

# Cauchy 分布を用いた正規分布の歪対称化

2017SS091 山田祐揮

指導教員：阿部俊弘

## 1 はじめに

正規分布や Cauchy 分布は対称分布であるが、実際のデータの分布は対称であるとは限らず、非対称であるほうが自然である。本研究では、Cauchy 分布の特徴と正規分布との比較を行い、相違点を明確にした後に Cauchy 分布を用いた正規分布の歪対称化を行うことで歪 Cauchy 正規分布のモデルを与える。また、同一データに対して Newton 法を用いたパラメータ推定を行うことにより、Azzalini の歪正規分布との数値的な比較を行う。

## 2 Cauchy 分布について

Cauchy 分布は  $t$  分布の自由度  $n = 1$  のときのものである。

期待値や分散が定義されず、最頻値と中央値は常に与えられ、Cauchy 分布は 2 つの標準正規分布の比を用いて生成される。(白石 [3])

Cauchy 分布の分布関数を

$$F_C(x) = \frac{1}{2} + \frac{\arctan x}{\pi}, \quad -\infty < x < \infty$$

密度関数を

$$f_C(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$

とする。Cauchy 分布の期待値は、

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_C(x) dx \\ &= \lim_{\alpha, \beta \rightarrow \infty} \frac{1}{\pi} \int_{-\alpha}^{\beta} \frac{x}{1+x^2} dx \\ &= \lim_{\alpha, \beta \rightarrow \infty} \frac{1}{2\pi} [\log(1+x^2)]_{-\alpha}^{\beta} \\ &= \lim_{\alpha, \beta \rightarrow \infty} \frac{r}{2\pi} \log \left( \frac{1+\beta^2}{1+\alpha^2} \right) \end{aligned}$$

となるので、定義できない。このとき、Cauchy 分布の期待値が定義できないので、分散も定義することはできない。推定を行う場合、外れ値が多い場合には正規分布よりも Cauchy 分布を用いる方が有効なことがある。

## 3 Azzalini の歪対称正規分布について

Azzalini[1] の提案した歪対称正規分布は、正規分布を含む確率分布である。歪パラメータにより正規分布以上に柔軟なデータのあてはめが可能だけでなく、正規分布と類似した性質を持つため、解析的に扱いやすい確率分布である。この分布は 3 パラメータで表現され、そのうちの 1 つである歪パラメータを 0 とするとき正規分布となることから、平均と分散の 2 パラメータで表現する正規分布の拡張であるといえる。

標準正規分布に対応する歪正規分布  $SN(0, 1, \lambda)$ ,  $-\infty < \lambda < \infty$  の確率密度関数は

$$f_{SN}(x) = 2\phi(x)\Phi(\lambda x)$$

与えられる。ここで、 $\phi$  と  $\Phi$  は標準正規分布の密度関数と分布関数であり、次により与えられる：

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

$$\Phi(x) = \int_{-\infty}^x \phi(t) dt.$$

この分布の特徴として、

- 特別な場合に正規分布を含む
- 数学的に扱いやすい
- 幅広い歪度と尖度を持っている

という点あげられる。

## 4 Cauchy 分布を用いた正規分布の歪対称化

Azzalini [1] の Lemma 1 の公式を用いて、Cauchy 分布の歪対称版について考える：

$$\begin{aligned} g_{SCN}(x) &= 2F_C(\lambda x)\phi(x) \\ &= \frac{2}{\sqrt{2\pi}} \left( \frac{1}{2} + \frac{\arctan \lambda x}{\pi} \right) e^{-\frac{x^2}{2}} \end{aligned}$$

ここで、位置パラメータ  $\mu$  と尺度パラメータ  $\sigma$  は通常の手法  $x \mapsto (x - \mu)/\sigma$  により導入できるが、歪パラメータについては、 $\lambda/\sigma \mapsto \lambda$  とリパラメトライズする。関数  $f_{SCN}(x)$  を

$$\begin{aligned} f_{SCN}(x) &= \frac{2}{\sigma} F_C(\lambda(x - \mu))\phi\left(\frac{x - \mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left( 1 + \frac{2}{\pi} \arctan \lambda(x - \mu) \right) e^{-\frac{(x - \mu)^2}{2\sigma^2}} \end{aligned}$$

とする。ここで、 $-\infty < \lambda < \infty$  である。 $\frac{2}{\pi} \arctan \lambda x$  も奇関数であるので、 $f_{SCN}(x)$  は密度関数になっていることを確認できる。密度関数が  $f_{SCN}(x)$  である分布を  $SCN(\mu, \sigma, \lambda)$  と表記する。リパラメトライズをしなかった密度関数  $g_{SCN}(x)$  とリパラメトライズをした密度関数  $f_{SCN}(x)$  の  $\lambda$  の値は、 $g_{SCN}(x)$  の  $\lambda$  を  $g_{SCN}(x)$  の  $\sigma$  で割った値として定義することによって同じ分布を描くことができるのでリパラメトライズを行うことによって数学的な問題は発生しないと言える。このリパラメトライズを行うことにより、密度関数に現れるパラメータを減らすことができるので、 $g_{SCN}$  を複数回微分し、Score 方程式と Fisher 情報行列を導出する場合に比べ、簡潔な形であらわすことができる。

## 5 歪 Cauchy 正規分布の対数尤度関数

$x_1, \dots, x_n$  がパラメータ  $(\mu, \sigma, \lambda)$  の歪 Cauchy 正規分布  $SCN(\mu, \sigma, \lambda)$  に従うとき、対数尤度関数を  $l(\mu, \sigma, \lambda) = \log L(\mu, \sigma, \lambda)$  とすると、

$$l(\mu, \sigma, \lambda) = \frac{n}{2} \log \frac{1}{2\pi} - n \log \sigma + n \log 2 + \sum_{i=1}^n \log \left( \frac{1}{2} + \frac{\arctan(\lambda(x_i - \mu))}{\pi} \right) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}$$

により与えられ、この関数を最大化する  $(\mu, \sigma, \lambda)$  が推定値として定まり、 $\frac{\partial l}{\partial \mu} = 0$ ,  $\frac{\partial l}{\partial \sigma} = 0$ ,  $\frac{\partial l}{\partial \lambda} = 0$  の条件を満たす。この尤度方程式の勾配ベクトルを Score 関数という。

## 6 歪 Cauchy 正規分布の Fisher 情報行列

本節では、歪 Cauchy 正規分布の Fisher 情報行列を与える。Fisher 情報行列の対角要素は以下ようになる：

$$\frac{\partial^2 l}{\partial \mu^2} = - \sum_{i=1}^n \frac{2\pi\lambda^3(x_i - \mu)f_C(\lambda(x_i - \mu))^2}{F_C(\lambda(x_i - \mu))} - \sum_{i=1}^n \frac{\lambda^2 f_C(\lambda(x_i - \mu))^2}{F_C(\lambda(x_i - \mu))^2} - \frac{n}{\sigma^2},$$

$$\frac{\partial^2 l}{\partial \sigma^2} = - \sum_{i=1}^n \frac{3(x_i - \mu)^2}{\sigma^4} + \frac{n}{\sigma^2},$$

$$\frac{\partial^2 l}{\partial \lambda^2} = - \sum_{i=1}^n \frac{2\pi\lambda(x_i - \mu)^3 f_C(\lambda(x_i - \mu))^2}{F_C(\lambda(x_i - \mu))} - \sum_{i=1}^n \frac{(x_i - \mu)^2 f_C(\lambda(x_i - \mu))^2}{F_C(\lambda(x_i - \mu))^2}.$$

$\lambda = 0$  のときに、Fisher 情報行列が退化しないという利点があることがわかる。

## 7 Newton-Raphson 法

$x_1, \dots, x_n$  を  $SCN(\mu, \sigma, \lambda)$  に従うデータとし、 $\mathbf{x} = (x_1, \dots, x_n)^\top$  とする。このとき、 $\xi = (\mu, \sigma, \lambda)^\top$  とすると、観測 Fisher 情報行列  $J(\xi)$  は

$$J(\xi) = - \frac{\partial^2 l(\xi)}{\partial \xi \partial \xi^\top}$$

で与えられる。 $J(\xi)$  を用いて Newton-Raphson 法を適用する。 $k$ th step のパラメータを  $\xi^{(k)}$  とすると、 $(k+1)$ th step のパラメータは

$$\xi^{(k+1)} = \xi^{(k)} + J(\xi^{(k)})^{-1} S(\mathbf{x}, \xi^{(k)})$$

で更新される。

## 8 実データへの適用

Cook & Weisberg (1994) [2] の  $n = 202$  のオーストラリアのアスリートデータを使用する。このデータは全部で 6 変量あり、それぞれ、202 人のアスリートの LBM, BMI, SSF, Pcbfat, Height, Weight 等のデータがまとめられている。今回は SSF data, Pcbfat data を用いて推定を行った。

(1) は SSF data についての歪 Cauchy 正規分布、(2) は SSF data についての歪正規分布、(3) は Pcbfat data についての歪 Cauchy 正規分布、(4) は Pcbfat data についての歪正規分布の推定結果である。

	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\lambda}$	$MLL$
(1)	29.60	51.80	6.93	-945.86
(2)	29.24	51.36	46.74	-945.33
(3)	5.63	10.01	$\infty$	-611.92
(4)	5.71	9.95	81.56	-612.39

表 1 SSF data と Pcbfat data に対する歪 Cauchy 正規分布と歪正規分布のパラメータ推定の結果

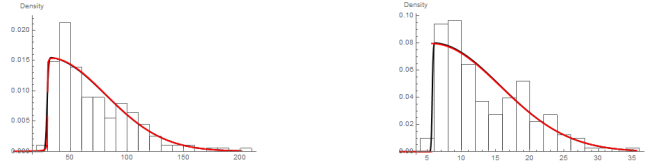


図 1 左図は SSF data であり、右図は Pcbfat data である。黒の曲線は歪正規分布の推定した密度関数、赤の曲線は歪 Cauchy 正規分布の推定した密度関数である

## 9 おわりに

本研究では、Cauchy 分布による正規分布の歪対称化について検討した。その結果、 $\lambda = 0$  のときに Fisher 情報行列が退化しないという利点があることが分かった。分布の単峰性についても検討したが、密度関数の微分の形が複雑になってしまうため、モードの位置を陽的に求めることは容易ではないことがわかった。歪 Cauchy 正規分布のパラメータ推定の例として、SSF data と Pcbfat data に対して当てはめをおこなった。

## 参考文献

- [1] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.
- [2] Cook, R. D. & Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.
- [3] 白石高章 (2012). *統計科学の基礎: データと確率の結びつきがよくわかる数理*, 日本評論社.