8.3　2標本問題への応用

08T4083T 真下飛瑠

- 特性的な正定値カーネルkを用いると、平均 m_X^k によって確率分布が識別可能
  →2標本の均一性判定に正定値カーネルが適用可能

- 2標本の均一性判定とは
  - 2つのサンプル(X1,…,Xl)と(Y1,…,Yn)を発生させた分布が同じかどうかを判定する

  - 以下ではX1,…,XlとY1,…,Ynは可測空間 (X, B) に値をとり、それぞれ独立に確率分布P、Qに従う
  - P=Q：帰無仮説
  - P≠Q：対立仮説
  - として検定を行う

- P=Qかどうかは次式が0か否かによって判定する

$$M^2(P,Q) \equiv \left\| m_P^k - m_Q^k \right\|_{H_k}^2$$

- 0とみなされればP=Qとして帰無仮説が採択される

- 特性的な実正定値カーネルであるkは、X~P、Y~Qなる独立な変数X,Yに対して

$$E\left[k(X,Y)^2\right] < \infty$$

を満たすとする

- 検定統計量

$$\hat{M}_{l.n} = \left\| \hat{m}_P - \hat{m}_Q \right\|_{H_k}^2$$

$$= \frac{1}{l^2} \sum_{a,b=1}^{l} k(X_a, X_b) + \frac{1}{n^2} \sum_{c,d=1}^{n} k(Y_c, Y_d) - \frac{2}{nl} \sum_{a=1}^{l} \sum_{c=1}^{n} k(X_a, Y_c)$$

- これを不偏化

$$U_{l.n} = \frac{1}{l(l-1)} \sum_{a=1}^{l} \sum_{b \neq a} k(X_a, X_b) + \frac{1}{n(n-1)} \sum_{c=1}^{n} \sum_{d \neq c} k(Y_c, Y_d)$$

$$- \frac{2}{nl} \sum_{a=1}^{l} \sum_{c=1}^{n} k(X_a, Y_c)$$

- また、$U_{l,n}$ はカーネルによる2標本U-統計量にもなる

$$h(x_1, x_2; y_1, y_2) = k(x_1, x_2) + k(y_1, y_2) - \frac{1}{2}\{k(x_1, y_1) + k(x_1, y_2) + k(x_2, y_1) + k(x_2, y_2)\}$$

- まず、仮説検定を行うために、帰無仮説P=Qのもとでの検定統計量Uの分布を知りたい
- $N = l + n$(総データ)とおき、

$$\frac{l}{N} \to \gamma, \qquad \frac{n}{N} \to 1 - \gamma \qquad (N \to \infty)$$

を仮定

- P=Qという帰無仮説のもと、Nを無限大にしたときの漸近分布は以下のように与えられる

$$NU_{l,n} \quad \Rightarrow \quad \sum_{i=1}^{\infty} \lambda_i \left(Z_i^2 - \frac{1}{\gamma(1-\gamma)} \right) \qquad (n \to \infty) \qquad (8.7)$$

Z_i: 正規分布 $\mathrm{N}\left(0, \frac{1}{\gamma(1-\gamma)}\right)$に従う独立な確率変数

$\{\lambda_i\}_{i=1}^{\infty}$:

$$\tilde{k}(x,y) = k(x,y) - E\left[k(x,X)\right] - E\left[k(X,y)\right] + E\left[k(X,\tilde{X})\right] \qquad (8.8)$$

を積分核にもつ$L^2(P)$上の積分作用素の非零固有値を重複度だけ並べたもの

ある単位ベクトル$\phi_i \in L^2(P)$に対して

$$\int \tilde{k}(x,y)\phi_i(y)dP(y) = \lambda_i \phi_i(x) \qquad (8.9)$$

を満たす非負実数λ_iを重複度だけ並べたもの

- kが特性的な場合、対立仮説P≠Qのもとでは

$$M^2(P,Q) \neq 0$$

であり、$\sqrt{N}\left(U_{l,m} - M^2(P,Q)\right)$ は正の分散をもつ正規分布となる

定理8.12

$$\sqrt{N}\left(U_{l,n} - M^2(P,Q)\right) \quad \Rightarrow \quad N\left(0,\sigma^2\right) \qquad \left(N \to \infty\right)$$

- 特に、$NU_{l,n}$ による検定は一致性をもつ

- 以上から、λ iが決定できれば棄却域が決定できる

- 式(8.8)の積分核は中心化された正定値カーネルに一致
  →固有値 λ iの一致推定量は中心化グラム行列 $\widetilde{K}$ の固有値によって与えられることがわかる

  → $\widetilde{K}$ の固有値を求め、カイ2乗分布に従うn-1個の独立なサンプルを発生させることにより、式(8.7)の極限分布の α -%点の近似値を求める

例

N＼a	M(P,Q) 1	0.75	0.5	0.25	0	Kolmogorov-Smirnov 1	0.75	0.5	0.25	0
200	0.996	0.898	0.788	0.964	0.882	0.962	0.910	0.730	0.956	0.940
500	0.990	0.868	0.544	0.118	0.038	0.990	0.752	0.382	0.112	0.124
1000	0.986	0.976	0.704	0.088	0	0.954	0.950	0.796	0.316	0.002

$$Q_a : \quad a\sqrt{\frac{3}{2\pi}}e^{-\frac{3}{2}x^2} + (1-a)\frac{1}{2}I_{[-1,1]}(x)$$

P:正規分布N(0,1/3)
Qa:区間[-1,1]上の一様分布とN(0,1/3)の混合分布
有意水準α=5%
とし、aを変化させてM(P,Q)による検定した場合の結果