Kernel Method and its Application

横浜市立大学　生命ナノシステム科学研究科　牧草 夏実
Natsumi Makigusa
Graduate School of Nanobioscience,
Yokohama City University

1　序

再生核ヒルベルト空間の理論を用いた統計解析手法にカーネル法と呼ばれる手法がある．カーネル法は様々な統計解析手法と組み合わせて用いられているが，本講究録では，適合度検定について述べる．特に，2 標本の均一性の検定について，再生核ヒルベルト空間における分散の違いを調べることにより，2 標本の分布間の違いを測る手法に基づく仮説検定について述べ，その漸近挙動について紹介する．この検定統計量の漸近挙動および，シミュレーション，実データへの適用結果の詳細は [6] にまとめられているので，詳細は，そちらを参照されたい．

2　カーネル法

カーネル法とは，1990 年代半ばごろから発展したデータ解析手法である．サポートベクターマシンと呼ばれる，2 値判別を与える手法において，非線形化を行うために，再生核ヒルベルト空間の理論を取り入れることが可能なカーネル関数を用いたことが，カーネル法の発端となっている．カーネル法は，正定値カーネルと呼ばれる関数により，データをより高次な再生核ヒルベルト空間と呼ばれる空間に写し，線形手法を用いる手法である．カーネル関数による非線形な変換によって，非線形な解析を行うことや，サンプル数よりも次元数のほうが大きい，高次元小標本データを解析することが可能であることが特徴である．サポートベクターマシン以外にも，線形判別分析，主成分分析，リッジ回帰，

適合度検定，ニューラルネットワーク等様々な手法と組み合わせて用いられている．

この節では，カーネル法においてデータを変換する関数である，正定値カーネル，および写した先の空間である，再生核ヒルベルト空間の定義を述べ，正定値カーネルと再生核ヒルベルト空間の関連性について述べる．

初めに，正定値カーネルおよび再生核ヒルベルト空間の定義は次の通りである．

定義 2.1（正定値カーネル） $\mathcal{X}$ を集合とするとき，次の 2 条件を満たすカーネル $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ を ($\mathcal{X}$ 上の) 正定値カーネル (positive definite kernel) という．

- (対称性) 任意の $x, y \in \mathcal{X}$ に対し $k(x, y) = k(y, x)$,
- (正値性) 任意の $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathbb{R}$ に対し

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

定義 2.2（再生核ヒルベルト空間） 集合 $\mathcal{X}$ 上の再生核ヒルベルト空間 (reproducing kernel Hilbert space, RKHS) とは，$\mathcal{X}$ 上の関数からなるヒルベルト空間 $\mathcal{H}$ で，任意の $x \in \mathcal{X}$ に対して $k_x \in \mathcal{H}$ が存在し，任意の $f \in \mathcal{H}$ に対して

$$\text{(再生性)} \quad \langle f, k_x \rangle_{\mathcal{H}} = f(x)$$

を満たすものをいう．ここで，$\langle \cdot, \cdot \rangle_{\mathcal{H}}$ はヒルベルト空間 $\mathcal{H}$ の内積である．

再生核ヒルベルト空間の定義の，k_x に対して，$k(y, x) = k_x(y)$ により定まるカーネル k を $\mathcal{H}$ の再生核 (reproducing kernel) という．

$\mathcal{X}$ 上の再生核ヒルベルト空間の再生核 k は $\mathcal{X}$ 上の正定値カーネルであり，その再生核が一意であることはすぐに導出される事実である．逆に，正定値カーネルが与えられているときに，再生核ヒルベルト空間が一意に存在していることについては，[1] で次の定理が述べられている．

定理 2.1（Moore–Aronszajn's Theorem [1]） k を $\mathcal{X}$ 上の正定値カーネルとする．このとき，k を再生核とする $\mathcal{X}$ 上の再生核ヒルベルト空間が一意に存在する．

定理 2.1 により，正定値カーネルと再生核ヒルベルト空間が 1 対 1 に対応していることが知られており，統計解析におけるカーネル法理論の根幹となっている．

カーネル法は，このような理論的背景により，データ $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ を正定値カーネル k により関数 $k(\cdot, \boldsymbol{X}_1), \ldots, k(\cdot, \boldsymbol{X}_n) \in \mathcal{H}_k$ に変換し，$\mathcal{H}_k$ 上の内積および再生性を用

いて統計的解析を行う．ここで，$\mathcal{H}_k$ は k に対応する再生核ヒルベルト空間を表すものとする．

3　カーネル法を用いた適合度検定

適合度検定とは，1 標本の枠組みでは，未知の分布 P と既知の分布 Q，データ $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ について，帰無仮説 $H_0 : P = Q$ と対立仮説 $H_1 : P \neq Q$ に基づく仮説検定であり，2 標本の枠組みでは，未知の分布 P, Q，データ $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Q$ について，帰無仮説 $H_0 : P = Q$ と対立仮説 $H_1 : P \neq Q$ に基づく仮説検定である．

本講究録では，P, Q をヒルベルト空間 $\mathcal{H}$ 上の確率分布とするとき，2 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．ユークリッド空間での二標本検定はすでに様々な検定方法が議論されているが，ヒルベルト空間に値をとる確率変数に対する二標本検定を考えることで，高次元データに対する二標本検定の議論を与える．カーネル法は，このような高次元データに対するアプローチの 1 つであり，カーネル法を用いた二標本検定として，Maximum Mean Discrepancy(MMD) に基づく二標本検定が [4] により既に議論されている．この MMD と同様の考え方により，Maximum Variance Discrepancy(MVD) という新たな指標を考え，この MVD による二標本検定を考える．

カーネル法ではデータ $X_1, \ldots, X_n \in \mathcal{H}$ を正定値カーネル k によって，関数 $k(\cdot, X_1), \ldots, k(\cdot, X_1) \in \mathcal{H}_k$ に変換し解析を行うが，データの分布を正定値カーネルで写したものは，カーネル平均埋め込みと呼ばれ，正定値カーネル k による分布 P の $\mathcal{H}_k$ への埋め込み $\mu(P)$ は Bochner 積分を用いて

$$\mu(P) = \mathbb{E}_{X \sim P}[k(\cdot, X)] = \int_{\mathcal{H}} k(\cdot, x) dP(x)$$

により定義される．とくに，分布 P を単射に埋め込むようなカーネルは characteristic kernel と呼ばれる．

定義 3.1（characteristic kernel, [3]） $k : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ を正定値カーネルとする．$\mu : P \mapsto \mu(P)$ が単射であるとき，k は characteristic kernel であるという．

ガウスカーネル ($k(s,t) = \exp(-\sigma \|s - t\|^2_{\mathcal{H}})$, $\sigma > 0$) やラプラスカーネル ($k(s,t) =$

$\exp(-\beta\|s-t\|_{\mathcal{H}}),\ \beta>0)$ は characteristi kernel であることが知られている．

確率分布 P,Q 間の違いを測る指標である Maximum Mean Discrepancy(MMD) は [4] において次の式で定義されている．

定義 3.2（Maximum Mean Discrepancy, [4]） $\mathcal{F}$ を関数 $f:\mathcal{X}\to\mathbb{R}$ の集合とする．このとき，

$$\mathrm{MMD}[\mathcal{F},P,Q]=\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{X\sim P}[f(X)]-\mathbb{E}_{Y\sim Q}[f(Y)]\right)$$

を Maximum Mean Discrepancy という．

MMD はある関数のクラス $\mathcal{F}$ における期待値の最大の違いを意味している．この関数のクラス $\mathcal{F}$ を characteristic kernel に対応する再生核ヒルベルト空間 $\mathcal{H}_k$ の単位球とする．このとき，MMD はカーネル平均埋め込み $\mu(P),\mu(Q)$ によって次のように得られる．

補題 3.1（[4]） $\mathcal{F}$ を characteristic kernel に対応する再生核ヒルベルト空間の単位球とし，$\mathbb{E}_{X\sim P}[\sqrt{k(X,X)}]<\infty$ とする．このとき，

$$\mathrm{MMD}[\mathcal{F},P,Q]=\sup_{\|f\|_{\mathcal{H}_k}\le 1}\langle\mu(P)-\mu(Q),f\rangle_{\mathcal{H}_k}=\|\mu(P)-\mu(Q)\|_{\mathcal{H}_k}\tag{1}$$

である．

補題 3.1 より，characteristic kernel による MMD とは，分布 P,Q をカーネル平均 $\mu(P),\mu(Q)$ によって $\mathcal{H}_k$ 内に 1 対 1 に埋め込み，この $\mathcal{H}_k$ 上で $\mu(P)$ と $\mu(Q)$ のノルムに関する距離として，分布 P,Q の距離を定めていることがわかる．

本稿では，この characteristic kernel による MMD と同様の考えで Maximum Variance Discrepancy(MVD) と呼ばれる分布の違いを測る指標を定め，MVD による二標本検定についての漸近的結果を紹介する．

MMD では期待値の違いを測っていたのに対して，MVD では分散の違いを測ることを考える．まず，$k(\cdot,X)$ の分散は，ヒルベルト空間の分散の定義に基づいて，

$$V_{X\sim P}[k(\cdot,X)]=\mathbb{E}_{X\sim P}\left[(k(\cdot,X)-\mu(P))^{\otimes 2}\right]\in\mathcal{H}_k^{\otimes 2}$$

である．ここで，任意の $h\in\mathcal{H}_k$ に対し，テンソル積 $h^{\otimes 2}$ は作用素 $\mathcal{H}_k\to\mathcal{H}_k,\ \ h'\mapsto\langle h,h'\rangle_{\mathcal{H}_k}h$ により定められているものであり，ヒルベルト空間 $\mathcal{H}_1,\mathcal{H}_2$ の任意の元 $h_1\in\mathcal{H}_1,\ h_2\in\mathcal{H}_2$ に対し，テンソル積 $h_1\otimes h_2$ は作用素 $\mathcal{H}_2\to\mathcal{H}_1,\ \ h'\mapsto\langle h',h_2\rangle_{\mathcal{H}_2}h_1$ により定められるもののことを表している．$\mathcal{H}_k^{\otimes 2}$ はテンソル積空間 $\mathcal{H}_k\otimes\mathcal{H}_k$ である．

$k(\cdot, X), k(\cdot, Y)$ の分散をそれぞれ $\Sigma_k(P), \Sigma_k(Q)$ とする．すなわち，

$$\Sigma_k(P) = \mathbb{E}_{X\sim P}\left[(k(\cdot, X) - \mu(P))^{\otimes 2}\right],$$
$$\Sigma_k(Q) = \mathbb{E}_{Y\sim Q}\left[(k(\cdot, Y) - \mu(Q))^{\otimes 2}\right]$$

とする．A を $\mathcal{H}_k^{\otimes 2}$ 内のノルムが 1 の作用素とする．このとき，分布 P, Q の違いを測る指標 MVD は次のように定められる ([6]).

$$\mathrm{MVD}[P, Q] = \sup_{\|A\|_{\mathcal{H}_k^{\otimes 2}} \le 1} \langle A, \Sigma_k(P) - \Sigma_k(Q)\rangle_{\mathcal{H}_k^{\otimes 2}} = \|\Sigma_k(P) - \Sigma_k(Q)\|_{\mathcal{H}_k^{\otimes 2}}.$$

この MVD を用いた 2 標本の均一性の検定を考える．2 標本 $X_1, \ldots, X_n \in \mathcal{H}$, $Y_1, \ldots, Y_m \in \mathcal{H}$ がそれぞれ独立に同一の分布 P, Q に従っているとする．すなわち，$X_1, \ldots, X_n \overset{i.i.d.}{\sim} P, Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Q$ に基づく

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

の仮説検定を考える．[2] では，カーネル k で写す前の空間において，共分散作用素が等しいかどうかの検定を考えているが，本稿では，カーネル k で写した先で，共分散作用素が等しいかどうかを調べることによって，分布 P, Q が等しいかどうかに着目している．

$T^2 = \mathrm{MVD}[P, Q]^2$ とすると，T^2 は

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

によって推定することができる．ここで，

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

である．この検定統計量 $\widehat{T}^2_{n,m}$ について，その漸近帰無分布は次の定理で得られている．

定理 3.1（[6]） $\mathbb{E}_{X\sim P}[k(X, X)^2] < \infty$, $\mathbb{E}_{Y\sim Q}[k(Y, Y)^2] < \infty$, $\lim_{n,m\to\infty} n/(n+m) \to \rho$, $0 < \rho < 1$ とする．このとき，$H_0 : P = Q$ のもとで，

$$(n+m)\widehat{T}^2_{n,m} \underset{D}{\to} \frac{1}{\rho(1-\rho)}\sum_{\ell=1}^{\infty}\lambda_\ell Z_\ell^2, \ n, m \to \infty$$

である．ここで，$Z_\ell \overset{i.i.d.}{\sim} N(0, 1)$ であり，λ_ℓ は作用素 $V_{X\sim P}[(k(\cdot, X) - \mu(P))^{\otimes 2}]$ の固有値であり，“$\underset{D}{\to}$” は分布収束を表している．

詳細および証明は [6] 参照.

命題 3.1（[6]） $\mathbb{E}_{X\sim P}[k(X,X)^2] < \infty$, $\mathbb{E}_{Y\sim Q}[k(Y,Y)^2] < \infty$, $\lim_{n,m\to\infty} n/(n+m) \to \rho$, $0<\rho<1$ とする．また，カーネル $h : \mathcal{H}\times\mathcal{H} \to \mathbb{R}$ を

$$h(x,y) = \left\langle (k(\cdot,x)-\mu(P))^{\otimes 2} - \Sigma_k(P), (k(\cdot,y)-\mu(P))^{\otimes 2} - \Sigma_k(P) \right\rangle_{\mathcal{H}_k^{\otimes 2}}, \quad x,y\in\mathcal{H}$$

とし，

$$h(\cdot,x) = (k(\cdot,x)-\mu(P))^{\otimes 2} - \Sigma_k(P) \in \mathcal{H}_k^{\otimes 2}$$

とする．また，自己共役作用素 $S_k : L_2(\mathcal{H},P) \to L_2(\mathcal{H},P)$ を

$$S_k g(x) = \int_{\mathcal{H}} h(x,y)g(y)dP(y), \quad g\in L_2(\mathcal{H},P) \tag{2}$$

とする．このとき，定理 3.1 の $V_{X\sim P}[(k(\cdot,X)-\mu(P))^{\otimes 2}]$ の固有値と (2) の S_k の固有値は同じである．

検定において，帰無分布が得られれば，帰無分布の上側 $\alpha\%$ 点を棄却点とすることで，帰無仮説が真であるにもかかわらず棄却してしまう誤りの確率 (Type I error の確率) を $\alpha\%$ に抑えたうえで，対立仮説が真であるにもかかわらず帰無仮説を採用してしまう誤りの確率 (Type II error の確率) を小さくするように考えることができる．実際には，特定のデータ数 n における帰無分布を得ることは困難であるため，この MVD に基づく 2 標本検定では，漸近帰無分布に基づき棄却点を決定することとなる．しかし，定理 3.1 で得られている漸近帰無分布に現れる，無限個の重み λ_ℓ それぞれを求めることは困難となった．そのため，[5] で報告されている *Spec* と呼ばれる手法と同様の方法を考える．

次の定理では，$V[(k(\cdot,X)-\mu(P))^{\otimes 2}]$ の推定量の固有値によって構成した分布が，漸近帰無分布に分布収束することを示している．

定理 3.2（[6]） $\mathbb{E}_{X\sim P}[k(X,X)^2] < \infty$ とする．

$$\Upsilon = V[(k(\cdot,X)-\mu(P))^{\otimes 2}], \quad \widehat{\Upsilon}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left\{(k(\cdot,X_i)-\widehat{\mu}(P))^{\otimes 2} - \widehat{\Sigma}_k(P)\right\}^{\otimes 2}$$

とし，$\{\lambda_\ell\}_{\ell=1}^{\infty}$ と $\{\widehat{\lambda}_\ell^{(n)}\}_{\ell=1}^{\infty}$ をそれぞれ Υ, $\widehat{\Upsilon}^{(n)}$ の固有値とする．このとき，

$$\sum_{\ell=1}^{\infty}\widehat{\lambda}_\ell^{(n)} Z_\ell^2 \underset{D}{\to} \sum_{\ell=1}^{\infty}\lambda_\ell Z_\ell^2, \quad n\to\infty$$

である．ここで，$Z_\ell \overset{i.i.d.}{\sim} N(0,1)$ である．

この結果により，$\widehat{\lambda}_{\ell}^{(n)}$ を計算できれば良いことがわかる．次の命題では，あるグラム行列 H の固有値と $\widehat{\Upsilon}^{(n)}$ の固有値が同じであることを主張している．

命題 3.2（[6]） グラム行列 $H = (H_{ij})_{1\leq i,j\leq n}$ を

$$H_{ij} = \left\langle (k(\cdot, X_i) - \widehat{\mu}(P))^{\otimes 2} - \widehat{\Sigma}_k, (k(\cdot, X_j) - \widehat{\mu}(P))^{\otimes 2} - \widehat{\Sigma}_k(P) \right\rangle_{\mathcal{H}_k^{\otimes 2}}$$

とする．このとき，H/n の固有値と $\widehat{\Upsilon}^{(n)}$ の固有値は同じである．

注意 3.1 命題 3.2 により，H/n の固有値 $\widehat{\lambda}_{\ell}^{(n)}$，$\ell = 1, \ldots, n-1$ を用いて $1/\{\rho(1-\rho)\}\sum_{\ell=1}^{n-1}\widehat{\lambda}_{\ell}^{(n)} Z_{\ell}^2$ を計算することによって棄却点を得ることができる．さらに，行列 H は

$$H = P_n(\widetilde{K}_X \odot \widetilde{K}_X)P_n$$

と展開することができるので，中心化グラム行列 $\widetilde{K}_X$ から計算することができる．ここで，$\odot$ はアダマール積を表している．

参考文献

[1] Aronszajn, N., Theory of reproducing kernels, Trans. Amer. Math. Soc., **68** (1950), 337–404.

[2] Boente, G., Rodriguez, D. and Sued, M., Testing equality between several populations covariance operators, Ann. Inst. Statist. Math., **70** (2018), 919–950.

[3] Fukumizu, K., Gretton, A., Sun, X. and Schö lkopf, B., Kernel measures of conditional dependence, Neural Inf. Process. Ser., **20** (2008).

[4] Gretton, A., Borgwardt, K., Rasch M., Schölkopf, B. and Smola A., A kernel method for the two sample problem, Neural Inf. Process. Ser., **19** (2007), 513–520.

[5] Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B. K., A Fast, Consistent Kernel Two-Sample Test, Neural Inf. Process. Ser., **22** (2009).

[6] Makigusa, N., Two-sample test based on maximum variance discrepancy; arXiv:math.ST/2012.00980.