再生核ヒルベルト空間における射影のモーメントによる二標本検定

千葉大・融合理工学府　牧草 夏実

P, Q をヒルベルト空間 $\mathcal{H}$ 上の確率分布とするとき，二標本 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Q$ に基づく検定

$$\text{帰無仮説 } H_0 : P = Q \quad \text{vs.} \quad \text{対立仮説 } H_1 : P \neq Q$$

を考える．

確率変数 $X \sim P$, $Y \sim Q$ を正定値カーネル k によって，この k に対応する再生核ヒルベルト空間 $H(k)$ 上に，それぞれ $k(\cdot, X)$，$k(\cdot, Y)$ により変換を行う．このとき，この $k(\cdot, X)$，$k(\cdot Y)$ の平均まわりの2次モーメント $\Sigma_k(P)$，$\Sigma_k(Q)$ は，それぞれヒルベルト空間 $H(k)^{\otimes 2} = H(k) \otimes H(k)$ での期待値 $\Sigma_k(P) = \mathbb{E}_{X\sim P}[(k(\cdot, X) - \mu(P))^{\otimes 2}]$，$\Sigma_k(Q) = \mathbb{E}_{Y\sim Q}[(k(\cdot, Y) - \mu(Q))^{\otimes 2}]$ によって定められている．ここで，$\mu(P)$，$\mu(Q)$ は $k(\cdot, X)$ の1次モーメント $\mu(P) = \mathbb{E}_{X\sim P}[k(\cdot, X)]$，$\mu(Q) = \mathbb{E}_{Y\sim Q}[k(\cdot, Y)]$ であり，$\otimes$ はテンソル積を表しており，任意の $f \in H(k)$ に対し，$f^{\otimes 2} = f \otimes f = \langle f, \cdot \rangle_{H(k)} f$ である．

この $k(\cdot, X)$ と $k(\cdot, Y)$ の $f \in H(k)$ への射影のモーメント差

$$\sup_{\|f\|_{H(k)}=1} |\langle f, \mu(P) - \mu(Q) \rangle_{H(k)}| = \|\mu(P) - \mu(Q)\|_{H(k)}$$

により，2つの分布の違いを測るのが，Maximum Mean Discrepancy (MMD) と呼ばれるものである．同様の考え方により，$(k(\cdot, X) - \mu(P))^{\otimes 2}$ と $(k(\cdot, Y) - \mu(Q))^{\otimes 2}$ の $A \in H(k)^{\otimes 2}$ への射影のモーメント差

$$\sup_{\|A\|_{H(k)^{\otimes 2}}=1} |\langle A, \Sigma_k(P) - \Sigma_k(Q) \rangle_{H(k)^{\otimes 2}}| = \|\Sigma_k(P) - \Sigma_k(Q)\|_{H(k)^{\otimes 2}}$$

により2つの分布の違いを測る．これは，MMD のようなある種の分布の違いを測っている．この違い $\|\Sigma_k(P) - \Sigma_k(Q)\|^2_{H(k)^{\otimes 2}}$ は

$$\widehat{T}^2 = \left\| \widehat{\Sigma}_k(P) - \widehat{\Sigma}_k(Q) \right\|^2_{H(k)^{\otimes 2}}$$

によって推定することができる．ただし，

$$\widehat{\Sigma}_k(P) = \frac{1}{n} \sum_{i=1}^{n} (k(\cdot, X_i) - \widehat{\mu}(P))^{\otimes 2}, \quad \widehat{\mu}(P) = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i),$$

$$\widehat{\Sigma}_k(Q) = \frac{1}{m} \sum_{i=1}^{m} (k(\cdot, Y_i) - \widehat{\mu}(Q))^{\otimes 2}, \quad \widehat{\mu}(Q) = \frac{1}{m} \sum_{i=1}^{m} k(\cdot, Y_i)$$

である．

本発表では，この $\|\Sigma_k(P) - \Sigma_k(Q)\|_{H(k)^{\otimes 2}}$ による二標本検定について，$\widehat{T}^2$ の帰無仮説の下での漸近分布，対立仮説の下での漸近分布の導出を行う．また，$\|\Sigma_k(P) - \Sigma_k(Q)\|_{H(k)^{\otimes 2}}$ と再生核との関連について，MMD と再生核との関連と対応させた議論を行う．さらに，MMD による二標本検定との比較を行う．