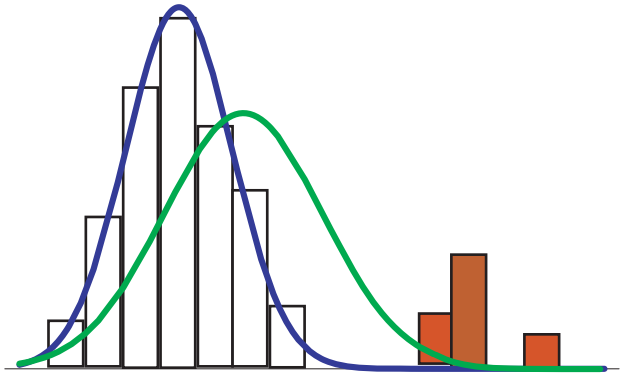
ロバスト推測の基礎と
ダイバージェンス型メソッドへの発展

藤澤 洋徳

統計数理研究所

fujisawa@ism.ac.jp



Contents

1．外れ値とロバスト推測

外れ値とは何であるのか？
ロバスト推測とは何であるのか？

ある実験をしていて，以下のような10個のデータ値が得られたとする：

$$5.6,\ 5.7,\ 5.4,\ 5.5,\ 5.8,\ 5.5,\ 5.3,\ 5.6,\ 5.4,\ 5.2.$$

このデータの**標本平均は5.5**である：

$$\frac{5.6+5.7+5.4+5.5+5.8+5.5+5.3+5.6+5.4+5.2}{10}=5.5.$$

何らかの原因で，**外れ値（異常値，outlier）**を観測することがある．
最後のデータ値がタイプされる時に，誤って5が二回タイプされてしまったとする．

$$5.6,\ 5.7,\ 5.4,\ 5.5,\ 5.8,\ 5.5,\ 5.3,\ 5.6,\ 5.4,\ 55.2.$$

このときの**標本平均は10.5**になる．

$$\frac{5.6+5.7+5.4+5.5+5.8+5.5+5.3+5.6+5.4+55.2}{10}=10.5.$$

5.6, 5.7, 5.4, 5.5, 5.8, 5.5, 5.3, 5.6, 5.4, 5.2.
5.6, 5.7, 5.4, 5.5, 5.8, 5.5, 5.3, 5.6, 5.4, 55.2.

外れ値の混入 ： 5.2 → 55.2.
標本平均 ： 5.5 → 10.5.

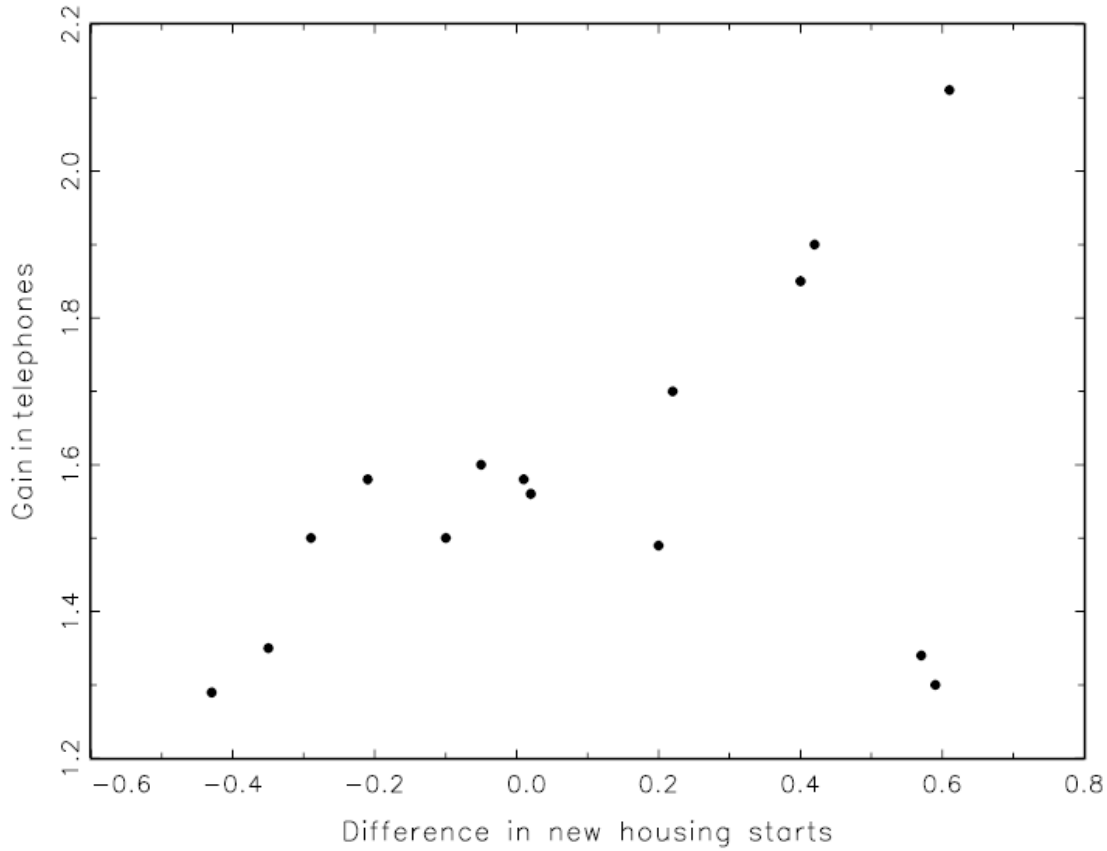
何が起きているのか？

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n} \quad \to \infty \qquad \text{as } x_n \to \infty$$

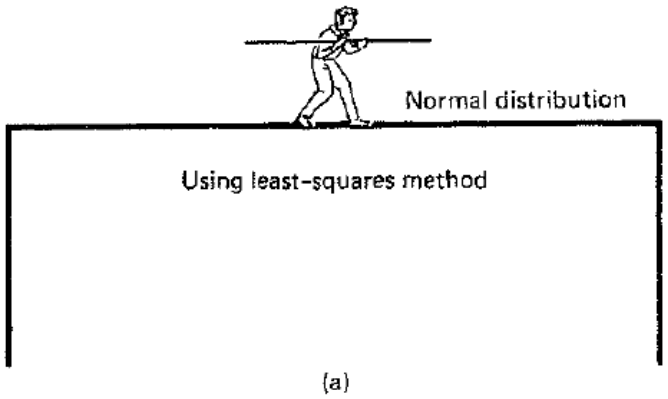
標本平均10.5は，本来の意図するところを見ているとは，とても言えない．
データに外れ値が入っている場合には，何らかの対策が必要となりやすい．

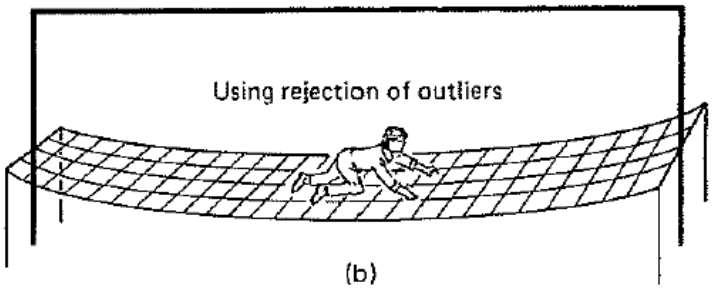
外れ値に影響されにくい推定を**ロバスト推定(robust estimation)**といい，外れ値に影響されにくい統計的推測全体を**ロバスト推測(robust inference)**という．

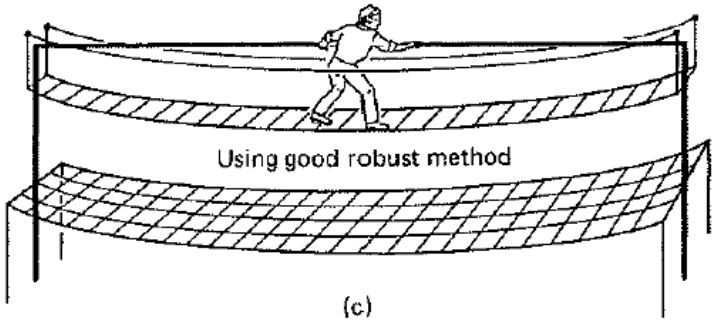
注意： なお，ロバスト推測という言葉は，広義の意味では，外れ値に限らず，何らかの悪影響に引きずられにくい統計的推測全体のことを言う．



$\hat{\rho} = 0.44$ $\hat{\rho}^* = 0.91$ $\hat{\rho}_{rob} = 0.85$







Hampel et al. (1986)

2．簡単なロバスト推定

外れ値にロバストである推定方法について，幾つかの簡単な例を挙げる．

平均μのロバスト推定

5.6, 5.7, 5.4, 5.5, 5.8, 5.5, 5.3, 5.6, 5.4, 55.2.

まずは平均μをロバスト推定することを考えよう．
最も単純で汎用的に用いられている推定値は，中央値 (median) である．

データ$x_1, \ldots, x_n$を小さい順に並び変えた順序統計値を以下で表す：

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

このとき，中央値は，次で定義される：

$$\begin{aligned} \mathrm{Med}(\{x_i\}) &= x_{(k)} && n = 2k-1 \text{のとき} \\ &= (x_{(k-1)} + x_{(k)})/2 && n = 2k \text{のとき} \end{aligned}$$

タイプミスのある例：

$$5.6,\ 5.7,\ 5.4,\ 5.5,\ 5.8,\ 5.5,\ 5.3,\ 5.6,\ 5.4,\ 55.2.$$

順序統計値に並び変える：

$$5.3,\ 5.4,\ 5.4,\ 5.5,\ 5.5,\ 5.6,\ 5.6,\ 5.7,\ 5.8,\ 55.2.$$

中央値は次になる：

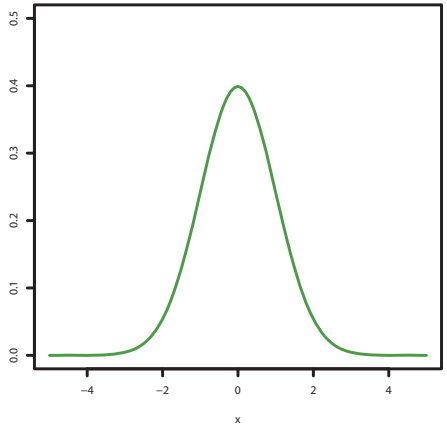
$$\mathrm{Med}(\{x_i\}) = (5.5 + 5.6)/2 = 5.55$$

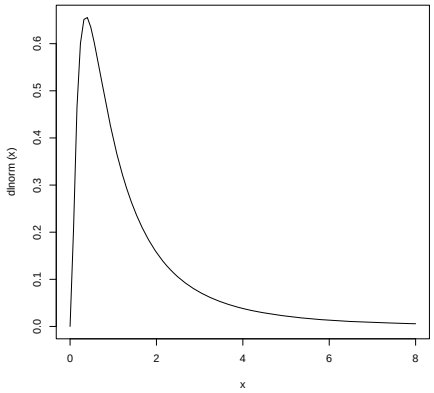
これは平均μの妥当な推定値であろう．

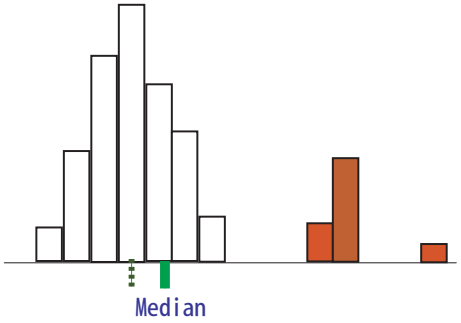
ただし，中央値を使うときには，いくつかの注意が必要である．

データの背後にある母集団は
対称分布である．

外れ値の割合が大きい場合は，
中央値は平均の妥当な推定値ではない．







標準偏差σのロバスト推定

5.6, 5.7, 5.4, 5.5, 5.8, 5.5, 5.3, 5.6, 5.4, 5.2.
5.6, 5.7, 5.4, 5.5, 5.8, 5.5, 5.3, 5.6, 5.4, 55.2.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\sqrt{S^2} = 0.185 \qquad \sqrt{S^2} = 15.7$$

$$\hat{\sigma}_{rob}(= \mathrm{MADN}) = 0.222$$

平均μの他の簡単なロバスト推定法

刈り込み平均(trimmed mean)

$$x_{(1)} \leq \cdots \leq x_{[n\alpha]} \leq \cdots \leq x_{n-[n\alpha]} \leq \cdots \leq x_{(n)}$$

上側100α%と下側100α%のデータを使わない平均.

$$\hat{\mu} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}$$

ただし$[a]$はaを超えない整数.

重み付きスコアに基づいたロバスト推定

母集団が$N(\mu, \sigma^2)$に従う場合を考える．
簡単のために標準偏差σは既知であるとする．
平均パラメータμを推定する問題を考えよう．

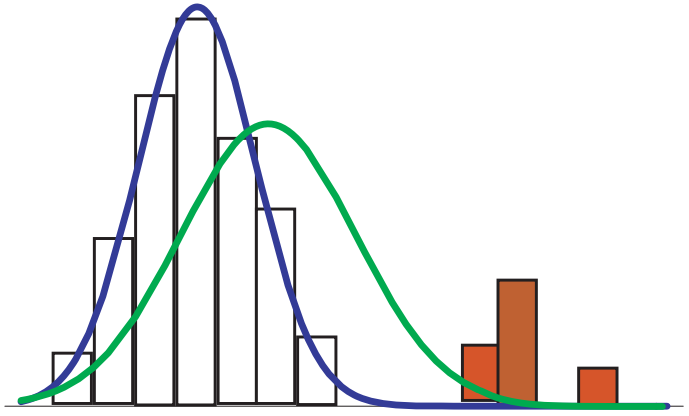
最尤推定量 尤度方程式

$$\sum_{i=1}^{n}(x_i - \mu) = 0 \qquad (\mu = \bar{x})$$

平均μをロバスト推定することを考えよう．
正規分布の密度関数を$\phi(x;\mu,\sigma^2)$で表す．

$$\phi(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

外れ値の典型的な性質の一つは，その生起確率が小さいことである．
外れ値x^*に対して$\phi(x^*;\mu,\sigma^2)$は小さくなるだろう．



尤度方程式において，外れ値の寄与を小さくするために，平均μの推定値$\hat{\mu}$を，次の推定方程式の解として与えることにしよう：

$$\sum_{i=1}^{n} \phi(x_i; \mu, \sigma^2)(x_i - \mu) = 0.$$

最初の例で，タイプミスがあった場合に，上記の推定方法を適用してみる．ただし，標準偏差σは，外れ値を除外したデータの標本標準偏差の近似値0.158であったと仮定してみる．そのとき，推定値は，$\hat{\mu} = 5.517$となった．

疑問

もしもσ^2が未知だったら？

もしも正規分布（対称分布）でなくて一般の分布$f(x;\theta)$だったら？

第二部で関連方法が提案される．

3. 重み付きスコアに基づいたロバスト法と関連したダイバージェンス

本節と次節では，データxに対して，何らかのパラメトリックモデル$f_\theta(x) = f(x;\theta)$が想定されている．

Basu et al. (1998, Biometrika)

Eguchi and Kano (1998, unpublished)

Jones et al. (2001, Biometrika)

平均パラメータ μ の重み付き推定方程式を思い出そう：

$$\sum_{i=1}^{n} \phi(x_i; \mu, \sigma^2)(x_i - \mu) = 0$$

コアはスコア関数だと考えよう：

$$\sum_{i=1}^{n} s(x_i; \theta) = 0 \qquad s(x; \theta) = \frac{d}{d\theta} \log f(x; \theta)$$

重み付き推定方程式を考えてみよう：

$$\sum_{i=1}^{n} f(x_i; \theta)^{\beta} s(x_i; \theta) = 0 \qquad \beta > 0$$

正規分布における平均パラメータの推定の場合は，このままで良かったが，一般のパラメータを推定する場合には，もう一工夫が必要である．

重み付き推定方程式：

$$\sum_{i=1}^{n} f(x_i;\theta)^{\beta} s(x_i;\theta) = 0.$$

推定方程式の不偏性を思い出す：

$\mathsf{E}_{f_\theta}\left[f(x;\theta)^{\beta} s(x;\theta)\right] = 0$ **とは限らない．**

（正規分布での平均μの推定の時は自動的に満たされていた．）

核の関数を修正する：

$$\psi(x;\theta) = f(x;\theta)^{\beta} s(x;\theta) - \mathsf{E}_{f_\theta}\left[f(x;\theta)^{\beta} s(x;\theta)\right]$$

これによって推定方程式の不偏性が満たされる： $\mathsf{E}_{f_\theta}[\psi(x;\theta)] = 0.$

Ｍ推定：

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i;\theta)$$

重み付きスコアに基づいた推定方程式：

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}\left\{ f(x_i;\theta)^{\beta} s(x_i;\theta) - \mathsf{E}_{f_\theta}\left[f(x;\theta)^{\beta} s(x;\theta)\right]\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta)^{\beta} s(x_i;\theta) - \int f(x;\theta)^{1+\beta} s(x;\theta) dx \end{aligned}$$

尤度方程式とKLダイバージェンスの関係を思い出そう：

$$\hat{\theta} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \theta)$$

これを微分すると尤度方程式が得られる：

$$0 = \frac{1}{n} \sum_{i=1}^{n} s(x_i; \theta)$$

右辺を積分して戻せば元に戻る：

$$\frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \theta)$$

これの極限にマイナスをつけたものを考える（**相互エントロピー**）.

相互エントロピー：

$$d_{KL}(g, f_\theta) = -\mathbf{E}_g[\log f(x;\theta)] = -\int g(x) \log f(x;\theta) dx$$

KL ダイバージェンスは次のように表現できる：

$$\begin{aligned} D_{KL}(g, f) &= \mathbf{E}_g \left[\log \frac{g(X)}{f(X)}\right] = \mathbf{E}_g \left[\log g(X)\right] - \mathbf{E}_g \left[\log f(X)\right] \\ &= -d_{KL}(g, g) + d_{KL}(g, f) \end{aligned}$$

ダイバージェンスの性質　(≈ 距離)

$$D_{KL}(g, f) \geq 0. \qquad D_{KL}(g, f) = 0 \;\Leftrightarrow\; g = f.$$

β-ダイバージェンス

重み付きスコアに基づく推定方程式から，
相互エントロピーとダイバージェンスを作る：

$$0 = \frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta)^{\beta} s(x_i;\theta) - \int f(x;\theta)^{1+\beta} s(x;\theta) dx$$

KLのときと同様に積分して極限を取ってマイナスをつける．
β-相互エントロピー：

$$d_{\beta}(g, f_{\theta}) = -\frac{1}{\beta}\int g(x) f(x;\theta)^{\beta} dx + \frac{1}{1+\beta}\int f(x;\theta)^{1+\beta} dx$$

β-ダイバージェンス (β-divergence, density power divergence)：

$$\begin{aligned} D_\beta(g,f) &= -d_\beta(g,g) + d_\beta(g,f) \\ &= \frac{1}{\beta(1+\beta)}\int g(x)^{1+\beta}dx - \frac{1}{1+\beta}\int g(x)f(x)^\beta dx \\ &\quad + \frac{1}{1+\beta}\int f(x)^{1+\beta}dx \end{aligned}$$

極限を考えるとKLダイバージェンスになる：

$$\lim_{\beta\to 0} D_\beta(g,f) = D_{KL}(g,f).$$

γ-ダイバージェンス

重み付きスコアに基づく推定方程式を見直す：

$$0 = \frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta)^{\gamma} s(x_i;\theta) - \mathbf{E}_{f_\theta}\left[f(x;\theta)^{\gamma} s(x;\theta)\right]$$

重みを１にする：

$$0 = \frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta)^{\gamma} s(x_i;\theta) \Big/ \frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta)^{\gamma} - \mathbf{E}_{f_\theta}\left[f(x;\theta)^{\gamma} s(x;\theta)\right] \Big/ \mathbf{E}_{f_\theta}\left[f(x;\theta)^{\gamma}\right]$$

先ほどと同じように行うと，対応するγ-相互エントロピーが得られる：

$$d_\gamma(g, f_\theta) = -\frac{1}{\gamma}\log\int g(x) f(x;\theta)^{\gamma} dx + \frac{1}{1+\gamma}\log\int f(x;\theta)^{1+\gamma} dx$$

γ-ダイバージェンス：

$$
\begin{aligned}
D_\gamma(g,f) &= -d_\gamma(g,g) + d_\gamma(g,f) \\
&= \frac{1}{\gamma(1+\gamma)} \log \int g(x)^{1+\gamma} dx - \frac{1}{1+\gamma} \log \int g(x) f(x)^\gamma dx \\
&\qquad + \frac{1}{1+\gamma} \log \int f(x)^{1+\gamma} dx
\end{aligned}
$$

極限を考えるとKLダイバージェンスになる：

$$
\lim_{\gamma \to 0} D_\gamma(g,f) = D_{KL}(g,f).
$$

二つの相互エントロピーの違い

β-相互エントロピー：

$$d_\beta(g,f) = -\frac{1}{\beta}\int g(x)f(x)^\beta dx + \frac{1}{1+\beta}\int f(x)^{1+\beta}dx$$

γ-相互エントロピー：

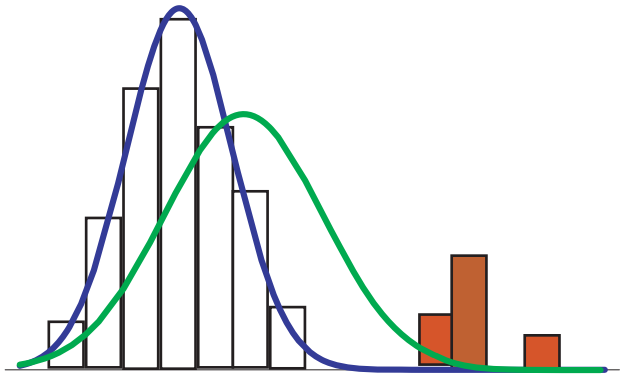
$$d_\gamma(g,f) = -\frac{1}{\gamma}\log\int g(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\log\int f(x)^{1+\gamma}dx$$

注意： 二つの相互エントロピーの違いは「$\log$」が付いているかどうかである．
これだけの違いだが大きな性質の違いも生み出す．

4. γ-ダイバージェンスに基づいたロバスト法の性質

Fujisawa and Eguchi (2008). Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis, Vol.99, 2053-2081.

外れ値とは何か



データを発生している分布（汚染された分布）

$$g(x) = (1-\varepsilon)f(x) + \varepsilon\delta(x)$$

f: **目的分布** $(= f_{\theta^*})$

δ: **外れ値の分布（汚染分布）**

ε: **外れ値の割合**

外れ値の分布が目的分布の裾にある：

$$\nu_f = \int f(x)\delta(x)dx \approx 0 \qquad \left(\left\{ \int f(x)^{\gamma_0}\delta(x)dx \right\}^{1/\gamma_0} \approx 0 \quad \gamma_0 > \gamma > 0 \right)$$

唯一つの本質的な仮定

外れ値の分布が目的とする分布の裾にある：

$$\nu_f = \left\{ \int f(x)^{\gamma_0} \delta(x) dx \right\}^{1/\gamma_0} \approx 0$$

外れ値を x^* として，外れ値の分布を $\delta(x) = \delta_{x^*}(x)$ とすると，

$$\nu_f = f(x^*) \approx 0$$

となるので，想定された条件は，外れ値の生起確率が小さいことを意味する．

Review: 経験推定

ディラック関数　$\delta_a(x)$

$$\int f(x)\delta_a(x)dx = f(a)$$

経験密度関数　**(厳密には経験分布関数で議論すべき)**

$$\bar{g}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x) \;\rightarrow\; g(x)$$

経験推定

$$\mathsf{E}_g[h(X)] = \int g(x)h(x)dx \;\leftarrow\; \int \bar{g}(x)h(x)dx = \frac{1}{n}\sum_{i=1}^{n}h(x_i)$$

γ-相互エントロピー：

$$d_\gamma(g, f) = -\frac{1}{\gamma}\log\int g(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\log\int f(x)^{1+\gamma}dx$$

γ相互エントロピーの経験推定

$$\begin{aligned} d_\gamma(\bar{g}, f) &= -\frac{1}{\gamma}\log\int \bar{g}(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\log\int f(x)^{1+\gamma}dx \\ &= -\frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)^\gamma\right) + \frac{1}{1+\gamma}\log\int f(x)^{1+\gamma}dx \end{aligned}$$

(γ-)ロバスト推定量

$$\begin{aligned}\hat{\theta}_\gamma &= \arg\min_{\theta} d_\gamma(\bar{g}, f_\theta) \\ &\left(= \arg\min_{\theta} \left\{ -d_\gamma(\bar{g}, \bar{g}) + d_\gamma(\bar{g}, f_\theta) \right\} = \arg\min_{\theta} D_\gamma(\bar{g}, f_\theta) \right) \\ &\rightarrow \theta^*_\gamma = \arg\min_{\theta} d_\gamma(g, f_\theta)\end{aligned}$$

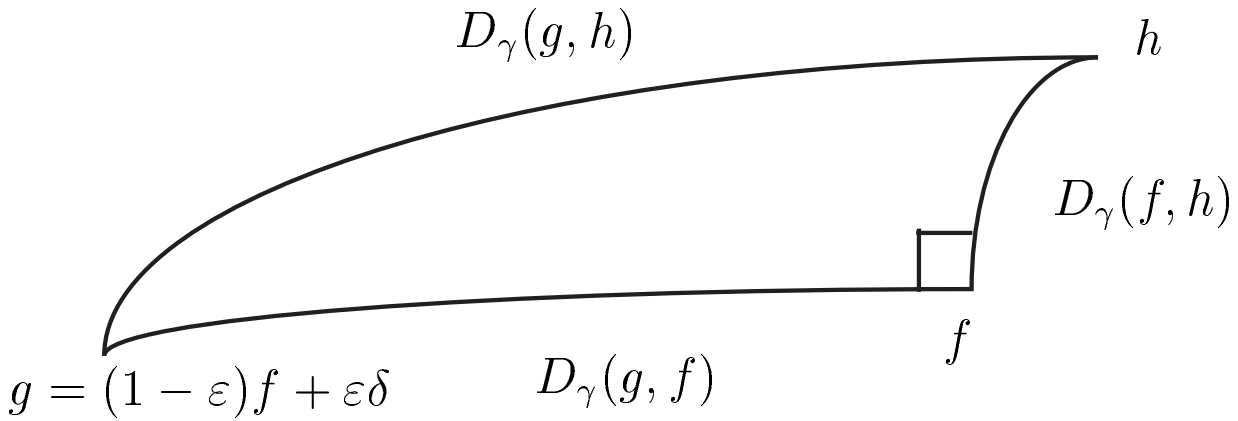
潜在的なバイアス

$$\theta^*_\gamma - \theta^* \quad (\approx 0) \quad = O\left(\varepsilon \nu^\gamma\right)$$

ピタゴリアン関係

$$D_\gamma(g, f) = -d_\gamma(g, g) + d_\gamma(g, f)$$

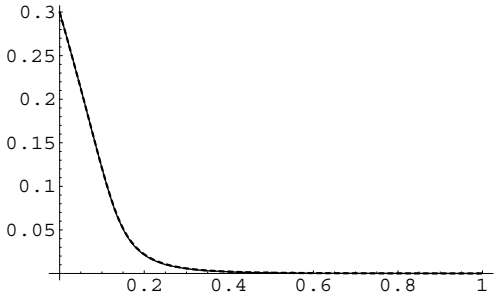
$$D_\gamma(g, f_\theta) = D_\gamma(g, f) + D_\gamma(f, f_\theta) + O(\varepsilon\nu^\gamma)$$



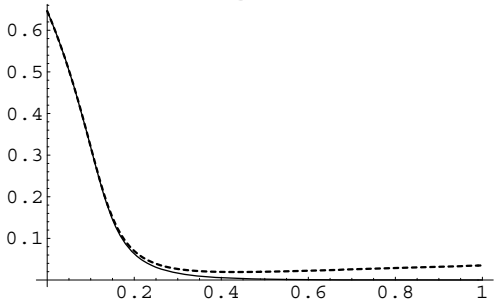
潜在的なバイアス $(\varepsilon = 0.05)$

$$g = (1-\varepsilon)N(0,1) + \varepsilon N(6,1) \quad \leftarrow \quad f_\theta = N(\mu, \sigma^2)$$

μ



σ



θ^*_γ: 実線 $d_\gamma(g,f)$

$\theta^{(m)}_\beta$: 点線 $d_\beta(g,f)$

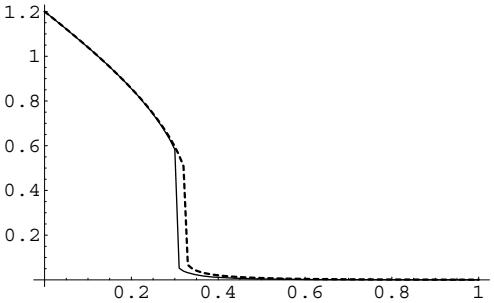
x軸: γ or β

y軸: 潜在バイアス

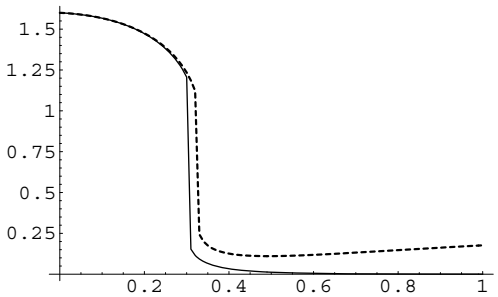
潜在的なバイアス $(\varepsilon = 0.2)$

$$g = (1-\varepsilon)N(0,1) + \varepsilon N(6,1) \quad \leftarrow \quad f_\theta = N(\mu, \sigma^2)$$

μ



σ



θ^*_γ: 実線 $d_\gamma(g, f)$

$\theta^{(m)}_\beta$: 点線 $d_\beta(g, f)$

x軸: γ or β

y軸: 潜在バイアス

更新アルゴリズム

パラメトリック分布を $N(\mu, \sigma^2)$ としたときは，次の繰り返しアルゴリズムの収束値によって推定値が得られます： $\theta = (\mu, \sigma^2)$.

$$
\begin{aligned}
w_i^{(a)} &= f(x_i; \theta^{(a)})^\gamma \Big/ \sum_{i=1}^{n} f(x_i; \theta^{(a)})^\gamma \\
\mu^{(a+1)} &= \sum_{i=1}^{n} w_i^{(a)} x_i \\
(\sigma^2)^{(a+1)} &= \left\{ \sum_{i=1}^{n} w_i^{(a)} x_i^2 - (\mu^{(a+1)})^2 \right\} (1+\gamma)
\end{aligned}
$$

このアルゴリズムは次のような**単調性**をもちます：

$$
d_\gamma(\bar{g}, f_{\theta^{(a)}}) \geq d_\gamma(\bar{g}, f_{\theta^{(a+1)}}) \geq \cdots \geq d_\gamma(\bar{g}, f_{\hat{\theta}_\gamma})
$$

漸近分散

漸近正規性：

$$\sqrt{n}\left(\hat{\theta}_\gamma - \theta^*_\gamma\right) \xrightarrow{d} N\left(0, \Sigma_g(\theta^*_\gamma)\right) \qquad \Sigma_g(\theta) = J_g(\theta)^{-1} I_g(\theta) J_g(\theta)'^{-1}$$

$$J_g(\theta) = \mathrm{E}_g\left[\frac{d}{d\theta'}\xi(x;\theta)\right] \qquad I_g(\theta) = \mathrm{E}_g\left[\xi(x;\theta)\xi(x;\theta)'\right]$$

漸近分散の関係：

$$\Sigma_g(\theta^*_\gamma) = \frac{1}{1-\varepsilon}\Sigma_f(\theta^*) + O(\varepsilon\nu^\gamma)$$

結論： **外れ値は自動的に無視されている.**

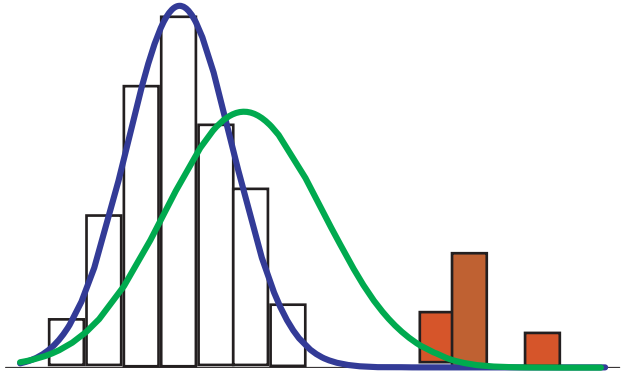
ある種のロバスト性を生じる相互エントロピーの唯一性

外れ値の割合が大きい場合にもバイアスが小さくなるロバスト推定をもたらす相互エントロピー $d(g, f)$ は，幾つかの条件の下では，**本質的に唯一つ**である：

$$d(g, f) = \phi(d_\gamma(g, f)).$$

ただし $\phi(u)$ は適当な単調増加関数である．

THANK YOU



Hironori Fujisawa

Institute of Statistical Mathematics