

自己回帰過程におけるロバスト推定

東京大学教養学部 ハンス ルドルフ クンシュ
チューリッヒ国立工科大学

(1983年5月 受付)

1. はじめに

私の出身のチューリッヒ国立工科大学 (ETH) ではロバストネスが統計学の中で最もさかんな研究分野である。私もいろいろな話をきいてロバストネスを考えはじめた。日本へ来る直前に自己回帰過程のパラメータのロバスト推定についてアイデアが浮かび、これを日本にいる間に研究して、結果を論文 (Künsch [4]) にまとめた。しかし日本ではロバストネスの議論があまりさかんでないから主な結果を technical などところを除いてわかりやすく書いてみようと思う。それには理論より例をあげて説明する方が説得力があるから、シミュレーションの結果でロバスト推定量を最尤推定量と比べることにする。最後に実際のデータの一つの例も議論する。

ロバストネスの数学的に厳密な定義がいくつかあるが、基本概念はデータの分布が僅かに異なったときでも推定量の性質があまり異ならないということである。こうすると「僅かに異なる」と言う意味が最も微妙な問題である。ロバスト統計ではたとえば $(1-\varepsilon)P + \varepsilon Q$ と P が僅かに異なっていることになる。 $(1-\varepsilon)P + \varepsilon Q$ はデータの中に確率 ε で Q の分布をもつ異常値がおこるようなモデル (gross error model) である。さらに異常値ばかりではなく、たとえばデータを丸めるときも分布は僅かに異なってくる。これらを表現するため分布の弱位相が適当である。任意の有界連続関数 f に対して $|\int f(x)P_1(dx) - \int f(x)P_2(dx)|$ が小さければ P_1 と P_2 が僅かに異なると言う。 f が有界だから異常値があっても積分があまり変わらないし、 f が連続だからデータを丸めても大きく変わらない。弱位相はエントロピーの距離より P に近い分布がたくさんあり、ロバストの方法を数学的に議論するのに適している。しかし、ここでは話を簡単にするため位相の問題を避けて、具体的にデータの中に異常値を入れたら推定値がどのくらい変わるかということ調べる。データを丸めるときには最尤推定値も大きく変わらないから、異常値の方が問題として重要である。

ロバストネスの話をするると二つの批判をよく受ける。一つは、経験ある統計家ならば異常値に気づくので修正すれば良いというものである。しかし、ここで主観的な処理をすると推定量にどんな影響を及ぼすかは全然わからない。たとえば異常ではないデータを修正すれば、相関に大抵バイアスが生じる。ロバストな方法でやるときは、異常値があってもなくても推定量の統計的な性質がわかる。その上にロバストな方法は客観的だから、そんなに経験がない人でもルーティンなデータ解析が可能である。

もう一つは異常値をモデリングして尤度法を使った方が良いと言う批判である。しかし時系列の場合には、異常値のいろいろなタイプがあってモデルの数が非常に多くなるので、データの数がそんなに大きくなければ、その方法に問題が起る。良いロバストな方法は、どんなタイプの異常値に対しても安定性をもつから、少なくともデータの特性がまだわからないときに有利である。ロバストな推定値が求まるとあとでデータの怪しいところを調べられるから、次のステップでモデリングがもっと簡単になるだろう。こうみると、ロバストな方法とモデリングは対立するどころかむしろ相補う関係にある。

2. AR-モデルにおけるロバスト推定論

本論文で取り扱うモデルは次のような自己回帰過程である.

$$(2.1) \quad x_{i-\xi} = \sum_{k=1}^p \beta_k (x_{i-k} - \xi) + u_i, \quad u_i \text{ i. i. d. } \sim N(0, \sigma^2),$$

ここに $\theta = (\beta_1, \dots, \beta_p, \xi, \sigma) \in \mathbb{R}^{p+2}$ は未知なパラメータである. (2.1) の定常性を保証するため, 方程式 $1 - \beta_1 z - \dots - \beta_p z^p = 0$ の根が全て $\{z, |z| > 1\}$ に入っていると仮定する. 以下に使用する $p \times p$ 行列 $R(\beta)$ と $A(\beta)$ を次のように定義する.

行列 R の (i, j) 要素は

$$\text{cov}_\theta(X_0, X_{i-j}) / \sigma^2 \quad (1 \leq i, j \leq p)$$

で, A は R の逆行列である.

Künsch [4] で提案したロバストな推定量は次の方程式の解である.

$$(2.2) \quad \sum_{i=1}^{n-p} w_{c_1} \{r(x_i - \hat{\xi}, \dots, x_{i+p-1} - \hat{\xi}, \hat{\beta}) \hat{u}_{i+p} / \hat{\sigma}^2\} (x_{i+h} - \hat{\xi}) \hat{u}_{i+p} = 0$$

$$(h = 0, 1, \dots, p-1),$$

$$(2.3) \quad \sum_{i=1}^{n-p} w_{c_2} (\hat{u}_{i+p} / \hat{\sigma}) \hat{u}_{i+p} = 0,$$

$$(2.4) \quad \sum_{i=1}^{n-p} w_{c_2} (\hat{u}_{i+p} / \hat{\sigma})^2 \hat{u}_{i+p}^2 = (n-p-1) \alpha \hat{\sigma}^2,$$

ここで

$$w_c(x) = \min(1, c/|x|) \quad (x \in \mathbb{R})$$

$$r(z_1, \dots, z_p, \beta) = \left(\sum_{k=1}^p \sum_{j=1}^p z_k A_{kj}(\beta) z_j \right)^{1/2}$$

$$\hat{u}_{i+p} = x_{i+p} - \hat{\xi} - \sum_{j=1}^p \hat{\beta}_j (x_{i+p-j} - \hat{\xi})$$

$$\alpha = (2\pi)^{-1/2} \int x^2 w_{c_2}(x)^2 \exp(-x^2/2) dx$$

であり, c_1 と c_2 は事前に決めなければならない定数である.

上記の方程式は $c_1 = c_2 = \infty$ のとき正規方程式となり, その解は近似的に最尤推定量になるが, c_1 と c_2 が有界のときは x_i, \dots, x_{i+p} のデータが当てはめているモデルに合わなければ 1 より小さな重み (weight) w_{c_1} と w_{c_2} が入る. この重みは次の理由で入っている: $(x_{i+h} - \hat{\xi})$ と u_{i+p} はともに x_{i+h} の有界な関数ではないから, ただ一つの x_i が大きくかわるとき $\sum_i (x_{i+h} - \hat{\xi}) \hat{u}_{i+p}$ と $\sum_i \hat{u}_{i+p}$ がゼロであるため $\hat{\theta}$ が大きく動かなければならない. だから, 極端に言えば最尤推定値がただ一つの異常値データで決まる可能性がある. w_{c_1} と w_{c_2} があるとこんなことがありえないから出て来る推定量のロバストネス (安定性) がすぐれている.

良いロバストな推定量はモデルが正しいときもデータの分布がモデルの分布と僅かに異なっているときも良い性質をもつ. このそれぞれの場合について (2.2) ~ (2.4) の解 $\hat{\theta}^{(*)}$ の性質を以下に述べる.

データ x_1, \dots, x_n が (2.1) のモデルにしたがっているときは $n^{1/2}(\hat{\theta}^{(*)} - \theta)$ が漸近的に平均ゼロで共分散行列 Γ の正規分布になる. この共分散行列 Γ の要素は

$$(2.5) \quad \begin{aligned} \Gamma_{ij} &= \lambda_1(c_1) A(\beta)_{ij} \quad (1 \leq i, j \leq p) \\ \Gamma_{p+1, p+1} &= \lambda_2(c_2) \sigma^2 \left(1 - \sum_{k=1}^p \beta_k\right)^{-2} \\ \Gamma_{p+2, p+2} &= \lambda_3(c_2) \sigma^2 / 2 \\ \Gamma_{ij} &= 0 \quad (\text{その他の } i, j \text{ の場合}) \end{aligned}$$

である。ここに $\lambda_1, \lambda_2, \lambda_3$ は常に 1 より大きくて単調減少で、無限大で 1 に収束する関数である。 $\lambda_1, \lambda_2, \lambda_3$ は一次元の積分として書けるがここでは重要でないので省略する。 λ_1 は $p=1$ のとき Fig. 1 に与えている。 Γ の形から次のことがわかる。 $\hat{\beta}^{(n)}, \hat{\xi}^{(n)}, \hat{\sigma}^{(n)}$ は互いに独立で $\hat{\beta}^{(n)}$ の誤差は c_2 と無関係で最尤推定量の誤差より僅かに大きい、 c_1 を適当にえらべばそれほど悪くはない。 $\hat{\xi}^{(n)}$ と $\hat{\sigma}^{(n)}$ についても同様である。

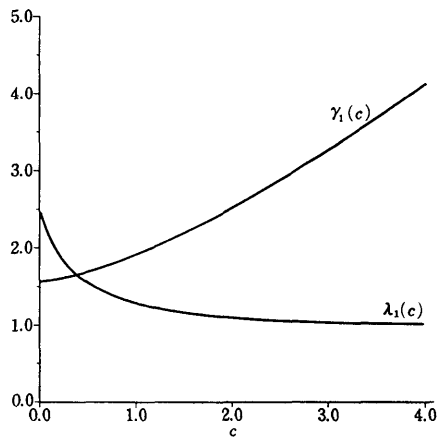


Fig. 1: $p=1$ のときの関数 λ_1 と γ_1 . λ_1 と γ_1 は式 (2.5) と (2.7) で定義されている。
Graph of the functions λ_1 and γ_1 for $p=1$. λ_1 and γ_1 are defined in (2.5) and (2.7).

次に異常値の影響を調べる。このために x_1, \dots, x_n が上と同じ様にモデル (2.1) にしたがっていることを仮定する。さらに y_1, \dots, y_n のデータは $[n(1-\varepsilon)]$ 個分の i では x_i と同じ値であるが、残りの $[\varepsilon n]$ 個の i では任意の値をとる ($[\cdot]$ はガウスの整数記号)。こうすると異常値による $\hat{\theta}^{(n)}$ の変更分、

$$(2.6) \quad \Delta \hat{\theta}^{(n)} = \hat{\theta}^{(n)}(y_1, \dots, y_n) - \hat{\theta}^{(n)}(x_1, \dots, x_n),$$

が小さければ推定量は良いと言える。 n が有限で ε が正のとき $\Delta \hat{\theta}^{(n)}$ は (x_1, \dots, x_n) と (y_1, \dots, y_n) の複雑な関数であるが、まず n を無限大にし、その後 ε をゼロに近づけると次の式が成り立つ。

$$(2.7) \quad \left(\sum_{j=1}^p \sum_{k=1}^p \Delta \hat{\beta}_j^{(n)} R_{jk}(\beta) \Delta \hat{\beta}_k^{(n)} \right)^{1/2} \leq \varepsilon \gamma_1(c_1),$$

$$(2.8) \quad |\Delta \hat{\xi}^{(n)}| \leq \varepsilon \gamma_2(c_2) \sigma \left(1 - \sum_{k=1}^p \beta_k\right),$$

$$(2.9) \quad |\Delta \hat{\sigma}^{(n)}| \leq \varepsilon \gamma_3(c_2) \sigma.$$

ここに $\gamma_1, \gamma_2, \gamma_3$ は単調増加で、無限大で発散する関数である。だからたとえば c_1 を大きくすると $\hat{\beta}$ のロバストネスが減少する。異常値が存在しても $\hat{\beta}$ の性質は c_2 と無関係であることが

(2.7)から明らかである。 $\lambda_1, \lambda_2, \lambda_3$ と同じ様に $\gamma_1, \gamma_2, \gamma_3$ の形がわかっているが省略する。 γ_1 は $p=1$ のときにFig. 1に与えている。

(2.7)で普通のユークリッド距離ではなくて、最尤推定量の漸近共分散の逆行列 $R(\beta)$ で定まった距離を使った理由は二つある。まず $|A\hat{\beta}^{(n)}| \leq \varepsilon \text{const.}$ の様な制約はパラメータシステムに依存するから、異なったパラメータシステムで同じ制約が成り立たない。パラメータシステムが異なると、最尤推定量の漸近共分散行列が $A\hat{\beta}$ と同時に異なってくるから(2.7)はどんなパラメータシステムでも成り立つ。さらに(2.7)の制約があると $A\hat{\beta}^{(n)}$ と異常値がないときの $\hat{\beta}^{(n)}$ の誤差とが比べられる。すなわち漸近共分散行列 Γ の式から

$$(2.10) \quad E \left[\sum_{j,k=1}^p (\hat{\beta}_j^{(n)} - \beta_j) R_{jk}(\beta) (\hat{\beta}_k^{(n)} - \beta_k) \right] \approx n^{-1} p \lambda_1(c_1)$$

言い替えると線形結合 $\eta = \sum_{j=1}^p \eta_j \beta_j$ に対して、「ずれ」と「誤差」について次のような評価が得られる。

$$(2.11) \quad |\hat{\eta}^{(n)}(y) - \hat{\eta}^{(n)}(x)|^2 \leq \varepsilon^2 \gamma_1^2(c_1) / \lambda_1(c_1) \cdot n E [\hat{\eta}^{(n)}(x) - \eta]^2.$$

これは任意の正定値行列 V と任意のベクトル b に対して

$$\sup_x |x^t b|^2 / (x^t V x) = b^t V^{-1} b$$

となることからすぐわかる(詳細はKrasker-Welsch [3] 参照)。同じ理由で(2.8)と(2.9)の $A\hat{\xi}^{(n)}$ と $A\hat{\sigma}^{(n)}$ の制約は β と σ に依存するようにした。

異常値があるときはもちろん $A\hat{\theta}^{(n)}$ をできるだけ小さくしたいから、 c_1 と c_2 が小さければ小さい程良い。しかし、こうすると異常値が存在しないときの推定効率が悪くなる。これらの競合関係のバランスをとらなければならない。もし異常値のパーセンテージ ε がわかっているならば(2.7)と(2.10)から

$$\varepsilon^2 \gamma_1(c_1)^2 = p/n \lambda_1(c_1)$$

になるような選択が自然だろう。しかしわれわれは ε を知らないから λ_1 と γ_1 の関数の変化(Fig. 1を参照)を見た上で次の方法をすすみたい。 $\lambda_1(c_1)$ が1.05と1.15の間に入る様な c_1

Table 1: λ_1 が1.05(cmax)と1.15(cmin)になるような値と対応する γ_1 。 λ_1 と γ_1 は式(2.5)と(2.7)で定義されている。

Values of cmin and cmax and corresponding values of γ_1 , cmin and cmax are defined by $\lambda_1(\text{cmin})=1.15$ and $\lambda_1(\text{cmax})=1.05$, and γ_1, λ_1 are defined in (2.5) and (2.7).

p	cmin	$\gamma_1(\text{cmin})$	cmax	$\gamma_1(\text{cmax})$
1	1.52	2.21	2.52	2.90
2	1.74	2.59	2.89	3.33
3	1.94	2.93	3.20	3.72
4	2.11	3.24	3.49	4.08
5	2.27	3.52	3.75	4.39
6	2.42	3.79	3.99	4.69
7	2.55	4.03	4.22	4.97
8	2.68	4.26	4.44	5.24
9	2.81	4.48	4.65	5.49
10	2.93	4.70	4.85	5.73

Table 2: $c=1.0$ と 1.5 の場合の $\gamma_2, \gamma_3, \lambda_2, \lambda_3$ の値. $\gamma_2, \lambda_2, \gamma_3, \lambda_3$ は式 (2.5), (2.8), (2.9) で定義されている.
 Values of $\gamma_2, \lambda_2, \gamma_3, \lambda_3$ for $c=1.0$ and $c=1.5$. The functions $\gamma_2, \lambda_2, \gamma_3, \lambda_3$ are defined in (2.5), (2.8) and (2.9).

c	$\gamma_2(c)$	$\lambda_2(c)$	$\gamma_3(c)$	$\lambda_3(c)$
1.0	1.46	1.11	1.28	2.12
1.5	1.73	1.04	1.54	1.32

を二つ三つえらんで推定値を比べる. 大きな差があればこの差の原因をデータを見て調べる. ここで 1.05 と 1.15 の値というのはそれぞれある程度任意ではあるが, こうすると最尤推定量よりロバストネスがすぐれて, しかも効率を大きく失わない様な推定量になる. λ_1 が 1.05 と 1.15 になる c_1 の値と対応する γ_1 の値は Table 1 に与えてある. c_2 には i. i. d. の location-scale 推定のときと同じ様に 1.0 と 1.5 の間の値をすすめたい. これに対応する $\gamma_2, \gamma_3, \lambda_2, \lambda_3$ の値は Table 2 に与えてある. $\hat{\sigma}$ の効率はかなり悪いが, σ は大抵 nuisance パラメータだから効率よりロバストネスが重要である. 次の章のシミュレーションの結果によると $\hat{\sigma}$ が異常値に最も鋭敏だから $c_2=1.0$ や 1.5 の値が適当である.

式 (2.2)~(2.4) により決まっている推定量は, (2.7)~(2.9) をみたくしている推定量の全体の中で最尤推定量に最も近いし, 異常値がない場合の $\hat{\beta}$ と $\hat{\xi}$ の漸近共分散を最小化する. このことの証明は容易ではないので省くが Künsch [4] にある. 異常値の影響と漸近共分散をともに表現する関数を見つけることが最も微妙な部分である. この関数は i. i. d. の場合に Hampel [1] が導入して influence function と呼ばれている. (2.4) は $\hat{\sigma}$ に対して最適性 (optimality) がないけれど $p=0$ (独立) の場合はロバスト統計で最もよく使われている推定量 (Huber の “Proposal 2”, Huber [2], p. 137) と同じ推定量になっている. また β, ξ, σ がみなそれぞれ違うタイプのパラメータであることから (2.7)~(2.9) の上限を $\hat{\beta}, \hat{\xi}, \hat{\sigma}$ に対して別々にとった. この三つのパラメータにおける合計のバイアスがある上限より小さくしたければ, 最適な推定量が少し違ってくる. またイノベーション ε_i が, 正規分布以外のゼロに対して対称な分布をもっているときも最適なロバスト推定量がえられるが, その形はさらに複雑になる.

式 (2.2)~(2.4) は反復化の方法で解ける. $\hat{\beta}$ を固定して $\hat{\xi}$ と $\hat{\sigma}$ を (2.3)~(2.4) から計算する. これは普通のロバスト location-scale 推定量になって Huber [2], Variant 4, pp. 147-148 の方法を使う. 次に前のステップでみつけた η と $\hat{\sigma}$ を固定して (2.2) を解く. このためにまた反復化の計算をしなければならない. w_{e_1} の中に出てくる $\hat{\beta}$ をさらに固定すると (2.2) が線形方程式になって簡単に解ける. この解は次のステップで w_{e_1} の中に入れて繰返す. 収束してからまた $\hat{\beta}$ を固定して (2.3)~(2.4) から η と $\hat{\sigma}$ の新しい値を計算する. 経験的に言って収束は非常に速い. ただ初期値 $\hat{\beta}$ で $A(\hat{\beta})$ が正定値行列になるように注意しなければならない.

最尤推定値とロバスト推定値が大きく異なったら, これの原因を知りたい. (2.2)~(2.4) から, 大きな差があれば, ある i で $w_{e_1}(r(x_i - \hat{\xi}, \dots, x_{i+p-1} - \hat{\xi}, \hat{\beta}) \hat{u}_{i+p}/\hat{\sigma}^2)$ や $w_{e_1}(\hat{u}_{i+p}/\hat{\sigma})$ が 1 よりずっと小さくなっていることがわかる. 即ち, x_{i+p} が前のデータ x_i, \dots, x_{i+p-1} に適合しないか, あるいは x_i, \dots, x_{i+p-1} のデータが当てはめたモデルに合わないことのいずれかまたは両方である. これを検討すれば大体データの怪しいところがわかる. しかし, ここでイノベーションの scale σ をロバストな方法で推定しないと $\hat{\sigma}$ が大きすぎて何もわからないことが多い (4章の例を参照).

3. シミュレーションの結果

前の章に述べたロバストな推定量(2.2)~(2.4)の性質は全部漸近的なので、有限の n や正の ε での性質を確かめたい。解析的な計算ができないからシミュレーションの方法しかない。しかし関係する因子をみな変化させると、その場合の数が非常に大きくなる。 n と p と θ の値だけではなくて、推定量のロバストネスを決める c_1 と c_2 , 異常値のパーセンテージ ε , 異常値の大きさ, その場所や符号もある。たとえば異常値が引き続いておこっているか, 散在しておこっているかによって, 推定値に対しての影響が違おうだろう。さらに異常値が続いておこっているときの符号も問題になる。その他に異常値が直接にデータ x_i の中でなく, イノベーション ε_i の所でおこっていることも考えられる。この場合には, 異常値がおこった所の x_i だけではなくてそのあとのいくつかのデータも異常になる。こんなタイプの異常値の場合でも(2.7)~(2.9)の制約が漸的に成り立つが, 有限のところをシミュレーションで確かめたい。

ここではシミュレーションは $p=1$ に限る。 $p>1$ のときは $\hat{\beta}_1, \dots, \hat{\beta}_p$ の相関のために結果が解釈しにくくなるからである。しかし, これは式(2.2)~(2.4)の推定量を $p=1$ に限ると言う意味ではない。 p 以外の因子は次の様に与えた, 異常値がないとき ($\varepsilon=0$) は, β を 0.8 として n を 100, 200 と 400 で与えた。このシミュレーション実験の主な目的は, 異常値の種類によって推定量に及ぼす影響がどういふふうに変るか調べることである。だから異常値が存在しているとき $n=100$ と $\varepsilon=0.05$ を固定して β を 0.8, 0.0 と -0.8 で与えて異常値のタイプ, 場所と符号を変化させた。いつも c_1 を 1.5 と 2.5, また c_2 を 1.0 と 1.5 とおいて四つのロバスト推定値を計算したが, c_2 を動かしても, 漸近的な理論と同様に, $\hat{\beta}$ の変動は確率的で小さいことがシミュレーションで確認されたので, 結果は省略した。 c_1 を動かしているときは $\hat{\xi}$ と $\hat{\sigma}$ についても同様である。

Table 3: 長さ n の異常値がない AR(1) 過程での $\hat{\beta}$ の平均 (上の数字) と標準誤差 (下の数字). シミュレーションの数=20.
Mean (upper figure) and standard error (lower figure) of $\hat{\beta}$ for an AR(1)-process of length n . Number of simulations=20.

c_1	c_2	$n=100$	$n=200$	$n=400$	$n \rightarrow \infty$
∞	∞	.783	.773	.787	.800
		.053	.049	.034	$.060n^{-1/2}$
2.5	1.0	.787	.775	.786	.800
		.054	.052	.037	$.063n^{-1/2}$
1.5	1.0	.790	.778	.787	.800
		.056	.053	.038	$.069n^{-1/2}$

Table 3 に異常値がないときの結果を与えた。Table 3 のために $\beta=0.8$ の AR(1) 過程を長さ 400 で 20 標本発生させて, 初めの 100 個と 200 個のデータも別途に使った。Table 3 では ξ と σ は nuisance パラメータということにして $\hat{\beta}$ の結果しか与えていない。最尤推定量とロバスト推定量の $\hat{\beta}$ の平均と分散はほとんど同じであるが, 漸近的な結果と同様に, c_1 が小さくなると分散が僅かに大きくなる傾向にある。

異常値が存在しているときの結果は Table 4~Table 7 に与えている。 $\beta=0.8$ のときには Table 3 に使ったデータの 20 標本を元にして異常値を入れた。これらの結果の上で $\Delta\hat{\theta}$ の大きさを調べるためにはシミュレーションがそんなに多く必要ないことがわかったので, $\beta=0.0$ と $\beta=-0.8$ のときには二つの標本しか発生させなかった。異常値の場所は確率的でなくて事前に決めたが, 結果は同じだと思う。

Table 4-Table 7 の注釈の前に、異常値の大きさを説明したい。| β |=0.8, $\sigma=1.0$ のときには異常値の大きさを 5.0 で与えた。そうすると異常値は他のデータの標準偏差の三倍になっているが、異常値の所のイノベーションは他の所のイノベーションの五倍程度になる。だから $\beta=0.0$ のときには、異常値の大きさを 3.0 と 5.0 の二つの値で与えた。異常値がこれ以上大きくなっても、最尤推定量のバイアスがどんどん大きくなるだけでロバスト推定量にはほとんど影響がないことがわかっている。たとえば Table 4-Table 7 の $\beta=0.0$ のときのシミュレーションの結果を見よ。

Table 4: 長さ 100 の AR(1) 過程のとき 5 個の異常値の推定値に及ぼす影響。符号が交互する異常値が続く。 $\Delta\beta$ と $\Delta\hat{\sigma}$ は式 (2.6) で定義されている。
Influence of 5 outliers on the parameter estimates of an AR(1)-process of length 100. Consecutive outliers with alternating sign. $\Delta\beta$ and $\Delta\hat{\sigma}$ are defined in (2.6).

		$\Delta\beta$ の			(2.7) の $ \Delta\beta $ の上限 bound of (2.7)	β の平均 mean of β	β の標準誤差 s.e. of β	
c_1	c_2	最 小 minimum	中央値 median	最 大 maximum				
∞	∞	-.652	-.562	-.321	∞	.238	.123	
2.5	1.0	-.132	-.096	-.062	.087	.694	.062	
1.5	1.0	-.113	-.071	-.041	.066	.721	.059	
		$\Delta\hat{\sigma}$ の			(2.9) の $ \Delta\hat{\sigma} $ の上限 bound of (2.9)			
c_1	c_2	最 小 minimum	中央値 median	最 大 maximum				
∞	∞	.766	.872	1.045	∞			
1.5	1.5	.045	.110	.142	.077			
1.5	1.0	.021	.059	.139	.064			
ii) $\beta=0.0, \sigma=1.0$, 異常値の絶対値=3, と 5., シミュレーションの数=2. Absolute value of outliers=3, and 5., number of simulations=2.								
		異常値=±3, $\Delta\beta$		異常値=±5, $\Delta\beta$		(2.7) の $ \Delta\beta $ の上限 bound of (2.7)		
c_1	c_2							
∞	∞	-.308	-.324	-.513	-.556	∞		
2.5	1.0	-.187	-.163	-.184	-.141	.144		
1.5	1.0	-.164	-.103	-.166	-.095	.110		
		$\Delta\hat{\sigma}$			(2.9) の $ \Delta\hat{\sigma} $ の上限 bound of (2.9)			
c_1	c_2							
∞	∞	.143	.168	.329	.344	∞		
1.5	1.5	.040	.084	.039	.083	.077		
1.5	1.0	.044	.079	.044	.077	.064		
iii) $\beta=-.8, \sigma=1.0$, 異常値の絶対値=5., シミュレーションの数=2. Absolute value of outliers=5., number of simulations=2.								
		$\Delta\beta$		(2.7) の $ \Delta\beta $ の上限 bound of (2.7)				
c_1	c_2							
∞	∞	-.008	-.032	∞				
2.5	1.0	-.080	-.064	.087				
1.5	1.0	-.084	-.052	.066				
		$\Delta\hat{\sigma}$			(2.9) の $ \Delta\hat{\sigma} $ の上限 bound of (2.9)			
c_1	c_2							
∞	∞	.186	.203	∞				
1.5	1.5	.009	.028	.077				
1.5	1.0	.015	.025	.064				

Table 4-Table 7 の結果によると, どんな種類の異常値でも $\hat{\sigma}$ に大きな影響を及ぼす. 最尤推定量の $\Delta\hat{\sigma}$ は σ の本当の値の半分か, 同じくらいの量であり, 明白に危険である. ロバストの場合には $\Delta\hat{\sigma}$ がずっと小さくて有利である. 例外もあるが, $\Delta\hat{\sigma}$ は $c_2=1.0$ のときの方が $c_2=1.5$ のときより小さい. シミュレーションに与えた異常値では平均の推定値 $\hat{\xi}$ にはあまり差が出てこなかったので省略する.

$\hat{\beta}$ に及ぼす影響は異常値の種類と β の値によって違う. 異常値の場所が続いて, 符号が交互するとき (Table 4) は, 最尤推定量の $|\Delta\hat{\beta}|$ が $\beta=0.8$ と $\beta=0.0$ の場合には大きい, $\beta=-0.8$

Table 5: 長さ 100 の AR (1) 過程のとき 5 個の異常値の推定値に及ぼす影響. 符号が同じ異常値が続く. $\Delta\hat{\beta}$ と $\Delta\hat{\sigma}$ は式 (2.6) で定義されている.

Influence of 5 outliers on the parameter estimates of an AR (1)-process of length 100. Consecutive outliers with the same sign. $\Delta\hat{\beta}$ and $\Delta\hat{\sigma}$ are defined in (2.6).

i) $\beta=.8, \sigma=1.,$ 異常値の絶対値=5., シミュレーションの数=20. Absolute value of outliers=5., number of simulations=20.		$\Delta\hat{\beta}$ の			(2.7) の $ \Delta\hat{\beta} $ の上限 bound of (2.7)	$\hat{\beta}$ の平均 mean of $\hat{\beta}$	$\hat{\beta}$ の標準誤差 s.e. of $\hat{\beta}$
c_1	c_2	最小 minimum	中央値 median	最大 maximum			
∞	∞	-.102	.010	.047	∞	.782	.054
2.5	1.0	-.009	.061	.098	.087	.838	.039
1.5	1.0	-.002	.056	.101	.066	.843	.043
		$\Delta\hat{\sigma}$ の			(2.9) の $ \Delta\hat{\sigma} $ の上限 bound of (2.9)		
c_1	c_2	最小 minimum	中央値 median	最大 maximum			
∞	∞	.040	.212	.386	∞		
1.5	1.5	-.067	.030	.090	.077		
1.5	1.0	-.044	.024	.105	.064		
ii) $\beta=.0, \sigma=1.,$ 異常値の絶対値=3. と 5., シミュレーションの数=2. Absolute value of outliers=3, and 5., number of simulations=2.		異常値=±3. Outliers=±3.		異常値=±5. Outliers=±5.		(2.7) の $ \Delta\hat{\beta} $ の上限 bound of (2.7)	
c_1	c_2	$\Delta\hat{\beta}$		$\Delta\hat{\beta}$			
∞	∞	.225	.219	.411	.391	∞	
2.5	1.0	.095	.158	.097	.128	.144	
1.5	1.0	.057	.096	.057	.076	.110	
		$\Delta\hat{\sigma}$		$\Delta\hat{\sigma}$		(2.9) の $ \Delta\hat{\sigma} $ の上限 bound of (2.9)	
∞	∞	.124	.122	.295	.301	∞	
1.5	1.5	.083	.071	.093	.074	.077	
1.5	1.0	.076	.070	.076	.067	.064	
iii) $\beta=-.8, \sigma=1.,$ 異常値の絶対値=5., シミュレーションの数=2. Absolute value of outliers=5., number of simulations=2.		$\Delta\hat{\beta}$		$\Delta\hat{\beta}$		(2.7) の $ \Delta\hat{\beta} $ の上限 bound of (2.7)	
c_1	c_2	$\Delta\hat{\beta}$		$\Delta\hat{\beta}$			
∞	∞	.559	.598			∞	
2.5	1.0	.052	.117			.087	
1.5	1.0	.031	.081			.066	
		$\Delta\hat{\sigma}$		$\Delta\hat{\sigma}$		(2.9) の $ \Delta\hat{\sigma} $ の上限 bound of (2.9)	
∞	∞	.927	.783			∞	
1.5	1.5	.110	.082			.077	
1.5	1.0	.061	.027			.064	

の場合にはほとんどゼロである。ロバスト推定量の $|A\hat{\beta}|$ は $\beta=0.8$ と $\beta=0.0$ のときに小さいが、 $\beta=-0.8$ の場合には僅かに大きい。 $A\hat{\beta}$ がこの種類の異常値で負になることは明らかである。

符号が同じ異常値が続くとき (Table 5) は、 $\beta=0.0$ の結果が Table 4 と同様であり、 $\beta=-0.8$ の場合 Table 4 の $\beta=0.8$ と同様である。しかし $\beta=0.8$ の場合に $A\hat{\beta}$ が負になるのは不思議かも知れない。これは、異常値の間の相関が 1.0 であるが異常値とその近くの良いデータとの相関が負になっていることが原因である。

Table 6: 長さ 100 の AR (1) 過程のとき 5 個の異常値の推定値に及ぼす影響。異常値が散在している。 $A\hat{\beta}$ と $A\hat{\sigma}$ は式 (2.6) で定義されている。
Influence of 5 outliers on the parameter estimates of an AR(1)-process of length 100. Isolated outliers. $A\hat{\beta}$ and $A\hat{\sigma}$ are defined in (2.6).

i) $\beta=.8, \sigma=1.$, 異常値の絶対値=5., シミュレーションの数=20. Absolute value of outliers=5., number of simulations=20.		$A\hat{\beta}$ の			(2.7) の $ A\hat{\beta} $ の上限 bound of (2.7)	$\hat{\beta}$ の平均 mean of $\hat{\beta}$	$\hat{\beta}$ の標準誤差 s.e. of $\hat{\beta}$
c_1	c_2	最 小 minimum	中央値 median	最 大 maximum			
∞	∞	-.412	-.282	-.178	∞	.497	.086
2.5	1.0	-.219	-.104	-.058	.087	.681	.068
1.5	1.0	-.181	-.072	-.035	.066	.721	.059
		$A\hat{\sigma}$ の			(2.9) の $ A\hat{\sigma} $ の上限 bound of (2.9)		
c_1	c_2	最 小 minimum	中央値 median	最 大 maximum			
∞	∞	.556	.678	.838	∞		
1.5	1.5	.109	.200	.293	.077		
1.5	1.0	.069	.140	.231	.064		
ii) $\beta=.0, \sigma=1.$, 異常値の絶対値=3. と 5., シミュレーションの数=2. Absolute value of outliers=3. and 5., number of simulations=2.		異常値=±3. Outliers=±3.		異常値=±5. Outliers=±5.		(2.7) の $ A\hat{\beta} $ の上限 bound of (2.7)	
c_1	c_2	$A\hat{\beta}$		$A\hat{\beta}$			
∞	∞	-.041	.062	-.059	.036	∞	
2.5	1.0	-.090	.064	-.119	.057	.144	
1.5	1.0	-.109	.045	-.115	.041	.110	
		$A\hat{\sigma}$		$A\hat{\sigma}$		(2.9) の $ A\hat{\sigma} $ の上限 bound of (2.9)	
∞	∞	.176	.182	.483	.497	∞	
1.5	1.5	.077	.066	.078	.077	.077	
1.5	1.0	.089	.080	.095	.087	.064	
iii) $\beta=-.8, \sigma=1.$, 異常値の絶対値=5., シミュレーションの数=2. Absolute value of outliers=5., number of simulations=2.		$A\hat{\beta}$		(2.7) の $ A\hat{\beta} $ の上限 bound of (2.7)			
c_1	c_2	$A\hat{\beta}$		$A\hat{\beta}$			
∞	∞	.241	.340	∞			
2.5	1.0	.029	.143	.087			
1.5	1.0	.013	.090	.066			
		$A\hat{\sigma}$		(2.9) の $ A\hat{\sigma} $ の上限 bound of (2.9)			
∞	∞	.646	.684	∞			
1.5	1.5	.221	.156	.077			
1.5	1.0	.162	.119	.064			

異常値が散在しているとき (Table 6) は, $\beta=0.8$ と $\beta=-0.8$ の $|4\hat{\beta}|$ が同じくらいでロバスト推定量の方がより小さい. Table 4 と Table 5 の結果とを比べると, 今の種類の異常値の方がロバスト推定量には強い影響を及ぼすが, 最尤推定量のときは逆である. $\beta=0.0$ の場合にはどんな推定量にもあまり害がない. Table 6 では異常値の符号が交互であるが, 符号が同じ場合でも変わらないと思う.

Table 7: 長さ 100 の AR(1) 過程のとき 5 のイノベーション異常値の影響. 異常値の絶対値=5.0. $4\hat{\beta}$ と $4\hat{\sigma}$ は式 (2.6) で定義されている.

Influence of 5 innovation outliers on the parameter estimates of an AR (1)-process of length 100. Absolute value of outliers=5.0. $4\hat{\beta}$ and $4\hat{\sigma}$ are defined in (2.6).

i) $\beta=.8, \sigma=1.,$ シミュレーションの数=20. Number of simulations=20.		$4\hat{\beta}$ の			(2.7) の $ 4\hat{\beta} $ の上限 bound of (2.7)	$\hat{\beta}$ の平均 mean of $\hat{\beta}$	$\hat{\beta}$ の標準誤差 s.e. of $\hat{\beta}$
c_1	c_2	最 小 minimum	中央値 median	最 大 maximum			
∞	∞	-.166	-.014	.160	∞	.756	.058
2.5	1.0	-.093	.007	.142	.087	.789	.052
1.5	1.0	-.093	.020	.101	.066	.793	.053
		$4\hat{\sigma}$ の			(2.9) の $ 4\hat{\sigma} $ の上限 bound of (2.9)		
c_1	c_2	最 小 minimum	中央値 median	最 大 maximum			
∞	∞	.372	.476	.521	∞		
1.5	1.5	.043	.097	.130	.077		
1.5	1.0	.000	.065	.173	.064		
ii) $\beta=-.8, \hat{\sigma}=1.,$ シミュレーションの数=2. Number of simulations=2.					(2.7) の $ 4\hat{\beta} $ の上限 bound of (2.7)		
c_1	c_2	$4\hat{\beta}$					
∞	∞	-.011	-.002	∞			
2.5	1.0	-.000	-.010	.087			
1.5	1.0	-.011	.012	.066			
					(2.9) の $ 4\hat{\sigma} $ の上限 bound of (2.9)		
c_1	c_2	$4\hat{\sigma}$					
∞	∞	.498	.492	∞			
1.5	1.5	.096	.068	.077			
1.5	1.0	.072	.048	.064			

イノベーションの中の異常値のとき (Table 7) は, どんな推定量にとっても $|4\hat{\beta}|$ がそんなに大きくないし, 最尤推定量とロバスト推定量の差も小さい. イノベーションが正規分布でなくても $\hat{\beta}$ の漸近的な分布が変わらないことは良く知られている. Table 7 はこれを反映している.

Table 4-Table 7 の結果を次の様に要約する. 最尤推定量の $\hat{\sigma}$ にはどんな種類の異常値でも大きな影響を及ぼすし, 種類によっては $\hat{\beta}$ に及ぼす影響も大きい. ロバスト推定量の場合には, $|4\hat{\beta}|$ と $|4\hat{\sigma}|$ が (2.7) と (2.9) の上限の二~三倍大きくなることもあるけれど最尤推定量よりずっと安定している. だから, ここで使った c_1 と c_2 の値で決まるロバスト推定量をすすめる.

4. 実際のデータの例

ここで議論するデータはある銀行の91ヶ月の間の毎月の貸利率である。データの経過は Fig. 2 に与えてある。そのままのデータに AR(p) 過程 ($p=1, 2$) を当てはめると Table 8 の

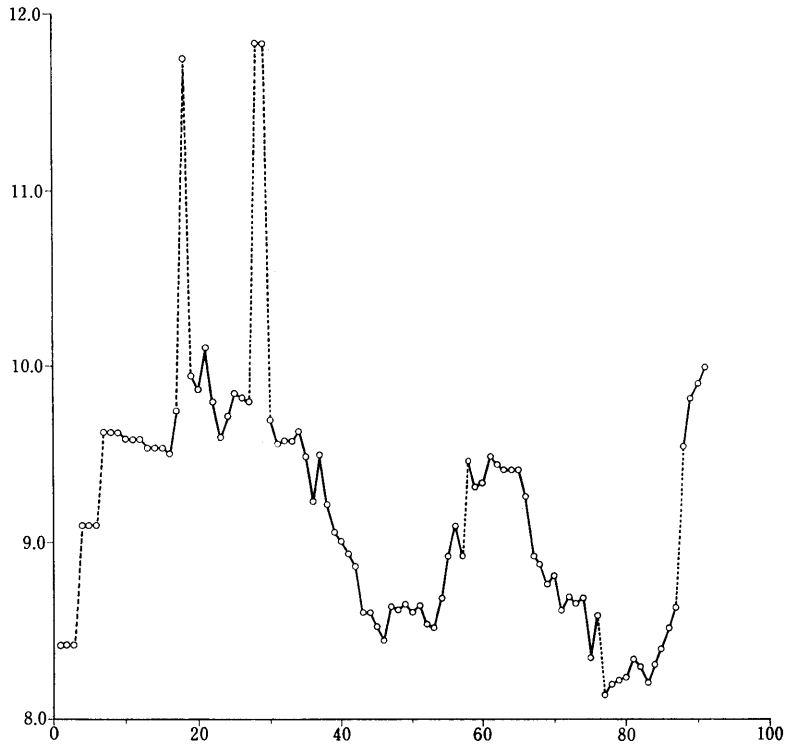


Fig. 2: 91 か月の間の毎月の貸利率のデータ。点線は異常値や怪しいところを表わす。
Monthly lending rate during 91 months. The dotted lines indicate doubtful parts in the data.

Table 8: Fig. 2 のデータに AR(p) 過程を当てはめたときのパラメータ推定値。
Parameter estimates obtained by fitting an AR(p)-process to the data of Fig. 2.

$p=1$:

c_1	c_2	β	ξ	$\hat{\sigma}$
∞	∞	.789	9.19	.443
2.5	1.5	.958	9.18	.154
2.5	1.0	.959	9.11	.133
1.5	1.5	.959	9.18	.154
1.5	1.0	.959	9.11	.133

$p=2$:

c_1	c_2	β_1	β_2	ξ	$\hat{\sigma}$
∞	∞	.766	.030	9.19	.442
2.5	1.5	.978	-.025	9.19	.154
2.5	1.0	.994	-.032	9.14	.126
1.5	1.5	1.000	-.035	9.23	.153
1.5	1.0	1.003	-.031	9.16	.124

推定値が出てくる。 $p=1$ と $p=2$ の推定値を比べると、 c_1 と c_2 がどんな値の場合でも $p=1$ が良いだろうと見当がつく。しかし最尤推定の $\hat{\sigma}$ と $\hat{\beta}$ とロバストな $\hat{\sigma}$ と $\hat{\beta}$ の値は大きく違う。データを見れば三つの異常値の存在がすぐわかる。 $x_{18}=11.75$ と $x_{28}=x_{29}=11.84$ は他のデータよりずっと大きい。法律の改正のあと銀行が利率の政策を一時変えたことが原因である。

Table 9: Fig. 2 のデータで三つの異常値を修正したあとのパラメータ推定値。
Parameter estimates for the data of Fig. 2 after correction of the three large outliers.

$p=1$:

c_1	c_2	$\hat{\beta}$	$\hat{\xi}$	$\hat{\sigma}$
∞	∞	.923	9.12	.212
2.5	1.5	.958	9.21	.136
2.5	1.0	.959	9.15	.126
1.5	1.5	.963	9.22	.136
1.5	1.0	.964	9.16	.125

この三つの異常値を修正するため、その近くのデータと同じ値を与えて $x_{18}=x_{28}=x_{29}=9.85$ とおいてみよう。そうすると、当てはめた AR(1) 過程のパラメータ推定値は Table 9 のとおりになる。ロバスト推定値は異常値を修正しても、ほとんど変わらず、最尤推定値はロバスト推定値に近づいてきた。しかしまだ $\hat{\sigma}=0.212$ と $\hat{\sigma}=0.136$ の大きな差が残っているから、データの中にその修正した異常値以外にも怪しいところがあると考えられる。それを調べるため、第2章に述べた方法を使って $|\hat{u}_i|$ と $|(x_i - \hat{\xi})\hat{u}_{i+1}|$ の大きな値をさがす。修正したデータを使うから、前の異常値のところには目立つことがないが、それ以外に $\hat{u}_4=0.654$, $\hat{u}_7=0.528$, $\hat{u}_{58}=0.532$, $\hat{u}_{77}=-.470$ と $\hat{u}_{88}=0.891$ が出てくる。Fig. 2 ではこの五か所や前の異常値の所で点線を使っている。他の \hat{u}_i は $\sigma=0.13$ や 0.14 の仮説と合っている。 -0.357 と -0.333 の二つの値以外は $|\hat{u}_i|$ がみな 0.3 より小さい。 $\hat{\sigma}=0.212$ の推定値を使うと \hat{u}_{87} しか目立たないし、 $\hat{\sigma}=0.443$ の推定値を使うと何にも目立たないから σ のロバスト推定値を求めなければならない。

この五か所で本当に異常が起きているかどうかを決めるためには、この銀行と経済の事情とをもっと詳しく知らなければならない。いずれにしても、ロバスト推定のためにイノベーションの特性がわかった。修正したところとその次のイノベーションは使えないから、85のイノベーションが残っている。その85の中の80のイノベーションは 0.13 の標準偏差を持ち、他の五つのイノベーションは 0.13 の 3.5 倍から 7 倍までになっている。

謝 辞

この研究は統計数理研究所での特別研究員として行ったものです。統計数理研究所の厚遇と様々な議論の機会を心から感謝いたします。また日本での研究の機会を与えて下さった学術振興会にも感謝いたします。4章のデータをくださいましたウィーン大学の Wolfgang Polasek さんと、特にこの論文の最初の原稿を見ていただいた尾形良彦さんにお礼を申し上げます。

参 考 文 献

- [1] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *J. Amer. Statist. Ass.*, **69**, 383-394.

- [2] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.
- [3] Krasker, W.S. and Welsch, R.E. (1982). Efficient bounded influence regression estimation, *J. Amer. Statist. Ass.*, **77**, 595-604.
- [4] Künsch, H. (1983). The influence function and optimal robust estimators for time series, submitted to *Ann. Statist.*

Robust Estimation for Autoregressive Processes

Hans Rudolf Künsch

(Tokyo University, College of General Education
and ETH Zurich)

In Künsch (1983, submitted to *Ann. Statist.*) an optimal robust estimator for the parameters of an autoregressive process was constructed. Here we describe this estimator and discuss its properties avoiding the technical difficulties of the subject. The estimator is defined as the solution of the equations (2.2)–(2.4) where A is the inverse of the covariance matrix of p consecutive observations and c_1, c_2 are two constants regulating the robustness of the estimator. Under the model, the asymptotic covariance Γ of the estimator is larger by a factor λ compared with maximum likelihood (Formula (2.5)), but this is compensated by a greater stability in the presence of outliers. Denoting by y_1, \dots, y_n a sample where an ε -percentage of good data x_1, \dots, x_n has been replaced by arbitrary values, the resulting bias $\Delta \hat{\theta}$ is asymptotically bounded as stated in formulas (2.7)–(2.9).

Simulation experiments with AR(1)-processes show how different kinds of outliers can change the maximum likelihood estimates drastically. In all cases the estimate for the variance of the innovations is increased while the estimate for the autoregressive parameter is seriously affected by certain kinds of outliers. Compared with this, the robust estimates are much better although the bounds (2.7)–(2.9) are often exceeded. Finally an example of real data containing several outliers is analyzed.