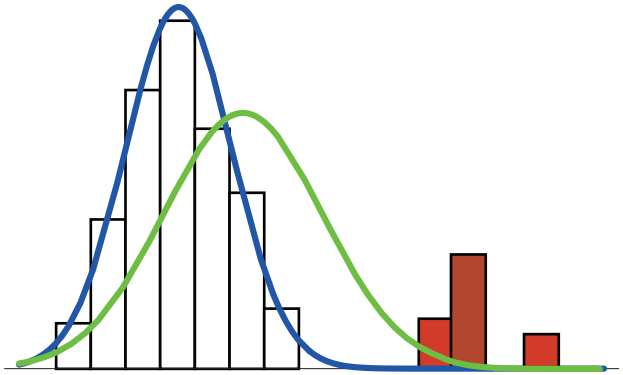
ロバストな推定を導く ダイバージェンス

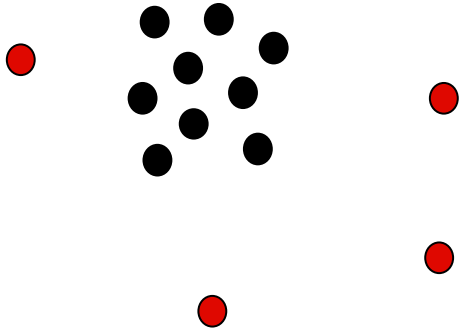
藤澤 洋徳

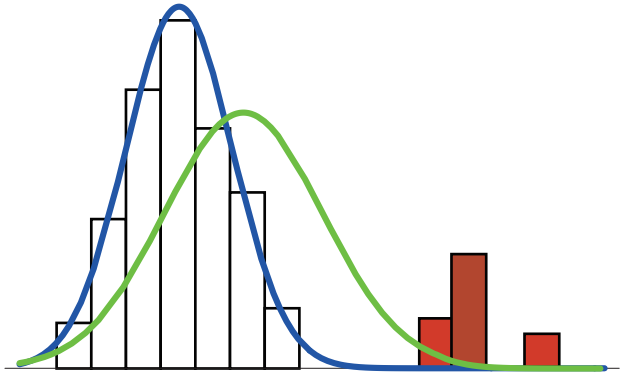
統計数理研究所／理研 AIP

fujisawa@ism.ac.jp

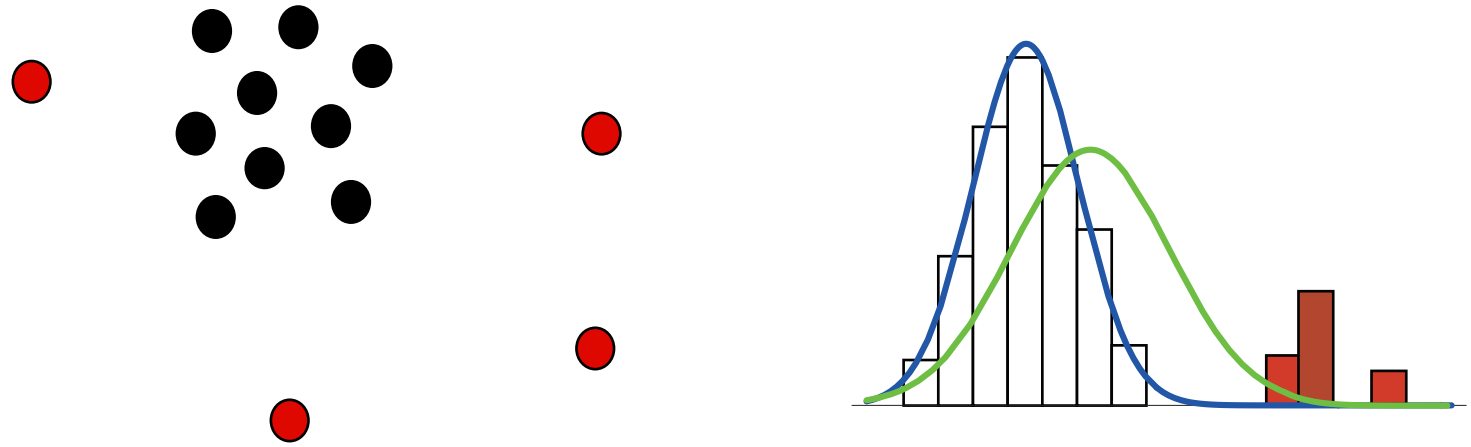


Contents





1. 外れ値とロバスト推定



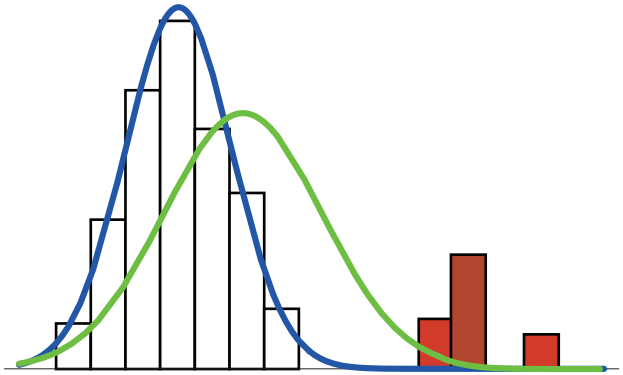
赤色が外れ値のイメージである.

目標： 青色の密度関数を同定する！

ロバスト推定

外れ値を事前にうまく取り除くというような余計な手間をかけることなく,
青色の密度関数が自動的に得られるようなパラメータ推定

1.1. 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：

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

ν_f の値が小さいほど，外れ値に起因する潜在バイアスは小さくできる，というような結論が導き出せる．

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

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

Review: ダイバージェンスと相互エントロピー

KL ダイバージェンスと相互エントロピー

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

ダイバージェンスの性質 ($\approx$ 距離)

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

経験密度関数

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

最尤推定量　$d_{KL}(g, f) = -\mathsf{E}_g[\log f] = -\int g(x) \log f(x) dx$

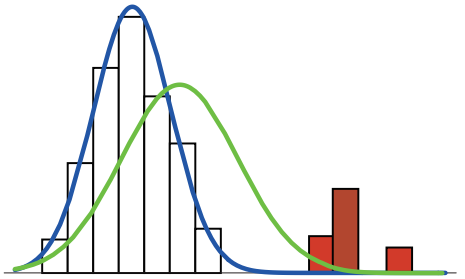
$$\begin{aligned}
\widehat{\theta} &= \arg\min_{\theta} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log f(x_i;\theta) \right\} = \arg\min_{\theta} \left\{ -\int \bar{g}(x) \log f(x) dx \right\} \\
&= \arg\min_{\theta} d_{KL}(\bar{g}, f_\theta) = \arg\min_{\theta} \{ -d_{KL}(\bar{g}, \bar{g}) + d_{KL}(\bar{g}, f_\theta) \} \\
&= \arg\min_{\theta} D_{KL}(\bar{g}, f_\theta) \\
&\rightarrow \; \theta^{\#} = \arg\min_{\theta} D_{KL}(g, f_\theta) = \arg\min_{\theta} d_{KL}(g, f_\theta)
\end{aligned}$$

1.2. (相互エントロピーに基づいた)ロバスト推定の目的

ロバスト推定としては本当はこうであって欲しい：

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

$$\begin{aligned} \widehat{\theta}_d = \arg\min_\theta d(\bar{g}, f_\theta) \quad &\to \quad \theta_d^* = \arg\min_\theta d(g, f_\theta) \\ &\approx \quad \theta^* = \arg\min_\theta d(f, f_\theta) \end{aligned}$$



ポイント： 本当に近づけたいのは，汚染された分布 g とパラメトリック分布 f_θ とではなくて，目的分布 f とパラメトリック分布 f_θ である．

目標： どのような相互エントロピー $d(g,f)$ に基づいた推定方法であれば，外れ値の割合が小さくなくても，潜在バイアス $\theta_d^* - \theta^*$ を十分に小さくできるだろうか？ （スーパーロバストネス）

2. Density Power Divergence (Basu et al. 1998)

凸関数

$$h(u) = \frac{1}{\gamma(1+\gamma)} u^{1+\gamma} \quad (\gamma > 0)$$

$$h(u) - h(v) - h'(v)(u-v) \geq 0$$

Density Power Divergence (β-divergence)

$$D_{\mathrm{pow}}(g, f) = \int [h(g(x)) - h(f(x)) - h(f(x))\{g(x) - f(x)\}] dx$$

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

相互エントロピー: $D_{\mathrm{pow}}(g, f) = d_{\mathrm{pow}}(g, f) - d_{\mathrm{pow}}(g, g)$

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

相互エントロピーの経験推定と推定

$$
\begin{aligned}
d_{\mathsf{pow}}(g, f) &= -\frac{1}{\gamma}\int g(x) f(x)^{\gamma} dx + \frac{1}{1+\gamma}\int f(x)^{1+\gamma} dx \\
d_{\mathsf{pow}}(\bar{g}, f_\theta) &= -\frac{1}{\gamma}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i;\theta)^{\gamma}\right) + \frac{1}{1+\gamma}\int f(x;\theta)^{1+\gamma} dx \\
\widehat{\theta}_{\mathsf{pow}} &= \arg\min_{\theta} d_{\mathsf{pow}}(\bar{g}, f_\theta)
\end{aligned}
$$

外れ値 x_1 の影響： $f(x_1;\theta^*)$ は小さくなる.

$$
\begin{aligned}
d_{\mathsf{pow}}(\bar{g}, f_\theta) &\approx -\frac{1}{\gamma}\frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=2}^{n} f(x_i;\theta)^{\gamma}\right) + \frac{1}{1+\gamma}\int f(x;\theta)^{1+\gamma} dx \\
&\approx -\frac{1}{\gamma}\frac{n-1}{n}\int f^*(x) f(x;\theta)^{\gamma} dx + \frac{1}{1+\gamma}\int f(x;\theta)^{1+\gamma} dx
\end{aligned}
$$

外れ値が多かったらずれそう.

3. γ-Divergence (Fujisawa and Eguchi 2008)

γ-cross entropy

$$d_\gamma(g,f) = -\frac{1}{\gamma}\log \int g(x) f(x)^\gamma dx + \frac{1}{1+\gamma}\log \int f(x)^{1+\gamma} dx \quad (\gamma > 0).$$

density power cross entropy

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

違いは log だけ

経験推定

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

外れ値 x_1 の影響： $f(x_1;\theta^*)$ は小さくなる.

$$d_\gamma(\bar{g}, f_\theta) \approx -\frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=2}^{n} f(x_i;\theta)^\gamma\right) + \frac{1}{1+\gamma}\log\int f(x;\theta)^{1+\gamma}dx$$

$$= -\frac{1}{\gamma}\log\frac{n-1}{n} - \frac{1}{\gamma}\log\left(\frac{1}{n-1}\sum_{i=2}^{n} f(x_i;\theta)^\gamma\right) + \frac{1}{1+\gamma}\log\int f(x;\theta)^{1+\gamma}dx$$

$$\approx -\frac{1}{\gamma}\log\frac{n-1}{n} + d_\gamma(f^*, f_\theta)$$

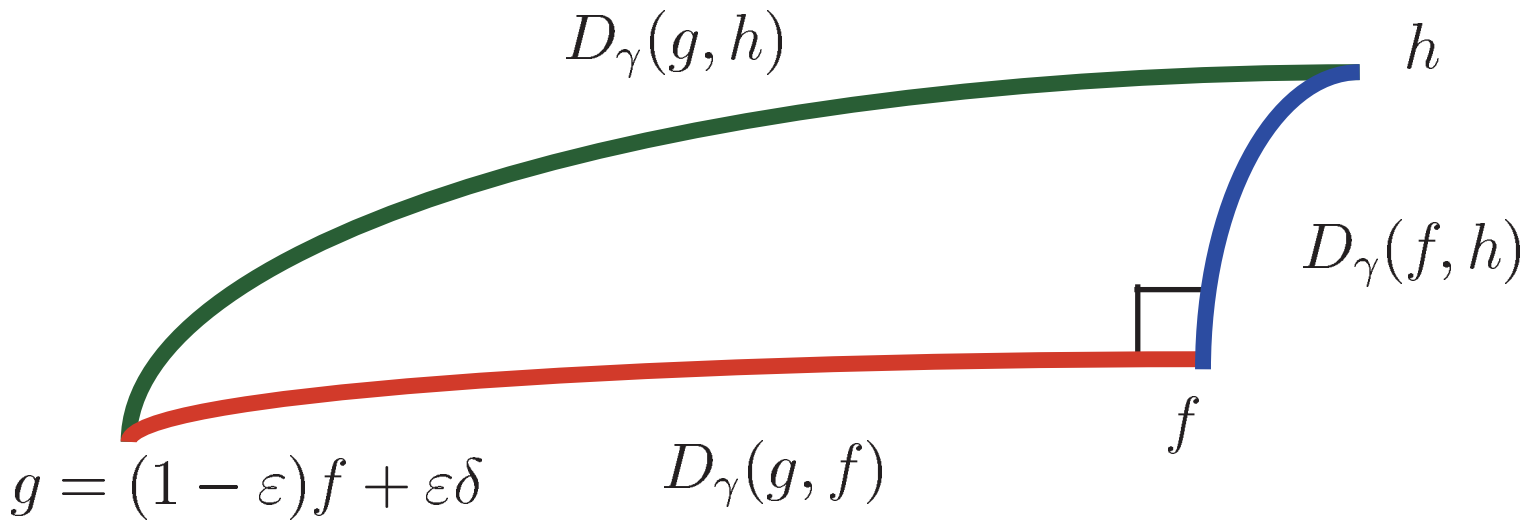
推定： 外れ値の割合が大きくても，次の論法は使えそう（使える）.

$$\hat{\theta}_\gamma = \arg\min_\theta d_\gamma(\bar{g}, f_\theta) \approx \arg\min_\theta d_\gamma(f_{\theta^*}, f_\theta) = \theta^*$$

ピタゴリアン関係

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

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



γ-Divergence の少し変わった性質

ダイバージェンス $D(g, f)$ の性質

(i) $D(g, f) \geq 0$

(ii) $D(g, f) = 0 \Leftrightarrow g = f$

2つのダイバージェンスの違い

$D_{\mathrm{pow}}(g, f)$: (i)(ii) は $\mathcal{F} = \{p(x); p(x) > 0\}$ において成り立つ.

$D_{\gamma}(g, f)$: (i) は $\mathcal{F}$ において成り立つ.
(ii) は $\mathcal{P}$ で成り立つ. $\mathcal{P} = \{p(x) > 0; \int p(x)dx = 1\}$
$\mathcal{F}$ では $g(x) = cf(x)$ と変わる. (c は定数.)

この冗長性（不変性）がとても重要. あとで見える.

最適化アルゴリズム

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：

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

ピタゴリアン関係，または，Majorization-Minimization Algorithm を利用する．

4. ロバスト推定の唯一性

まずは相互エントロピーのクラスを制限しよう：

$$d(g, f_\theta) = \psi\left(\int g\chi(f_\theta)dx, \int \rho(f_\theta)dx\right).$$

相互エントロピーのgに絡む部分がgの線形なので扱いやすい.
経験推定が可能. 推定値を次で定義できる：

$$\begin{aligned}\hat{\theta} &= \arg\min_\theta d(\bar{g}, f_\theta) \\ &= \arg\min_\theta \psi\left(\frac{1}{n}\sum_{i=1}^{n}\chi(f_\theta(x_i)), \int \rho(f_\theta)dx\right).\end{aligned}$$

外れ値を自動的に無視したい！

＃簡単のために$\delta = \delta_{x^*}$としよう．

$$
\begin{aligned}
d(g, f_\theta) &= \psi\left(\int g\chi(f_\theta)dx, \int \rho(f_\theta)dx\right) \\
&= \psi\left(\int\{(1-\varepsilon)f + \varepsilon\delta_{x^*}\}\chi(f_\theta)dx, \int \rho(f_\theta)dx\right) \\
&= \psi\left(\int(1-\varepsilon)f\chi(f_\theta)dx + \varepsilon\chi(f_\theta(x^*)), \int \rho(f_\theta)dx\right)
\end{aligned}
$$

$\chi(f_\theta(x^*))$が消えればよい．$f_\theta(x^*)$は十分に小さいだろう．
もしも$\chi(0) = 0$という感じの性質があれば，外れ値に関する部分は消える！

例えば $\chi(s) = s^\gamma$ $(\gamma > 0)$ であればよい．＃実はこれが答えになった
KL-divergence はここで駄目になる： $\chi(s) = \log(s)$.

現在位置：

$$\begin{aligned} d(g, f_\theta) &\approx \psi\left(\int (1-\varepsilon) f \chi(f_\theta) dx, \int \rho(f_\theta) dx\right) \\ &= d((1-\varepsilon)f, f_\theta). \end{aligned}$$

もしも $(1-\varepsilon)$ がなければ嬉しいのに！

$$\theta_d^* = \arg\min_\theta d(g, f_\theta) \quad \approx \quad \theta^* = \arg\min_\theta d(f, f_\theta).$$

じゃあ $(1-\varepsilon)$ があってもなくても同じとみなせる同値類が入るような相互エントロピーにすれば良い.

$$\arg\min_\theta d((1-\varepsilon)f, f_\theta) = \arg\min_\theta d(f, f_\theta) \qquad \text{for any } \varepsilon.$$

ポイント： ここまでの話で重要なことは，外れ値の割合 ε が小さくなくてはいけないと言う想定はない！

主要定理

これまでの話（＋α）をきちんと数学的に整理していくと次の命題が生まれる.

相互エントロピーは次のように表現される：

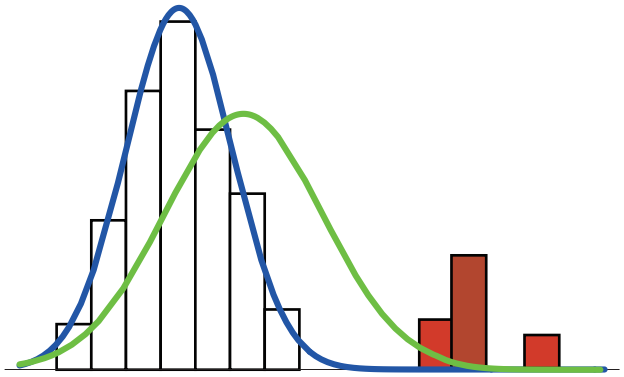
$$d(g,f) = \phi(d_\gamma(g,f)).$$

ただし，$\phi(u)$ は適当な単調増加関数であり，また，

$$d_\gamma(g,f) = -\frac{1}{\gamma}\log\int gf^\gamma dx + \frac{1}{1+\gamma}\log\int f^{1+\gamma}dx \quad (\gamma > 0),$$

である．つまり，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は，幾つかの条件の下では，本質的に唯一つである.

冗長性（不変性）の必要性



$$g = (1-\varepsilon)f + \varepsilon\delta \quad \leftarrow \quad f_\theta$$
$$g \approx (1-\varepsilon)f \quad \leftarrow \quad f_\theta$$

解釈：

外れ値が自然に無視される機構を入れ込むと，パラメトリック分布 f_θ のターゲットは $(1-\varepsilon)f$ になる．しかし本当のターゲットは $(1-\varepsilon)f$ ではなくて f なので定数倍を無意味にする機構は必要．ただしパラメトリック分布も目的分布も密度関数なので，そのノルムは等しいので，定数倍を無意味にしても，その二つは近づくことになる．

Kanamori and Fujisawa (2015): $cf(x;\theta)$ という拡張モデルを用意すれば，c が $1-\varepsilon$ を吸収するから，普通のダイバージェンスでも大丈夫かも？

5. 拡張モデル

拡張モデル

$$cf(x;\theta)$$

Theorem

Let

$$(\widehat{c},\widehat{\theta}) = \arg\min_{c,\theta} d_{\mathsf{pow}}(\overline{g}, cf_\theta).$$

Then,

$$\begin{aligned}\widehat{\theta} &= \arg\min_{\theta} d_\gamma(\overline{g}, f_\theta) \\ \widehat{c} &= \frac{\int g(x) f(x;\widehat{\theta})^\gamma dx}{\int f(x;\widehat{\theta})^{1+\gamma} dx} \quad \approx 1-\varepsilon\end{aligned}$$

Result

$$\begin{aligned}\widehat{\theta} &= \arg\min_{\theta} d_{\gamma}(\overline{g}, f_{\theta}) \\ \widehat{c} &= \frac{(1/n)\sum_{i=1}^{n} f(x_i;\widehat{\theta})^{\gamma}}{\int f(x;\widehat{\theta})^{1+\gamma}dx} \quad \approx 1-\varepsilon\end{aligned}$$

Density Power Divergence でも，拡張モデルの利用で，γ-Divergence の最小化と同じになる．
潜在バイアスを十分に小さくできる．推定アルゴリズムが簡単になる．

この結果は，Hölder Divergence (Kanamori and Fujisawa 2014) でも可能．

注意：Density Power Divergence だけだと江口先生が過去に指摘している．ただし，外れ値の割合の推定に使えるという考え方は，そこにはない．

6. Hölder Divergence (Kanamori and Fujisawa 2014)

$$d_H(p,q) \;=\; \phi\left(\frac{a_\gamma(p,q)}{a_\gamma(q,q)}\right) a_\gamma(q,q) \quad \text{for } \gamma > 0$$

$$a_\gamma(p,q) = \int p(x)q(x)^\gamma dx$$

$$\phi(1) = -1 \qquad \phi(z) \geq -z^{1+\gamma} \quad (z \geq 0)$$

Example

γ-type: $\phi(z) = -z^{1+\gamma}$ 下限

density power type: $\phi(z) = \gamma - (1+\gamma)z$

Hölder Divergence は主に次の２つの仮定から得られる：
(i) 相互エントロピーの経験推定可能性
(ii) ある種のデータのアフィン不変性

$$D(p_y, q_y) = a(\sigma) D(p_x, q_x) \qquad y = \mu + \sigma x$$

Theorem

Let

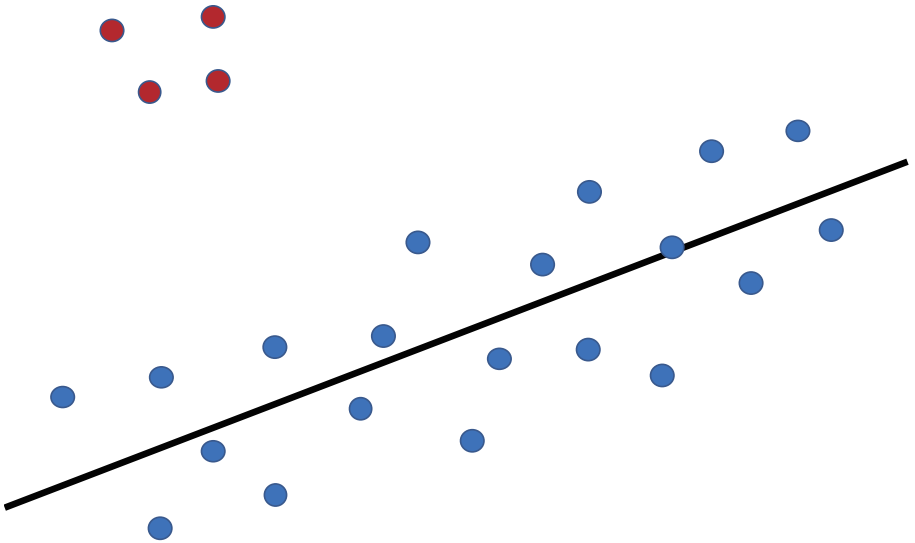
$$(\widehat{c}, \widehat{\theta}) = \arg\min_{c,\theta} d_H(\overline{g}, cf_\theta).$$

Then,

$$\begin{aligned} \widehat{\theta} &= \arg\min_{\theta} d_\gamma(\overline{g}, f_\theta) \\ \widehat{c} &= \frac{\int g(x) f(x;\widehat{\theta})^\gamma dx}{\int f(x;\widehat{\theta})^{1+\gamma} dx} \quad \approx 1 - \varepsilon \end{aligned}$$

Hölder Divergence の関数ϕに無関係に得られている.

7. Regression Problem (Kawashima and Fujisawa 2018)



$$y = \beta^\top x + e \qquad e \sim N(0, \sigma^2)$$

$$f(y|x;\theta) = \phi(y; \beta^\top x, \sigma)$$

Find a conditional pdf.

Density Power Cross Entropy For i.i.d. Case

$$d_{\mathsf{pow}}(p,q) = \int \left\{ -\frac{1}{\gamma} p(x) q(x)^{\gamma} + \frac{1}{1+\gamma} q(x)^{1+\gamma} \right\} dx$$

Density Power Cross Entropy For Regression

(between two conditional functions)

$$\begin{aligned}
& d_{\mathsf{pow}}(p_{y|x}, q_{y|x}; p_x) \\
&= \int d_{\mathsf{pow}}(p_{y|x}, q_{y|x})\, p_x(x) dx \\
&= \int \left\{ -\frac{1}{\gamma} p_{y|x}(y|x) q_{y|x}(y|x)^{\gamma} + \frac{1}{1+\gamma} q_{y|x}(y|x)^{1+\gamma} \right\} dy\, p_x(x) dx \\
&= -\frac{1}{\gamma} \int p(x,y) q_{y|x}(y|x)^{\gamma} dx dy + \frac{1}{1+\gamma} \int \left\{ \int q_{y|x}(y|x)^{1+\gamma} dy \right\} p_x(x) dx
\end{aligned}$$

γ-Cross Entropy For i.i.d. Case

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

Two Types of γ-Cross Entropy For Regression

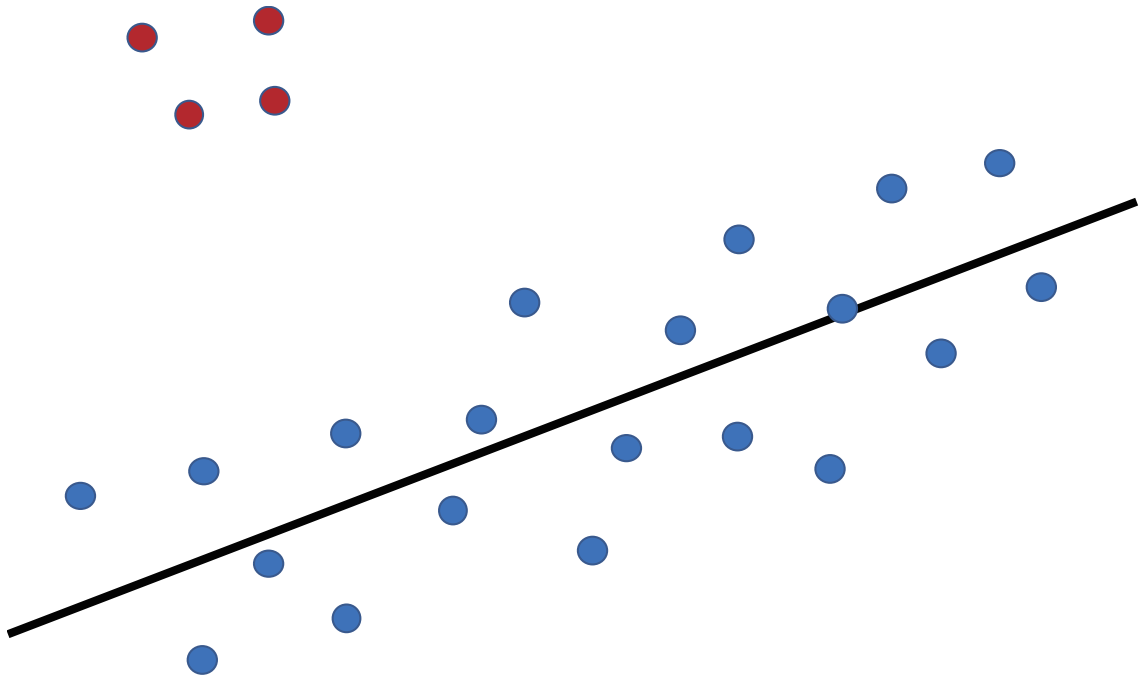
(between two conditional functions)

$$\begin{aligned} d_{1,\gamma}(p_{y|x}, q_{y|x}; p_x) &= -\frac{1}{\gamma}\log\int p_{y|x}(y|x)q_{y|x}(y|x)^\gamma dy\, p_x(x)dx \\ &\quad + \frac{1}{1+\gamma}\log\int q_{y|x}(y|x)^{1+\gamma}dy\, p_x(x)dx \\ d_{2,\gamma}(p_{y|x}, q_{y|x}; p_x) &= -\frac{1}{\gamma}\log\int \exp\{-\gamma d_\gamma(p_{y|x}, q_{y|x})\}\, p_x(x)dx \\ &= -\frac{1}{\gamma}\log\int \frac{\int p_{y|x}(y|x)q_{y|x}(y|x)^\gamma dy}{\int q_{y|x}(y|x)^{1+\gamma}dy}\, p_x(x)dx \end{aligned}$$

Remark: Both are empirically estimable using the i.i.d. dataset $\{(x_i, y_i)\}$.

$$
\begin{aligned}
d_{1,\gamma}(g_{y|x}, q_{y|x}; g_x) &= -\frac{1}{\gamma}\log\int g_{y|x}(y|x)q_{y|x}(y|x)^{\gamma}dy\, g_x(x)dx \\
&\quad + \frac{1}{1+\gamma}\log\int q_{y|x}(y|x)^{1+\gamma}dy\, g_x(x)dx \\
&= -\frac{1}{\gamma}\log\int q_{y|x}(y|x)^{\gamma}\, g(x,y)dydx \\
&\quad + \frac{1}{1+\gamma}\log\int q_{y|x}(y|x)^{1+\gamma}dy g_x(x)dx \\
d_{2,\gamma}(g_{y|x}, q_{y|x}; g_x) &= -\frac{1}{\gamma}\log\int \frac{\int g_{y|x}(y|x)q_{y|x}(y|x)^{\gamma}dy}{\int q_{y|x}(y|x)^{1+\gamma}dy} g_x(x)dx \\
&= -\frac{1}{\gamma}\log\int \frac{q_{y|x}(y|x)^{\gamma}}{\int q_{y|x}(y|x)^{1+\gamma}dy} g(x,y)dydx
\end{aligned}
$$

Two Types of Contamination For Regression



Homogeneous Contamination

$$g(y|x) = (1-\varepsilon)f(y|x;\theta^*) + \varepsilon\delta(y|x)$$

Theorem (We can show in a similar manner to in the i.i.d. case.)

$$\widehat{\theta}_{\gamma,n=\infty} - \theta^* \approx 0 \text{ for any parametric model.}$$

Pythagorian Relation

$$\begin{aligned} D_{1,\gamma}(g_{y|x}, f_{y|x;\theta}; g_x) &\approx D_{1,\gamma}(g_{y|x}, f_{y|x}; g_x) + D_{1,\gamma}(f_{y|x}, f_{y|x;\theta}; g_x) \\ D_{2,\gamma}(g_{y|x}, f_{y|x;\theta}; g_x) &\approx D_{2,\gamma}(g_{y|x}, f_{y|x}; g_x) + D_{2,\gamma}(f_{y|x}, f_{y|x;\theta}; g_x) \\ & \qquad -\frac{1}{\gamma}\log(1-\varepsilon) \end{aligned}$$

Heterogeneous Contamination

$$g(y|x) = (1 - \varepsilon(x))f(y|x;\theta^*) + \varepsilon(x)\delta(y|x)$$

Theorem

Type 1: $\widehat{\theta}_{\gamma,n=\infty} - \theta^* \approx 0$ for a location-scale family $p\left(\frac{y - h(x^\top\beta)}{\sigma}\right)\frac{1}{\sigma}$

Type 2: $\widehat{\theta}_{\gamma,n=\infty} - \theta^* \approx 0$ for any parametric model

Pythagorian Relation

$$D_{2,\gamma}(g_{y|x}, f_{y|x;\theta}; g_x) \approx D_{2,\gamma}(g_{y|x}, f_{y|x}; g_x) + D_{2,\gamma}(f_{y|x}, f_{y|x;\theta}; (1 - \varepsilon(x))g_x)$$

Remark: This is a more general result than Hung et al. (2018), who treated a logistic regression model and assumed γ is large enough.

mislabel: modeling $\longrightarrow$ outlier

Assumption

$$\int \delta(y|x) f(y|x)^{\gamma} dy \approx 0$$

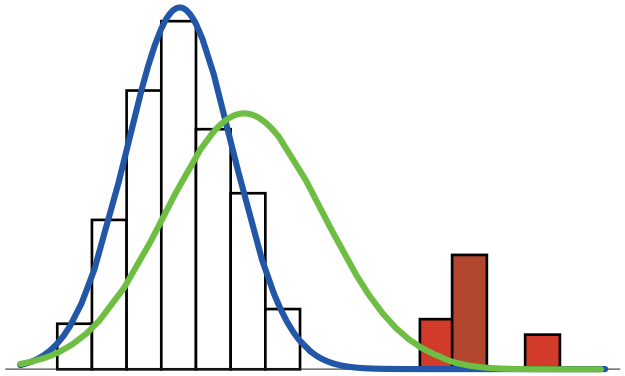
Proof

$$\begin{aligned}
& d_{2,\gamma}(g_{y|x}, f_{y|x;\theta}; g_x) \\
&= -\frac{1}{\gamma} \log \int \frac{\int g_{y|x}(y|x) f_{y|x}(y|x;\theta)^{\gamma} dy}{\int f_{y|x}(y|x;\theta)^{1+\gamma} dy} g_x(x) dx \\
&= -\frac{1}{\gamma} \log \int \frac{\int \{(1-\varepsilon(x)) f(y|x) + \varepsilon(x)\delta(y|x)\} f_{y|x}(y|x;\theta)^{\gamma} dy}{\int f_{y|x}(y|x;\theta)^{1+\gamma} dy} g_x(x) dx \\
&= -\frac{1}{\gamma} \log \Bigg[\int \frac{\int f(y|x) f_{y|x}(y|x;\theta)^{\gamma} dy}{\int f_{y|x}(y|x;\theta)^{1+\gamma} dy} (1-\varepsilon(x)) g_x(x) dx \\
&\qquad\qquad + \int \frac{\int \delta(y|x) f_{y|x}(y|x;\theta)^{\gamma} dy}{\int f_{y|x}(y|x;\theta)^{1+\gamma} dy} \varepsilon(x) g_x(x) dx \Bigg] \\
&\approx -\frac{1}{\gamma} \log \int \exp\{-\gamma d_{\gamma}(f_{y|x}, f_{y|x;\theta})\} (1-\varepsilon(x)) g_x(x) dx \\
&= d_{2,\gamma}(f_{y|x}, f_{y|x;\theta}; (1-\varepsilon(x)) g_x)
\end{aligned}$$

まとめ

2. Density Power Divergence (Bregman Divergence)
3. γ-Divergence: super-robustness. パラメータ推定アルゴリズム.
4. 相互エントロピーの唯一性
5. Extended Model: $cf(x;\theta)$
6. Hölder Divergence (Affine Invariant Divergence)
7. 回帰モデル用のγ-Divergence

THANK YOU



Hironori Fujisawa

The Institute of Statistical Mathematics