数理情報学９ (数理統計学)　5．最尤推定量

解説

データを生成する確率分布の形が $p(x;\theta)$ であるとする．パラメータ θ は未知とする．観測データからパラメータ θ を推定するためには，どのような方法を使えばよいだろうか．

平均や分散の推定では不偏推定量で推定することができた．不偏推定量が簡単に構成できるときには，そのような方法が適用可能だが，任意の統計モデルのパラメータに対して，簡単な不偏推定量が作れるとは限らない．どのような統計モデルにも (原理的には) 適用可能な普遍的な方法が，最尤推定である．

1　最尤推定量

観測値 $x_1,\ldots,x_n$ が得られているとする．これらの値は，密度関数が $p(x_1,\ldots,x_n;\theta)$ であるような確率分布に独立にしたがう確率変数の実現値とする．確率変数が離散のときには $p(x_1,\ldots,x_n;\theta)$ は確率関数とする．このとき

$$p(x_1,\ldots,x_n;\hat{\theta}_n)=\max_{\theta}\ p(x_1,\ldots,x_n;\theta)$$

もしくは

$$\log p(x_1,\ldots,x_n;\hat{\theta}_n)=\max_{\theta}\ \log p(x_1,\ldots,x_n;\theta)$$

を満たすパラメータ $\hat{\theta}$ を θ の最尤推定量という．対数関数は単調増加関数なので最大値を与えるパラメータの値は同じであることに注意．データ $x_1,\ldots,x_n$ が独立に同一の分布 $p(x;\theta)$ にしたがうときには $p(x_1,\ldots,x_n;\theta)=\prod_{i=1}^n p(x_i;\theta)$ となるので最尤推定量は

$$\sum_{i=1}^n \log p(x_i;\hat{\theta}_n)=\max_{\theta}\ \sum_{i=1}^n \log p(x_i;\theta)$$

となる．

最尤推定量

データ x が得られたとき，統計モデル $p(x;\theta)$ のもとで $p(x;\theta)$ (または $\log p(x;\theta)$) の最大値を達成するパラメータ $\widehat{\theta}_n$.

観測データ x が与えられているとき，確率密度関数 $p(x;\theta)$ を θ の関数とみなして

$p(x;\theta)$：尤度関数 (likelihood)

$\log p(x;\theta)$：対数尤度関数 (log-likelihood)

という．実際に計算するときには対数尤度を用いるほうが便利な場合が多い．

最尤推定量は以下のような直観にもとづいている．

最尤推定の考え方

データ x が観測されたとき「x が出現する確率が大きい」から観測された，と考える

⇓

x の出現確率が大きい分布 $\max_{\theta} p(x;\theta)$ を推定量とする (図 1)

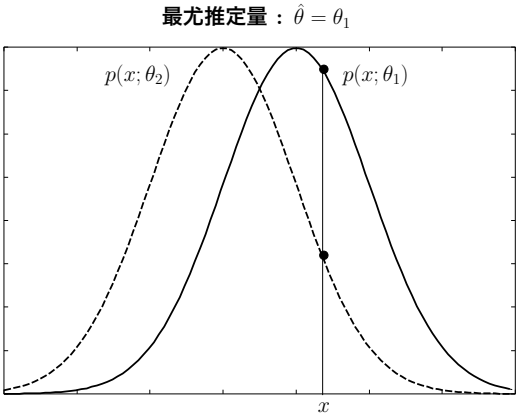

図 1: 最尤推定と尤度の最大化.

2 最尤推定量の計算法

対数尤度 $\log p(x;\theta)$ が，パラメータ θ について微分可能か不可能かで場合分けして解説する.

対数尤度がパラメータ θ について微分可能なとき： 確率密度関数 $p(x;\theta)$ が適当な数学的条件を満たすとする．最大値を求めるために，関数の極値条件を考える．たとえば θ が 1 次元パラメータのとき，最尤推定量 $\hat{\theta}$ は

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log p(x_i;\hat{\theta}) = 0$$

の解となる．上の式を 尤度方程式 とよぶ．より一般に $\theta = (\theta_1,\ldots,\theta_d) \in \boldsymbol{R}^d$ のときには

$$\text{尤度方程式：}\quad \frac{\partial}{\partial\theta_k}\sum_{i=1}^{n}\log p(x_i;\hat{\theta}) = 0,\quad k = 1,\ldots,d$$

を解くことで最尤推定量が求まる．正規分布の平均と分散の最尤推定量などは，このような計算から求めることができる.

練習問題 **1.** データ $X_1,\ldots,X_n$ が独立に正規分布 $N(\mu,1)$ から得られているとする．このとき μ の最尤推定量 $\hat{\mu}$ を計算せよ.

練習問題 **2.** データ $X_1,\ldots,X_n$ が独立に指数分布 $Ex(1/\lambda)$ から得られているとする．確率密度関数は

$$p(x;\lambda) = \frac{1}{\lambda}e^{-x/\lambda},\quad (0 \le x)$$

で与えられる．このときパラメータ λ の最尤推定量を計算せよ.

対数尤度がパラメータ θ について微分不可能なとき： 尤度の極値条件から最尤推定量を求めることができないので，尤度を最大にするパラメータを直接計算する必要がある．例を示す.

例 **1** (一様分布のパラメータの最尤推定量). 一様分布 $U[0,\theta]$ から独立に得られた観測値を $x_1,x_2,\ldots,x_n$ とする．これらの観測値からパラメータ θ の最尤推定量 $\hat{\theta}$ を求める．パラメータ θ の範囲は $\theta > 0$ とする．一様分布の密度関数は

$$f(x;\theta) = \frac{1}{\theta}\,\mathrm{I}\,(0 \le x \le \theta)$$

である．ここで $\mathrm{I}(A)$ は定義関数であり A が真なら 1，偽なら 0 をとる．観測値 $x_1, x_2, \ldots, x_n$ のもとでの尤度関数 $L(\theta)$ は

$$\begin{aligned} L(\theta) &= \frac{1}{\theta^n} \prod_{i=1}^{n} \mathrm{I}\left(0 \le x_i \le \theta\right) \\ &= \begin{cases} \dfrac{1}{\theta^n} & 0 \le x_1, \ldots, x_n \le \theta \\ 0 & \text{その他} \end{cases} \\ &= \begin{cases} \dfrac{1}{\theta^n} & 0 \le \min_i x_i \ \text{ かつ } \ \max_i x_i \le \theta \\ 0 & \text{その他} \end{cases} \end{aligned}$$

となる．尤度関数のグラフを図 2 に示す．尤度を最大にするパラメータ (最尤推定量) $\hat{\theta}$ は

$$\hat{\theta} = \max_i x_i$$

となることが分かる．図 2 の尤度関数は確かに微分不可能であり，極値条件からパラメータを求めることはできないことが分かる． □

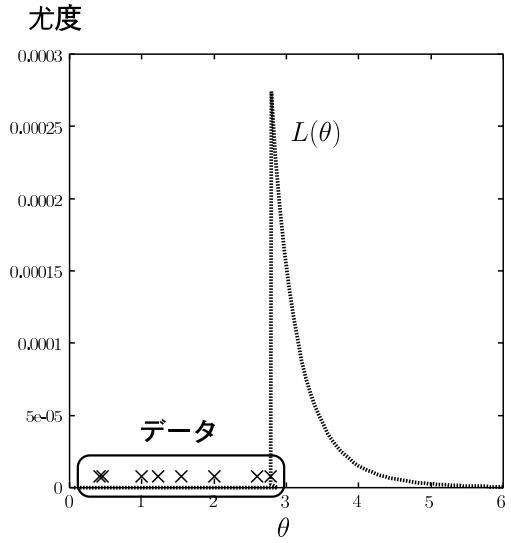

図 2: 一様分布のパラメータ θ に関する尤度関数．$U[0,3]$ から 8 個のデータが得られたときの尤度関数を示している．このデータでは，最尤推定量は $\hat{\theta} \doteqdot 2.78$ となる．

実際の多くの例の場合，最尤推定を解析的に導出することはできず (つまり簡単な式で表すことができず)，ニュートン法などの数値最適化法が必要になる．

3 最尤推定量の一致性・漸近分布

確率密度 $p(x;\theta)$ をもつ分布から，n 個のデータが独立に得られているとする．パラメータ θ の推定量 $\widehat{\theta}_n$ が

$$\widehat{\theta}_n \xrightarrow{p} \theta \quad (n \to \infty)$$

を満たすとき $\widehat{\theta}_n$ を θ の 一致推定量 という (または「推定量 $\widehat{\theta}_n$ には一致性がある」という)．つまりデータ数が多いときには，推定量 $\widehat{\theta}_n$ は真のパラメータに近い値をとる．一致性があるような推定量を構成することは非常に重要である．適当な条件のもとで 最尤推定量は一致推定量である ことが証明される．

また十分 n が大きいとき，最尤推定量 $\widehat{\theta}_n$ の分布は真のパラメータ θ を期待値とする正規分布に法則収束することが示される．これを厳密に表現すると以下のようになる：

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N\left(0, I(\theta)^{-1}\right), \quad (n \to \infty)$$

ここで $I(\theta)$ は統計モデル $\{p(x;\theta) \mid \theta \in \Theta \subset \mathbb{R}\}$ のフィッシャー情報量

$$I(\theta) = \int \left\{ \frac{\partial}{\partial \theta} \log p(x;\theta) \right\}^2 p(x;\theta) dx$$

である．

不偏推定量の場合には，推定量の分散は I^{-1} 以下にはならないことを以前紹介した (クラメール・ラオの下限)．最尤推定量は一般にパラメータの不偏推定量ではないが，$n \to \infty$ という漸近的な状況で，推定量の漸近分散がフィッシャー情報量の逆数に一致する．

4 指数型分布族と最尤推定量

4.1 指数型分布族

指数型分布族とは正規分布や指数分布，ポアソン分布などを含む確率分布のクラスのことであり，以下のように確率分布の集合として定義される．一般に d 次元指数型分布族が定義できるが，ここでは簡単のため 1 次元と 2 次元の場合を示す．

1 次元の指数型分布族

$$M_1 = \left\{ p(x;\theta_1) = h(x) \exp\left(\theta_1 t_1(x) - \psi(\theta_1)\right) \mid \theta_1 \in S_1 \subset \boldsymbol{R} \right\}$$

2 次元の指数型分布族

$$M_2 = \left\{ p(x;\theta) = h(x) \exp\left(\theta_1 t_1(x) + \theta_2 t_2(x) - \psi(\theta_1, \theta_2)\right) \;\middle|\; \theta = (\theta_1, \theta_2) \in S_2 \subset \boldsymbol{R}^2 \right\}$$

S_1, S_2 はそれぞれ $\boldsymbol{R}^1, \boldsymbol{R}^2$ の開集合とする．d 次元の指数型分布族では，d 次元パラメータ $(\theta_1, \ldots, \theta_d)$ によって確率密度関数が指定される．確率変数 X の次元は何次元でもよい．θ_1 や θ_2 を 自然パラメータ と呼ぶ．

具体例を示す．関数 $t_k(x)$ やパラメータ θ_k を適当に選ぶことで，様々な統計モデルが表現できることを説明する．

例 **2** (指数分布)．指数分布の確率密度関数は $p(x;\lambda) = \lambda e^{-\lambda x}, (\lambda > 0)$ と書ける．したがって

$$p(x;\lambda) = \exp\{-\lambda x + \log \lambda\}$$

となる．ここで

$$\begin{aligned} t_1(x) &= x, \quad \theta_1 = -\lambda, \\ h(x) &= 1, \quad \psi(\theta_1) = -\log(-\theta_1) \end{aligned}$$

とする．このとき指数分布の集合は

$$M_1 = \{ p(x;\theta) = h(x) \exp\{\theta_1 t_1(x) - (-\log(-\theta_1))\} \mid \theta_1 < 0 \}$$

と書けるので 1 次元指数型分布族であることが分かる．

例 **3** (正規分布). 平均 μ と分散 σ^2 をパラメータとする正規分布の集合は指数型分布族として表現できる. これは正規分布を次のように変形すると分かる.

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

ここで

$$t_1(x)=x,\quad t_2(x)=x^2,$$
$$\theta_1=\frac{\mu}{\sigma^2},\quad \theta_2=-\frac{1}{2\sigma^2},\quad h(x)=1$$

とおく. すると

$$\mu=-\frac{\theta_1}{2\theta_2},\ \sigma^2=-\frac{1}{2\theta_2}$$

となる. このようにパラメータを置き換えると正規分布の密度関数は

$$M_2=\left\{p(x;\theta_1,\theta_2)=h(x)\exp\left\{\theta_1t_1(x)+\theta_2t_2(x)-\left(-\frac{\theta_1^2}{2\theta_2}-\frac{1}{2}\log(-2\theta_2)+\frac{1}{2}\log 2\pi\right)\right\}\ \middle|\ \theta_1\in\boldsymbol{R},\theta_2\in(-\infty,0)\right\}$$

となる. したがって平均 μ, 分散 σ^2 の正規分布全体の集合は 2 次元指数型分布族であることが分かる. なお, パラメータ $\theta=(\theta_1,\theta_2)$ の範囲は $\mu\in\boldsymbol{R},\ \sigma^2>0$ の条件から導出される. □

練習問題 **3.** 離散確率変数の場合には, (密度関数ではなく) 確率関数の集合が M_1 や M_2 のように表現できるときに, 指数型分布族になっていると考える. 二項分布の集合

$$\left\{P(X=k;q)=\binom{n}{k}q^k(1-q)^{n-k}\ \middle|\ 0<q<1\right\}$$

は指数型分布族として表現できること示せ.

4.2 指数型分布族に対する最尤推定量

簡単のため 1 次元指数型分布族の最尤推定量について考える (例えば指数分布やポアソン分布).

1 次元指数型分布族を

$$M_1=\left\{p(x;\theta)=h(x)\exp\left(\theta t(x)-\psi(\theta)\right)\ \middle|\ \theta\in S\subset\boldsymbol{R}\right\}\tag{1}$$

とする. この統計モデルに含まれる分布 $p(x;\theta^*)$ から観測値 $x_1,\dots,x_n$ が独立に得られたとする. 観測値からパラメータ θ^* を最尤推定量で推定する.

対数尤度は

$$\begin{aligned}\ell(\theta)&=\sum_{i=1}^n\log h(x_i)+\sum_{i=1}^n\left\{\theta t(x_i)-\psi(\theta)\right\}\\&=\sum_{i=1}^n\log h(x_i)+n\,\theta\,\frac{1}{n}\sum_{i=1}^n t(x_i)-n\psi(\theta)\end{aligned}$$

となる. ここで

$$\bar{T}_n=\frac{1}{n}\sum_{i=1}^n t(x_i)$$

とおくと, 尤度方程式は

$$\frac{\partial\ell}{\partial\theta}(\widehat{\theta})=0\iff\bar{T}_n=\frac{\partial\psi}{\partial\theta}(\widehat{\theta})$$

となる. 上の方程式を解くことで最尤推定量 $\widehat{\theta}$ を得る.

演習問題

1. データ $X_1, \ldots, X_n$ が独立に正規分布 $N(\mu, \sigma^2)$ から得られているとする．

   (a) μ の値は $\mu = 0$ で既知のとき σ^2 を推定する．分散の範囲が $1 \leq \sigma^2$ であることが分かっているとき，σ^2 の最尤推定量 $\widehat{\sigma^2}$ を求めよ．

   (b) σ^2 の値は既知として μ を推定する．期待値 μ の範囲が $0 \leq \mu \leq 1$ であることが分かっているとき，μ の最尤推定量 $\widehat{\mu}$ を求めよ．

   ヒント：$1 \leq \sigma^2$ または $0 \leq \mu \leq 1$ の範囲で尤度を最大化する．

2. データ $X_1, \ldots, X_n$ は独立に正規分布 $N(\theta, \theta^2)$ (ただし $\theta < 0$) の正規分布にしたがっているとする．

   (a) θ の最尤推定量 $\widehat{\theta}_n$ を求めよ．

   (b) $\widehat{\theta}_n$ が θ の一致推定量になっているかどうか調べよ．

3. ある農園一帯のキャベツの葉に，モンシロチョウの卵がいくつ産みつけられてるかというデータ $x_1, \ldots, x_n$ が得られているとする．ここで x_i は i 番目の葉に産みつけられている卵の個数を表し，各葉に産みつける卵の個数には独立性を仮定する．このデータからモンシロチョウが葉に卵を産みつけるときの個数の分布を推定する．ここで気をつける点は，ある葉に卵が産みつけれられていない $(x_i = 0)$ とき，そもそもモンシロチョウがその葉に留まらなかったのか，留まったが産卵はしなかったか区別できないことである．

   それぞれの葉に対して，モンシロチョウが葉に留まる確率を q, 留まらない確率を $1 - q$ とする．また，モンシロチョウが葉に留まった条件のもとで x 個の卵を産む確率は，ポアソン分布にしたがうとする．ポアソン分布の確率は

   $$P(X = x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots, \quad \lambda > 0$$

   と書ける．パラメータ λ はキャベツの葉に依らずに一定の値とする．またモンシロチョウはすでに卵が産みつけられている葉には留まらないとする．以下の問に答えよ．

   (a) $x_i \geq 1$ であるようなデータの個数を a, $x_i = 0$ であるようなデータの個数を $n - a$, また $x_i \geq 1$ となっているデータに関する標本平均を

   $$z = \frac{1}{a} \sum_{i\,:x_i \geq 1} x_i$$

   とする．このとき対数尤度は

   $$\ell(\lambda, q) = (n - a) \log(1 - q + qe^{-\lambda}) + a \log q - a\lambda + az \log \lambda + \text{const.}$$

   となることを示せ．ここで const はパラメータ λ, q に依存しない項である．

   (b) λ と q の最尤推定量 $\widehat{\lambda}, \widehat{q}$ は

   $$\widehat{q} \;=\; \min\Big\{\frac{a}{n(1 - e^{-\widehat{\lambda}})},\, 1\Big\}$$

   を満たすことを示せ．

4. 平均が $(0, 0)$ で分散共分散行列が

   $$\Sigma = \begin{pmatrix} \sigma^2 & \rho \\ \rho & \sigma^2 \end{pmatrix}, \quad |\rho| < \sigma^2$$

であるような 2 次元正規分布の確率密度関数を $p(x, y; \sigma^2, \rho)$ とする．また 2 次元正規分布の集合を次のような 2 次元指数型分布族

$$M_2 = \left\{ \exp\left\{\theta_1(x^2+y^2) + \theta_2 xy - \psi(\theta_1, \theta_2)\right\} \;\middle|\; (\theta_1, \theta_2) \in S_2 \subset \boldsymbol{R}^2 \right\}$$

として表現する．以下の問に答えよ．

(a) データ $(X_1, Y_1), \ldots, (X_n, Y_n)$ が独立に 2 次元正規分布 $p(x, y; \sigma^2, \rho)$ から得られているとする．また

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 + Y_i^2\right) \quad T_2 = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

とする．自然パラメータ (θ_1, θ_2) の最尤推定量 $(\widehat{\theta_1}, \widehat{\theta_2})$ を T_1, T_2 を用いて表せ．ただし $2|T_2| < T_1$ が成立しているとする．

(b) データ $(X_1, Y_1), \ldots, (X_n, Y_n)$ が独立に確率密度関数 $p(x, y; \sigma^2, \rho_0)$ をもつ確率分布から得られているとする．このとき ρ_0 の値は既知として σ^2 を推定する．パラメータ σ^2 に関する最尤推定量の尤度方程式を ρ_0, T_1, T_2 を用いて表せ．

(c) (b) と同じ設定でデータが得られているとする．大数の法則から $T_2 \xrightarrow{p} \rho_0$ が成立するので，データ数が十分多いとき，データから計算される T_2 に対して $T_2 = \rho_0 + \varepsilon$ を仮定する．ここで ε は微小量とする．このとき σ^2 の最尤推定量 $\widehat{\sigma^2}$ に関して

$$\widehat{\sigma^2} = A(T_1) + B(T_1)\varepsilon + O(\varepsilon^2)$$

が成り立つように $A(T_1)$，$B(T_1)$ を定めよ．

ヒント：$T_2 = \rho_0 + \varepsilon$, $\widehat{\sigma^2} = A + B\varepsilon$ を尤度方程式に代入したときに $O(1)$ と $O(\varepsilon)$ の項が消えるように A, B を定める．

5. データ $X_1, \ldots, X_n$ が独立に密度関数 $f(x; \theta) = (1/\theta)e^{-x/\theta}$ をもつ分布から得られているとする．ここで $f(x; \theta)$ は $0 < x \leq \infty$ 上の分布である．パラメータ θ のもとで $X \leq 2$ となる確率を

$$\eta = P_\theta(X \leq 2) = \int_0^2 f(x; \theta)dx$$

とする．η の最尤推定量を求めよ．(ヒント：パラメータを θ から η に変換して最尤推定量を計算する)

6. データ $X_1, \ldots, X_n$ が独立に正規分布 $N(\mu, \sigma^2)$ から得られているとする．ここで $n \geq 2$ とする．パラメータ (μ, σ^2) の最尤推定量 $(\widehat{\mu}, \widehat{\sigma^2})$ を計算せよ．

7. 1 次元指数型分布族を

$$M = \left\{ p(x; \theta) = h(x)\exp\{\theta\, t(x) - \psi(\theta)\} \;\middle|\; \theta \in S \subset \boldsymbol{R} \right\}$$

と定める．S は $\boldsymbol{R}$ の適当な開集合とする．以下の問では x に関する積分と θ に関する微分は交換可能と仮定して計算してよい．

(a) 確率分布 $p(x; \theta)$ のもとでの $t(X)$ の期待値 $E_\theta[t(X)]$ は関数 $\psi(\theta)$ の微分に等しいこと，すなわち以下の等式が成立することを示せ．

$$\int t(x)p(x; \theta)dx = \frac{\partial \psi}{\partial \theta}(\theta)$$

(b) 統計モデル M のフィッシャー情報量 $I(\theta)$ を

$$I(\theta) = \int p(x; \theta)\left(\frac{\partial \log p(x; \theta)}{\partial \theta}\right)^2 dx$$

とする．また確率分布 $p(x;\theta)$ のもとでの $t(X)$ の分散を $V_\theta[t(X)]$ とする．次の 2 つの等式

$$I(\theta) = \frac{\partial^2 \psi}{\partial \theta^2}(\theta) = V_\theta[t(X)]$$

が成り立つことを示せ．

(c) データが確率分布 $p(x;\theta^*)$ から独立に得られるとする．n 個のデータから求めた最尤推定量を $\hat{\theta}_n$ とする．$\psi(\theta)$ は S 上で 2 階微分可能として，さらに $I(\theta^*) \neq 0$ を仮定する．このとき

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N\left(0, \frac{1}{I(\theta^*)}\right)$$

となることを示せ．(ヒント：デルタ法で計算)

8. データ $X_1, \ldots, X_n$ は独立であり，それぞれ $X_i \sim N(\mu, \sigma_i^2)$ とする．ただし σ_i^2 はすべて既知とする．平均パラメータ μ の最尤推定量 $\hat{\mu}$ を求めよ．また $\hat{\mu}$ は不偏推定量であるか調べよ．

9. データ $X_1, \ldots, X_n$ が独立に指数分布 $\mathrm{Ex}(\lambda)$ にしたがっているとする．指数分布 $\mathrm{Ex}(\lambda)$ の密度関数は

$$f(x;\lambda) = \begin{cases} \lambda \exp(-\lambda x) & 0 \leq x \\ 0 & \text{その他} \end{cases}$$

で与えられる．このときパラメータ λ の最尤推定量 $\hat{\lambda}$ を求めよ．

10. 一様分布 $U(0,\theta)$ は 1 次元指数型分布族として表現できるかどうか調べよ．