}(x) \to 0$ となるときの速さが正規分布のように指数オーダーなのか、コーシー分布のようにべきオーダーなのか、或は対数オーダーなのか、そういったことが効いてくる。0 へ行くのが速れば、極端な値が出にくく、 遅ければ極端な値が出やすいことを意味し、速い分布のことを裾が軽い(light tail)、遅い分布のことを裾が重い(heavy tail)という。防災の観点からは、裾が重く、極端な値が出やすいパレート分布のようなものを考えることが多い。

さて、問題となるのは水害を起こすような雨の降り方であり、長期間における降水量の時系列の最大値を表す最も基本的モデルは、$X_1, X_2, \ldots$ 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 $M_n = \max\{X_1, \ldots, X_n\}$ であり、 M_n の $n \to \infty$ のときの挙動を考える。このとき、$n \to \infty$ のとき $M_n \to x_F$ は明らかである（従って、この事実は意味がない）から、確率変数の和に対する中心極限定理と同じように、M_n を正規化したときの（非退化）極限分布（極値分布）を求めるのが基本的な問題となる。すなわち、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）$\bar{F}(x)$ 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。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、ヒル（Hill）推定量をはじめいくつかの推定量が提案されている。推定の際に更に詳細な裾の性質である次の２次の正則変動性が仮定されることがある。

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

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。

【研究動機】

極値統計に限らず、数理統計の数学的結果を現実のデータに適用する際の問題点として、現実のデータが必ずしも数学的仮定を満たさないということがある。典型例として、現実のデータは必ず誤差を含むことがあげられる。誤差には様々なものがあり、観測精度を上げることをはじめ、その悪影響を軽減すべく、対応策が講じられてきたことは言うまでもないが、どのような工夫をしても、数値にする以上、丸め誤差は避けて通れない。以下では、極値統計における丸め誤差（打切り誤差）を意識しながら、離散分布に対する極値理論の問題点を考える。

【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め誤差を含んだデータ、連続化は丸め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。

【離散化の問題点の例】

１.「分布が吸引領域から外れる＝正規化された最大値が収束しない」

吸引領域への属性と分布 F の局所的な滑らかさには密接な関係があり、離散性は吸引領域への属性を損なうことがある。離散性の影響は、裾が軽いほど大きく、重いほど少ない。具体的には、パレート分布や対数正規分布は離散化されてもそれぞれフレシェ分布とグンベル分布の吸引領域から外れないが、指数分布や正規分布のような裾が軽いものは離散化されるとグンベル分布の吸引領域には入らなくなる。

２．ヒル推定量が激しく振動する。

離散化されても吸引領域から外れなくても、ヒル推定量が離散データでは激しく振動してしまうことが報告されている。これは離散化によって、分布が２次正則変動性を失ったことが原因と考えられる。

以上の例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の開発が望まれる。

【これまでの成果と今後の課題】

これまでに得られている主な結果：

・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。

・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。ポアソン分布をはじめ、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。

・指数型分布（逆正規分布など裾が指数分布に近い挙動をする分布）の離散化分布に対しては、適当な分布との合成積（別の独立な確率変数を加えることによる数値補正）により、吸引領域に戻すことができる。

今後の課題：

・三つ目に書いた補正方法を他の分布に拡張する（指数分布より軽い裾を持つ分布の補正方法)。

・ヒル推定量の振動理由の解明と正確なパラメータ推定のための対策。

【極値統計関連情報】

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ます。今年度は来月７月２０日（金)、２１日（土）に開催されるので、ご興味のある方はご参加ください（事前登録・参加料などは必要ありません)。

極値理論の国際会議である Extreme Value Analysis が隔年で開催されています。昨年2017年の第10回はオランダのデルフト（デルフト工科大学）で行われ、来年2019年はクロアチアのザグレブで計画されています。他にも「第３回極値解析の進展と自然災害への応用」がイギリスのサウサンプトンで開催されました。

【無限分解可能過程関連情報】

極値理論以外に、統数研共同研究集会「無限分解可能過程に関連する諸問題」を長らく開催しています。現時点では日程未定ですが、秋から初冬に開催予定されます。

共同研究集会の情報は統数研ホームページイベント欄及び発表者のホームページhttp://www.ism.ac.jp/ shimura/ に掲載します。