社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－MPS－62 (20)
2006－BIO－7 (20)
2006／12／22

ブートストラップ法を用いた分布の裾指数の推定手法の改良

澤村 めぐみ 吉田 裕亮
お茶の水女子大学 大学院人間文化研究科

本研究では裾が重たい分布の裾の厚さを測る指数の推定を行う. 裾が重たい分布が現れる現象としては, 自然現象や金融市場など様々な場面で観測される. t-分布や安定分布は裾が重たい分布として知られている. 安定分布は 1925 年に Levy により定式化された. 金融工学の分野でリスク管理を行う際には, 観測された事象のデータからモデル化を行うことが必要である. 最適なモデルを求めることができれば, より損失を最小にすることが可能になる. 多くの場合, 裾の厚い分布があらわれ, より正確なパラメータを決めることが求められている. 本研究では分布の裾を測るパラメータをリサンプリング手法であるブートストラップ法を用いて求める手法に関する考察を行った.

An improvement of the estimation for the tail index by the bootstrap method

Megumi Sawamura, Hiroaki Yoshida
Ochanomizu University
Graduate school of Humanitics and Sciences

The t-distributions and the stable distributions are known to have fat tail, which appear in some natural phenomena. On the risk management in financial field, it is strongly required to estimate the optimized parameters of the stochastic models more accurately in order to make the loss be minimum. In this study we investigate the method for the estimation of the tail indices of probability density functions of fat tail type by the bootstrap method.

1 はじめに

いわゆる裾の重たい (fat tail) 確率分布の裾の厚さを示す指数 (パラメータ) を推定することはとても困難であり, 任意に, 例えばサンプル数の上位 1 % 点のデータを選ぶといった方法は適さない.

一般に, 確率分布関数 $F(x)$ 正則変動を持つとは, ある $\gamma > 0$ が存在して, 緩慢変動関数 L を用いて

$$1 - F(x) = x^{-1/\gamma} L(x) \qquad (x \to \infty)$$

と表されるときという. ただし,L は, 固定された x に対して

$$\frac{L(tx)}{L(t)} \to 1 \qquad (t \to \infty)$$

を満たす. このとき $1/\gamma$ をテール指数 (tail index) あるいは裾の厚さの指数（以下, 裾指数）とよぶ. $\gamma = \infty$ ならば, 裾は減少する. つまり正規分布とほぼ同じとなる. $\gamma < \infty$ ならば, 裾は厚いことになる. 例えば t-分布の場合は γ は自由度を表す. また正規分布を除く, 安定分布の場合は γ は特性指数を表すことが知られている.

一般的な確率分布のパラメータの推定方法では最尤法がよく用いられる. 最尤法とは確率密度関数から尤度関数求め, 尤度を最大にするようにパラメータを推定する手法である. この手法は確率密度関数が簡単な式で定義される分布に適している. しかし, 以下で述べる安定分布やパラメータが厳密に陽に解けない場合などは裾指数を最尤法で推定することができない. そこで最尤法に代わる効率的な推定方法が必要とされている.

先行研究として, もちろんさまざまな手法の提案がされている. 最近リサンプリングを行うブートストラップ法を 2 回用いたダブルブートストラップ法が Danielsson [1] らにより提案された. この手法は計算量は多いが, かなり正確な推定が可能であることがシミュレーションで示されている. 本研究では, こ

のダブルブートストラップ法を簡素化した手法による裾指数の推定の可能性をシミュレーションにより調べた.

2 裾推定

2.1 裾の重たい分布

ここで, 本研究で用いた, いわゆる裾の重たい分布の代表的な例を挙げておこう.

t-分布

統計解析でよく用いられる分布であるが密度関数は, かなり煩雑である. 自由度 m の t-分布の確率密度関数 f_m は

$$f_m(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})}\Big(1+\frac{x^2}{m}\Big)^{-\frac{m+1}{2}}$$

で与えられる. 自由度 m が大きくなると正規分布に近づく. ただし,Γ はガンマ関数である.

安定分布

安定分布とはフラクタル幾何学の創始者である Mandelbrot が株式市場において見出したものである. 特徴として正規分布より裾が厚く中心の周辺で尖りの急な分布であり, 急尖的分布, パレート分布などとも呼ばれている. 株式市場の収益率など金融の分野や自然界など様々なところで表れる分布である. 安定分布は分布関数が陽に定義されず, 密度関数の特性関数 (Fourier 変換) がパラメータ α によって以下のように表される.

$$\widehat{f}(x) = e^{-c|x|^{\alpha}} \ (0 < \alpha < 2)$$

ただし, c は正規化定数である. パラメータ α は裾を表す尺度であり, 値が小さいほど裾が重い. 安定分布は正規分布 $(\alpha = 2)$ やコーシー分布 $(\alpha = 1)$ の概念を含む広いクラスの分布である. 一般に, 密度関数は特別な場合を除いて, 単純な関数で与えられない分布であるため, 最尤法で推定することが難しい.

2.2 Hill 推定量

Hill(1975) が提唱したアルゴリズムで, 分布の裾の形状の推定方法として代表的なものである. 裾が丁度, パレート分布に従っているとき, すなわち, ある定数 $a > 0$ が存在して

$$1 - F(x) = ax^{-1/\gamma}$$

となるならば, パラメータ $1/\gamma$ の推定量は, 以下の通りになる.

$$\frac{1}{\gamma} = \frac{1}{k}\sum_{i=1}^{k}\log X_{n,n-i+1} - \log X_{n,n-k}$$

ここで,$X_{n,i}$ は観測値の順序統計量, つまり ($X_{n,1}$ □ $\cdots$ □ $X_{n,n}$) である. また k とは裾部分に属する観測値の数を表している. Hill 推定量では k を的確に選択することで裾指数を推定することが可能になる. したがって, この裾領域を示す k の推定が重要な問題となる.

2.3 ブートストラップ法

一般に, リサンプリング法はデータからのサンプリングによってばらつきを評価する一般的な手法であり, 一種の確率シミュレーション技法である. 計算機の発展が伴って初めて有効な手法と成り得た手法でもある. その中でもブートストラップ法は代表的な手法で, Eflron(1979) によって定式化された.

たとえば, データ $(x_1, \ldots, x_n)$ を各要素は独立同分布に従う確率変数の実現値とする. その期待値 μ の推定値は平均 $\overline{x} = (x_1 + \cdots + x_n)/n$ によって与えられる. $\overline{x}$ のバラツキを調べるために, データと同じ確率分布に従う確率変数の実現値を擬似乱数を使って生成 (リサンプリング) したものを, 複製データ $(x_1^*, \ldots, x_m^*)$ とする. この手続きを多数回繰り返し実行して, $\overline{x}^* = (x_1^* + \cdots + x_m^*)/m$ のばらつきを観測することにより, $\overline{x}$, すなわち μ の推定値のバラツキが推定できる.

一般にブートストラップ法を用いることで実験データに内在するばらつきの影響により導かれる誤差を防ぐことも可能となる.

3　裾領域の推定方法

3.1　推定のアルゴリズム

1. 全集合 $(x_1, \ldots, x_n)$ からリサンプリングによりサイズ m $(m < n)$ の標本 $(x_1^*, \ldots, x_m^*)$ を多数作成する.
2. それぞれの標本において二乗誤差 (AMSE) を計算し, 最小とする値の平均を k_m^* を求める. ここで,AMSE とは

   $$\widehat{AMSE}(m, k_i) = E\big(M_m(k_i) - 2(\gamma_m(k_i))^2 |(x_i)\big)$$

   なる期待値のことで $M_m(k_i)$ と $\gamma_m(k_i)$ は, 以下の式で定義される.

   $$M_m(k_i) = \frac{1}{k_i}\sum_{j=1}^{k_i}\big(\log x_{m-j+1} - \log x_{m-k_i}\big)^2$$

   $$\gamma_m(k_i) = \frac{1}{k_i}\sum_{j=1}^{k_i}\log x_{m-j+1} - \log x_{m-k_i}$$

3. 上の 2. で定まる k_m^* はリサンプリングサイズ m に対する裾部分に属する観測値の推定された数である. 全集合のデータ数 n に対する裾部分に属する観測値の数 k に換算して Hill 推定量を計算する.

   $$\frac{1}{\gamma} = \frac{1}{k}\sum_{i=1}^{k}\log X_{n,n-i+1} - \log X_{n,n-k_m^*}$$

ここで,k_m^* を的確に選ぶことが重要である. しかし, リサンプリングするだけでは AMSE のブレを消去することはできない. Danielsson らのダブルブートストラップ法はリサンプリングのサイズを変えて (1),(2) の手順を繰り返し行うことでブレを消去する手法を提案した. 本研究ではリサンプリングサイズを変えることで,Hill 推定量で求まる推定値がどのように変化するか調べ, 推定手法の簡素化を図った.

3.2　実験条件

今回の乱数実験の母集合のデータ数 $N = 12000$ に対して 50 回のリサンプリングによるブートストラップ標本を用いた. リサンプリングサイズとしては, $n = 1000$, $n = 2000$, $n = 4000$, $n = 8000$ について実験を行った. リサンプリングにより得えられた値 50 個の平均値を推定値とした.

3.3　t-分布での実験

t-分布乱数

標準正規分布 $N(0,1)$ に従う乱数 X と 自由度 m の χ^2 分布 χ_m^2 に従う乱数 Y を生成し,

$$T = \frac{X}{Y/m}$$

によって, 自由度 m の t-分布に従う実験乱数として使用した. ここで, 標準正規乱数は $[0,1)$ 一様擬似乱数からの Box-Muller 法で, また, 自由度 m の χ^2 乱数は独立な標準正規乱数の m 個の 2 乗和で与えた.

t-分布の場合の実験結果

自由度 $n = \{1, 2, 4, 8, 11, 16, 20\}$ の場合について実験を行った. 先に述べた通り Hill 推定量 $(\frac{1}{\gamma})$ は t-分布の自由度になる. リサンプリングサイズによる推定値の変化を以下に示す.

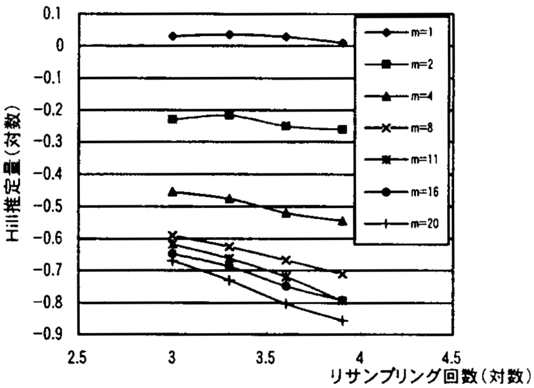

図1 対数プロット

t-分布の場合, リサンプリングサイズが大きくなるにつれて推定値が真値へと対数プロットで直線的に近づいていく様子が見られた. そこで真値へ近づく傾き, すなわち収束のオーダーを調べることでリサンプリング数を減らし, 計算量を少なくした推定方法が可能であると考えられる.

Hill 推定値とサンプリング数の対数の関係をそれぞれの自由度 m について近似した傾きは, 以下のようになった.

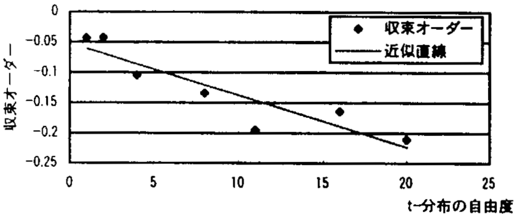

図2 収束オーダーと自由度の関係

本研究では, 裾の重たい分布のもう一つの例として安定分布を扱った.

3.4 安定分布での実験

安定分布乱数

安定分布に従う乱数生成アルゴリズムに Kanter(1975) のアルゴリズムがある. $[-\frac{\pi}{2}, \frac{\pi}{2})$ の一様乱数を V, 平均 1 の指数分布に従う乱数を W とし,V と W は互いに独立であるとする. このとき

$$X = \frac{\sin(\alpha V)}{\cos(V)^{\frac{1}{\alpha}}} \times \left(\frac{\cos(V - \alpha V)}{W}\right)^{\frac{(1-\alpha)}{\alpha}}$$

とおくと, X は特性指数 α 対称安定分布に従う乱数となる.

安定分布の場合の実験結果

$\alpha = \{1.0, 1.2, 1.4, 1.6\}$ の場合について実験を行った. リサンプリングサイズによる推定値の変化を以下に示す.

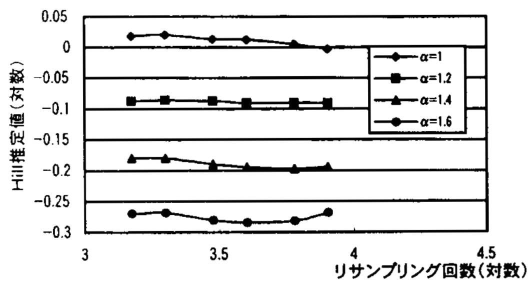

図3 対数プロット

安定分布ではリサンプリング数の違いが推定値に大きく影響をしていない. これは,Hill 推定量が裾が丁度, パレート分布になる場合 (緩慢変動関数 L が定数である場合に相当) の推定量であるためと考えられる. したがって, リサンプリングと共に収束のオーダが一定でないような場合は, 分布の裾指数に付随した緩慢変動関数が定関数でないような場合と考えられる. Hill 推定量では安定分布のパラメータ α は, やはり推定できないと考えられる.

4 実データへの応用例

為替レートの前日比などのような金融商品の分布は正規分布より裾が重たい分布だと知られている.Hill 推定値が収束するようなら緩慢変動関数が定関数ということが予想され, Hill 推定量で裾指数を推定することが可能になる. そこで 1994 年 1 月 1 日から 2006 年 11 月 16 日間の為替レートの前日比での対数分布で実験を試みた.

母集合のデータ 2090 個に対して, リサンプリングサイズ 500, 800, 1000, 1400, 1800 で計算を行ってみた. すると収束のオーダーが一定になるような結果が得られ, パレート分布の裾を持つ分布としての裾指数を推定が可能になる. 実際 t-分布で自由度 17 程度であると推定される.

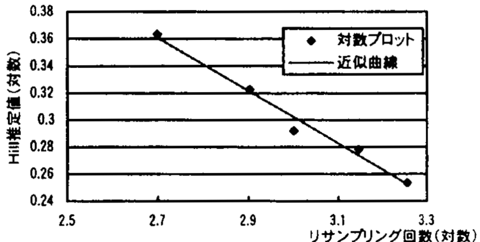

図4 対数プロット

5 まとめ

今回の実験シュミレーションでは Hill 推定値のリサンプリングサイズによる収束のオーダーを調べることで, パレート分布の裾を持つ分布の裾指数が推定が可能であると考えられる. これは, ダブルブートストラップ法より計算量の軽減が可能でもあると言える. また, 裾がパレート分布ではないタイプの安定分布等では, この方法ではパラメータが推定が困難であることも分かった.

参考文献

1) J.Daniellson,L. de Haan,L.Peng,C.G. de Vries," Using a Bootstrap Method to Choose the Sample Franction in Tail Index Estimation",Econometric Institute Report EI 2000-19/A