

仮説に制約条件がある場合の

Bivariate Sign Test

九大理学部 野間口謙太郎

Hodges (1955 AMS) は 1 次元 data に対する sign 検定を union-intersection principle を用いて、2 次元 data に対して拡張した。これがいわゆる bivariate sign test である。彼の考え方は parameter space に cone-制約条件が付いた場合でも適用可能である。

R^2 上の分布関数 F を、原点を境界に持つ任意の half space 上の確率が常に $1/2$ であるものとする。 C を R^2 での真の closed convex cone とする。一般性を失わずに一つの edge を x -軸に取ってよい。さらにもう一つの edge を vector L 方向とする。ここで vector の argument 関数 $\arg(\cdot)$ を通常の argument 関数を π で割ったものとする。例えば $\arg((0,1)')$ は $1/2$ となる。これを用いて C は $\{x \in R^2 \mid 0 \leq \arg(x) \leq \arg(L)\}$ とかける。このとき sample $X_1, X_2, \dots, X_n \sim F(x - \theta)$ に基づいて、

仮説検定問題 $H_0: \theta = 0$ vs. $H_1: \theta \in C - \{0\}$ に対する、Hodges 流の検定統計量 T_n を提案したい。ここでの取り扱い方は Joffe & Klotz (1962 AMS) に従っていて、この統計量の exact 分布、漸近分布を求め、Bahadur efficiency を計算する。

まず Hodges が、どのように考えたか、みてみたい。彼は

(1) 任意の $\mu \in R^2$ に対して $H_0(\mu) \equiv H_0$, $H_1(\mu)$

$\equiv \{\mu\}$ と置き、仮説を次のように分解した。

$$H_0 = \bigcap_{\mu \in R^2} H_0(\mu), \quad H_1 = \bigcup_{\mu \in R^2} H_1(\mu)$$

(2) 次に、部分仮説検定問題 $H_0(\mu)$ vs. $H_1(\mu)$

に対する検定統計量として $\mu'X_1, \dots, \mu'X_n$ を

用いた sign 検定を採用する。つまり

$$T_\mu \equiv (1/n) \sum_{i=1}^n I[\mu'X_i > 0]$$

を用いる。

(3) union-intersection principle の考え方により本来

の検定統計量として次の

$$T \equiv \sup_{\mu \neq 0} T_\mu$$

を採用する。 T が大きいとき H_0 が棄却されることになる。

これが、Hodges のとった approach である。これになら
 って制約条件付の場合も同様にしたいのであるが、よくなが
 めると (2) において必ずしも X_i と対立仮説の方向 μ との内
 積をとる必要はないということが解る。内積をとる相手と
 してはもっと良い方向 ν がありうる。次の例を考えてみる。
 各 X_i が、elliptical 分布に従っているとする。つまり X_i
 $= V_i \Sigma^{1/2} U_i$ という形をしているとしよう。ただし、 Σ :
 positive definite matrix、 U_i : 単位円周上の一様分布に
 従う r.v.、 V_i : 正の r.v. とする。このとき、 $H_0(\mu)$
 vs. $H_1(\mu)$ に対する sign 検定の中で most powerful な
 方向は、

$$\begin{aligned} P(\nu' X_i \geq 0 \mid \theta = \mu) \\ &= P(\nu' (V_i \Sigma^{1/2} U_i + \mu) \geq 0) \\ &= E(P(\nu' \Sigma^{1/2} U_i \geq V_i^{-1} \nu' \mu \mid V_i)) \end{aligned}$$

を最大にするものであるから、 $\nu = \Sigma^{-1} \mu$ とすれば良いこと
 がわかる。勿論 Σ は未知であるから直接にこれとの内積は
 とれないのであるが、 $\Sigma = I$ の場合は μ との内積をとれば良
 いのだから、data X_i をなるべく spherical 分布に従うよう
 に変換した後で対立仮説方向との内積を取ることにすれば良
 い。この考え方により次ぎの検定統計量を提案する。

(1) 各 data の原点からの距離を一定にする。

$$Y_i \equiv X_i / |X_i|$$

(2) 第3、4象限の data は原点に関して対称な点に移す。

$$Z_i \equiv \begin{cases} Y_i & \text{if } 0 \leq \arg(Y_i) < \pi \\ -Y_i & \text{otherwise} \end{cases}$$

(3) Z_i をその argument の順序を守りながら第1象限の半円周上に等間隔に並べる。 Z_i に対応するものを U_i とおく。つまり Z_i の argument が j 番目ならば $\arg(U_i) = (j/n) - (1/2n)$ であるようにとる。このとき $\rho \equiv P(X_i \in C \cup (-C))$ 、この一致推定量を

$$\hat{\rho} \equiv (1/n) \sum_{i=1}^n I[Z_i \in C]$$

とおき, vector L' , cone C' を次のように定義する。

$$\arg(L') = \hat{\rho},$$

$$C' \equiv \{x \mid 0 \leq \arg(x) \leq \arg(L')\}.$$

L' と C' は、 Z_i から U_i への変換に伴う L と C の変換に対応する。任意の $v \in R^2$ に対しても同様に、 $v' \in R^2$ を定義する。

(4) (2)において対称な点に移したものは再度対称な点に返す。

$$V_i \equiv \begin{cases} U_i & \text{if } Z_i = Y_i \\ -U_i & \text{otherwise} \end{cases}$$

$$(5) \quad R_t \equiv \{X \in R^2 \mid -1/2 < (\arg(X) - t) \bmod 2 \leq 1/2\}$$

と置き、

$$T_n(t) \equiv \sum_{i=1}^n I[V_i \in R_t]$$

を定義して、検定統計量として

$$T_n \equiv \sup_{0 \leq t \leq \rho} T_n(t)$$

を採用する。 α -point を $k_\alpha(\rho)$ と置くと、検定は

$$T_n \geq n/2 + \sqrt{n}k_\alpha(\hat{\rho})$$

のとき、 H_0 を棄却する事になる。

なお、制約条件が無い場合には optimal な方向などというようなものを考える必要はない。なぜならば、全ての方向との内積をとってその sup をみるのだから、結局 optimal な方向との内積も考慮されていることになるからである。

*** EXACT 分布 の 計算 ***

まず、 $C = \{X \in R^2 \mid 0 \leq \arg(X) \leq 1\}$ の場合を考えてみる。この場合の統計量を特別に T_n^* と書くことにする。 $T_n(i/n) - T_n(0)$, $i=0,1,2,\dots,n$ が random walk であることは容易に確かめることができるので、 S_i , i

$= 0, 1, 2, \dots, n$ を random walk とするとき

$$\begin{aligned} P \left(2 \left(T_n^* - n/2 \right) \leq t \right) \\ = P \left(2 \sup_{0 \leq l \leq n} S_l - S_n \leq t \right) \end{aligned}$$

と書ける。ところで、鏡像原理によると

$$\begin{aligned} P \left(S_1 < a, \dots, S_n < a, S_n = b \right) \\ = P \left(S_n = b \right) - P \left(S_n = 2a - b \right) \end{aligned}$$

であるから、 n, t が偶数のとき、

$$\begin{aligned} P \left(2 \left(T_n^* - n/2 \right) \leq t \right) \\ = \sum_{l=-t}^t P \left(S_j < (t+i)/2, j=1, \dots, n, S_n = i \right) \\ = P \left(-t \leq S_n \leq t \right) - (t+1) P \left(S_n = t+2 \right) \end{aligned}$$

n, t が奇数の場合も同様にして、

$$= P \left(-t \leq S_n \leq t \right) - t P \left(S_n = t+2 \right)$$

を得る。

一般の $C = \{ X \in R^2 \mid 0 \leq \arg(X) \leq \arg(L) \}$ の場合は、 $C \cup (-C)$ に落ちた data の数 N_0 の条件付のもとで求められる。明らかに $N_0 \sim B(n, \rho)$ であるが、 $N_0 = n_0$ が与えられたとき、 $\bigcap_{0 \leq t \leq \arg(L')} R_t$ に含まれる

data U_i は常に T_n に +1 として貢献するので、その数を Y_{n-n_0} と置く、 $R^2 - C \cup (-C)$ に含まれる n_0 個の U_i に関

しては $C = \{x \in R^2 \mid 0 \leq \arg(x) \leq 1\}$ の場合に考察したものと同様であるから $T_{n_0}^*$ と書く。すると $T_n = Y_{n-n_0} + T_{n_0}^*$ と表現でき、 $N_0 = n_0$ の条件付で、 $Y_{n-n_0} \perp T_{n_0}^*$ 、 Y_{n-n_0} である。これより T_n の分布を求めるには、 N_0 が与えられたもとでの Y_{n-n_0} と $T_{n_0}^*$ との convolution をとり、しかるのちに N_0 の分布で mixture をとらなければならない。この計算は不可能とは言わないが、大変にうるさい。そこで、次ぎに漸近分布の計算をやってみる。

*** 漸 近 分 布 の 計 算 ***

まず T_n^* の漸近分布を求めてみる。exact 分布より n が偶数のとき、

$$\begin{aligned} P(2(T_n^* - n/2) \leq \sqrt{n}t) \\ = P(-[\sqrt{n}t] \leq S_n \leq [\sqrt{n}t]) \\ = ([\sqrt{n}t] + 1)P(S_n = [\sqrt{n}t] + 2) \end{aligned}$$

である。ただし $[\sqrt{n}t]$ は $\sqrt{n}t$ より小さい最大の偶数である。

また n が奇数の場合も同様であるから、これより n について極限をとって、次を得る。

$$\lim_{n \rightarrow \infty} P(2(T_n^* - n/2) \leq \sqrt{n}t)$$

$$= \Phi(t) - \Phi(-t) - t \phi(t)$$

ただし Φ 、 ϕ はそれぞれ標準正規分布 $N(0,1)$ の分布関数と密度関数である。この極限分布は、 Z を自由度3のカイ二乗分布に従うとするときの Z^2 の従う分布に等しいので、これを $\sqrt{\chi_3^2}$ 、この分布関数を G と書くことにする。また $\sqrt{\rho}$ Z の分布を $\sqrt{\rho} \sqrt{\chi_3^2}$ 、 $N(0,1-\rho)$ を $\sqrt{1-\rho} N(0,1)$ 、これらの convolution を $\sqrt{1-\rho} N(0,1) + \sqrt{\rho} \sqrt{\chi_3^2}$ 、この分布関数を G_ρ と書くことにする。exact 分布の表現から、

$$2(T_n - n/2)/\sqrt{n} \underset{a}{\sim} \sqrt{1-\rho} N(0,1) + \sqrt{\rho} \sqrt{\chi_3^2}$$

となるのはすぐに解るのであるが、ここでは別の approach を試みてみよう。

$$(T_n - n/2)/\sqrt{n}$$

$$= \sup_{0 \leq t \leq \bar{\rho}} (T_n(t) - T_n(0))/\sqrt{n} - (T_n(1) - T_n(0))/2\sqrt{n}$$

と変形できるが、 $\{(T_n(t) - T_n(0))/\sqrt{n}, 0 \leq t \leq 1\}$ は process として Wiener process W_t に、 $\bar{\rho}$ は ρ に弱収束するので、これより

$$2(T_n - n/2)/\sqrt{n} \underset{a}{\sim} 2 \sup_{0 \leq t \leq \rho} W_t - W_1$$

$$= (2 \sup_{0 \leq t \leq \rho} W_t - W_\rho) + (W_\rho - W_1)$$

を得る。明らかに第1項と第2項とは独立であり、第2項

は $\sqrt{1-\rho} N(0,1)$ に従う。 よって、第1項が $\sqrt{\rho} \sqrt{\chi^2_3}$ に従うことを示せばよい。 このために $Bt \equiv W_{t\rho} / \sqrt{\rho}$, $0 \leq t \leq 1$, $(M, B) \equiv (\sup_{0 \leq t \leq 1} Bt, B_1)$ と定義する。

Bt はまた Wiener process となり、この時の (M, B) の同時密度関数が次ぎで与えられる事はよく知られている。

$$f(m, b) = \begin{cases} \sqrt{2/\pi} (2m-b) \exp(-(2m-b)^2/2) & \text{if } m \geq 0, m \geq b \\ 0 & \text{otherwise} \end{cases}$$

これより $2M - B = (2 \sup_{0 \leq t \leq \rho} Wt - W_\rho) / \sqrt{\rho}$ の密度関数は

$$\sqrt{2/\pi} t^2 \exp(-t^2/2)$$

と計算できる。 これは $\sqrt{\chi^2_3}$ の密度関数である。

** Approximate Bahadur Efficiency の 計算 **

検定統計量 S_n の Bahadur slope は、

(1) $\theta = 0$ の下で、 S_n はある確率変数 S に弱収束し、

S の分布関数を F_S と置くとき、

$$\log(1 - F_S(x)) = (-a x^2/2) (1 + o(1))$$

となり、

(2) $\theta = \mu$ の下で、 $S_n / \sqrt{n}$ が $b(\mu)$ に確率収束するとき、

$a(b(\mu))^2$ で与えられる。

ここでは $S_n = (2/\sqrt{n})(T_n - n/2)$ と置く。さらに $\theta = 0$ の下での分布を parameter $(\Sigma, 0)$ の elliptical 分布として、 $\Sigma = I$ のときの X -座標の周辺分布を Ψ で表すことにする。

a の計算

T_n の定義から明らかに、

$$T_n(0) \leq T_n \leq T_n^*$$

である。よって

$$\begin{aligned} (2/\sqrt{n})(T_n(0) - n/2) &\leq S_n \\ &\leq (2/\sqrt{n})(T_n^* - n/2) \end{aligned}$$

これより次ぎの stochastic larger 性を得る。

$$N(0,1) \underset{St.}{<} \sqrt{1-\rho} N(0,1) + \sqrt{\rho} \sqrt{\chi^2_3} \underset{St.}{<} \sqrt{\chi^2_3}$$

これは分布関数で書けば

$$\Phi(x) \geq G_\rho(x) \geq G(x)$$

となるが、

$$\log(1 - \Phi(x)) = (-x^2/2)(1 + o(1))$$

$$\log(1 - G(x)) = (-x^2/2)(1 + o(1))$$

が成り立つのは明らかだろうから、次がえられる。

$$\log(1 - G_p(x)) = (-x^2/2)(1 + o(1))$$

つまり $a=1$ である。

$b(\mu)$ の計算

これを求めるために、 $\nu' \mu > 0$ であり、 $\nu' Z_i > 0$ となる Z_i の個数が丁度 $[n/2]$ 個となる様な方向 ν で、この長さが $|\Sigma^{-1} \mu|$ に等しくなるものを ν_n とおく。 $\mu \in C$ を変換した $\mu' \in C'$ に対する $R_{\mu'}$ を考えると、この内部に含まれる V_i と、 $\nu' Z_i > 0$ となる Z_i とは互いに 1 対 1 に対応するので、

$$W_n(\nu) \equiv (1/n) \sum_{i=1}^n I[\nu' X_i > 0]$$

と置くと、 $W_n(\nu_n) \leq (1/n) T_n$

であるという事が解る。 $\Sigma^{-1} \mu$ に直交する vector で上半平面にあるものを τ とするとき、領域 $\{x \mid \arg(\mu) \leq \arg(x) \leq \arg(\tau) \text{ or } \arg(-\mu) \leq \arg(x) \leq \arg(-\tau)\}$ の確率は elliptical 性から $1/2$ に等しいので、定義より ν_n は $\Sigma^{-1} \mu$ の一致推定量になっていることも解る。 また

$$(1/n) T_n \leq (1/n) \sup_{0 \leq t \leq 2} T_n(t)$$

$$= \sup_{\nu \neq 0} W_n(\nu)$$

である。 $W_n(\nu)$ は、 ν の argument の関数としてながめるとき 定数関数 $P(\nu'X_i > 0 \mid \theta = \mu) = \Psi(\nu'\mu / (\nu'\Sigma\nu)^{1/2})$ に弱収束する。このことより

$$\begin{aligned}\Psi(\Delta) &\leq \lim_{n \rightarrow \infty} (1/n) T_n \\ &\leq \sup_{\nu \neq 0} \Psi(\nu'\mu / (\nu'\Sigma\nu)^{1/2}) \\ &= \Psi(\Delta)\end{aligned}$$

を得る。ただし $\Delta = (\mu'\Sigma^{-1}\mu)^{1/2}$ である。ゆえに

$$\begin{aligned}b(\mu) &= \lim_{n \rightarrow \infty} (1/\sqrt{n}) S_n \\ &= \lim_{n \rightarrow \infty} (2/n) (T_n - n/2) \\ &= 2\Psi(\Delta) - 1\end{aligned}$$

となる。

以上により S_n の Bahadur slope は $(2\Psi(\Delta) - 1)^2$ となることが解る。これは結局、制約条件がない場合の Bahadur slope と同じである。よって、この量によっては、制約条件がある場合と、ない場合との差はみつけれない。また Hotelling の T^2 検定の slope は、分散が存在して Σ に等しいとき、 Δ^2 になるので、我々の検定の T^2 検定に対する

Bhadur efficiency は

$$(2\Psi(\Delta) - 1)^2 / \Delta^2$$

で得られる。これは 1 次元で $\Delta = \mu / \sigma$ と置いたときの t 検定に対する sign 検定の Bhadur efficiency に等しい。 $\Psi = \Phi$ で $\mu \rightarrow \infty$ としたとき、この値がその時の Pitman efficiency $2/\pi$ になるという事実も良く知られている。