統計数理(2010)
第 58 巻 第 2 号 185–206

特集「統計的機械学習」
[研究詳解]

正定値カーネルによるノンパラメトリック推論

福水 健次†

(受付 2010 年 5 月 6 日；改訂 6 月 22 日；採択 6 月 29 日)

要　　旨

正定値カーネルないしは再生核ヒルベルト空間を用いたデータ解析の方法論である，いわゆる「カーネル法」は，データを再生核ヒルベルトに写像し，この空間(特徴空間)において線形のデータ解析手法を適用する点に特徴があり，さまざまな手法のカーネル化が提案されてきた．最近になって，再生核ヒルベルト空間において平均や分散といった基本的な統計量を考えることによって，分布の均一性，独立性，条件付独立性といった統計的概念が，カーネル法によって扱えることがわかってきた．本論文では，この新しいノンパラメトリック推論の方法論の基本的な考え方を説明し，特に独立性や条件付独立性に関して今まで得られている結果の概要を解説する．

キーワード：正定値カーネル，再生核，ヒルベルト空間，ノンパラメトリック，独立性，条件付独立性．

1. はじめに

本論文は，正定値カーネルないしは再生核ヒルベルト空間を用いたデータ解析の方法論，いわゆる「カーネル法」の最近の展開である，正定値カーネルによって分布の性質をノンパラメトリックに推論する方法に関して紹介する．特に，分布の均一性，独立性，条件付独立性を議論するための正定値カーネルの方法に関する研究を解説する．

「カーネル」という用語は，カーネル密度推定をはじめとして，必ずしも正定値性を仮定しないカーネル関数を意味するものとして古くから統計学で用いられてきた．しかし「カーネル法」という呼び名が正定値カーネルによる方法を指すものとして既に広く普及しているため，本稿でも「カーネル法」と呼ぶことにする．

カーネル法は，サポートベクターマシン(Boser et al., 1992)が注目された 1990 年代の半ばから，主として計算機科学の分野で急速に発展したデータ解析の方法論である．その後，主成分分析，Fisher 判別分析，正準相関分析など，さまざまな線形のデータ解析手法がカーネルにより非線形化され，カーネル法の研究が盛んとなった．

カーネル法の新規性は，カーネルの正定値性を積極的にデータ解析に応用し，系統的手法を構築した点にある．正定値カーネルは再生核ヒルベルト空間というクラスの関数空間を定め，データをこの関数空間に写像する標準的方法が定まる．さらに，この関数空間の特別な内積を用いてデータ解析アルゴリズムを構築することにより，効率的な計算によって高次元データの高次モーメントを扱うことが可能となる．計算機科学で発達した方法ではあるが，一方でカーネル法は古典的な多変量解析の自然な拡張という側面も持つ．本論文は後者の立場に立った解

† 統計数理研究所：〒190–8562 東京都立川市緑町 10–3

説を試みる.

本論文は,まず 2 章においてカーネル法全般の基本的考えを述べ,3 章で,再生核ヒルベルト空間における平均によって,確率分布を一意に定めることが可能であることを説明する.4 章では 2 つの確率変数の独立性を特徴づけるための方法を紹介し,5 章で条件付独立性の特徴づけについて述べる.

2. カーネル法の概要

カーネル法は,データを(非線形)写像することによってデータの高次モーメントを扱う方法論である.データに何らかの変換を施してから解析する手法は古くから存在するが,カーネル法の特徴は,特殊な内積を持つ関数空間への写像を用いることにより,写像後のデータに対する線形の処理が効率的に行える点にある.本論文ではカーネル法の一般的方法論に関しては簡単に触れるだけなので,より詳しく知りたい読者は,例えば Schölkopf and Smola (2002) や福水(2010) などを見ていただきたい.

2.1 正定値カーネルと再生核ヒルベルト空間

データの写像に用いる空間を導入するために,正定値カーネルとそれが定める再生核ヒルベルト空間についてまとめておく(詳しくは Aronszajn, 1950 参照).なお,以下では実数値カーネルの場合のみを説明する.

集合 Ω に対し,$k:\Omega\times\Omega\to\mathbb{R}$ が Ω 上の**正定値カーネル**であるとは,対称性 $k(x,y)=k(y,x)$ を満たし,かつ任意の n 個の点 $x_1,\ldots,x_n\in\Omega$ と実数 $c_1,\ldots,c_n$ に対し,

$$\sum_{i,j=1}^{n} c_i c_j k(x_i,x_j)\geq 0 \tag{2.1}$$

が成り立つことをいう.行列 $(k(x_i,x_j))$ は**グラム行列**と呼ばれる.

Ω 上の正定値カーネル k に対し,Ω 上の実関数からなる(実)ヒルベルト空間 $\mathcal{H}$ で,以下の 2 つの性質を満たすものが一意的に存在する.

(i) 任意の $x\in\Omega$ に対して $k(\cdot,x)\in\mathcal{H}$ であり,$\{k(\cdot,x)\in\mathcal{H}\,|\,x\in\Omega\}$ の張る線形空間は $\mathcal{H}$ で稠密である.

(ii) 任意の $f\in\mathcal{H}$ と $x\in\Omega$ に対し,再生性

$$\langle f,k(\cdot,x)\rangle_{\mathcal{H}}=f(x) \tag{2.2}$$

が成り立つ.ここで $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ は $\mathcal{H}$ の内積を表す.

このようなヒルベルト空間のことを(k が定める)**再生核ヒルベルト空間**といい,$(\mathcal{H},k)$ であらわす.(ii)の再生性は再生核ヒルベルト空間をデータ解析に応用する上で最も重要な性質であり,ヒルベルト空間内での内積計算を容易にする.例えば $f=\sum_{i=1}^{n}a_i k(\cdot,x_i)$ と $g=\sum_{j=1}^{m}b_j k(\cdot,y_j)$ という 2 つの $\mathcal{H}$ の要素の内積は

$$\langle f,g\rangle_{\mathcal{H}}=\sum_{i=1}^{n}\sum_{j=1}^{m}a_i b_j k(x_i,y_j)$$

で与えられ,k の値の評価に還元される.これは,内積計算に積分を必要とする 2 乗可積分関数のなす関数空間などと大きく異なる点である.

k_n $(n=1,2,\ldots)$ を Ω 上の正定値カーネルとするとき,以下で定義される関数がまた正定値カーネルとなることは,比較的容易に示される.(i)非負結合 $c_1k_1+c_2k_2$ $(c_1,c_2\geq 0)$,(ii)積 k_1k_2,(iii)各点収束先 $k(x_1,x_2)=\lim_{n\to\infty}k_n(x_1,x_2)$(各点収束を仮定する).

$\mathcal{X}$ 上の再生核ヒルベルト空間 $(\mathcal{H}_1,k_1),(\mathcal{H}_2,k_2)$ に対し，k_1+k_2 により定まる再生核ヒルベルト空間は，ベクトル空間として $f+g$ $(f\in\mathcal{H}_1, g\in\mathcal{H}_2)$ の形の関数 ($f+g$ は関数値の和で定義する) からなることが知られている．これを $\mathcal{H}_1$ と $\mathcal{H}_2$ の直和といい，$\mathcal{H}_1+\mathcal{H}_2$ で表す．また，$(\mathcal{H}_1,k_1),(\mathcal{H}_2,k_2)$ をそれぞれ $\mathcal{X},\mathcal{Y}$ 上の再生核ヒルベルト空間とするとき，積 k_1k_2 の定める $\mathcal{X}\times\mathcal{Y}$ 上の再生核ヒルベルト空間はテンソル積 $\mathcal{H}_1\otimes\mathcal{H}_2$ と一致し，$\sum_{i=1}^n f_i(x)g_i(y)$ $(f_i\in\mathcal{H}_1, g_i\in\mathcal{H}_2)$ の形の関数集合は $\mathcal{H}_1\otimes\mathcal{H}_2$ で稠密である．

ユークリッド空間 $\mathbb{R}^m$ 上の正定値カーネルの代表的な例は，通常の内積 $k(x_1,x_2)=x_1^Tx_2$ のほかに，多項式カーネル

$$k_{d,c}^{poly}(x_1,x_2)=(x_1^Tx_2+c)^d$$

$(c\geq 0,\ d\in\mathbb{N})$ や，ガウス RBF (Radial Basis Function) カーネル

$$k_\sigma^G(x_1,x_2)=\exp\Big(-\frac{\|x_1-x_2\|^2}{2\sigma^2}\Big)$$

$(\sigma>0)$ などである．これらが正定値であることは，上で述べた 3 つの性質を用いると比較的容易に証明できる．また，多項式カーネル $k_{d,c}^{poly}$ $(c>0)$ の定める再生核ヒルベルト空間は，ベクトル空間として d 次以下の多項式全体と一致することが示される．内積は $k_{d,c}^{poly}$ により定まる．ガウス RBF カーネルが定める再生核ヒルベルト空間は無限次元になることが知られている．

2.2 正定値カーネルによるデータ解析の方法論

正定値カーネルおよび再生核ヒルベルト空間をデータ解析に用いる方法について述べる．データ $x_1,\ldots,x_n$ が集合 Ω の点として与えられているとする．これに対して Ω 上の正定値カーネル k とそれが定める再生核ヒルベルト空間 $\mathcal{H}$ を用意し，特徴写像

$$\Phi:\Omega\to\mathcal{H},\qquad x\mapsto k(\cdot,x) \tag{2.3}$$

によって，関数データ $\{\Phi(x_i)\}_{i=1}^n=\{k(\cdot,x_i)\}_{i=1}^n$ を作成する．例えば，ガウス RBF カーネルを用いると，$\{\Phi(x_i)=e^{-\frac{1}{2\sigma^2}\|x-x_i\|^2}\}_{i=1}^n$ という関数データを得る．

カーネル法の方法論の核心は，$\mathbb{R}^m$ のベクトルデータに対して適用可能な手法を，関数データ $\{\Phi(x_i)\}_{i=1}^n$ に拡張するというものである．この方法論は線形手法のカーネル化と呼ばれ，主成分分析，フィッシャー (Fisher) 判別分析，正準相関分析など様々な手法のカーネル化が行われてきた．SVM も，マージン最大化を尺度とする線形識別器のカーネル化として定義される (Schölkopf and Smola, 2002)．

近年になって，もっと基本的な平均や分散といった統計量を再生核ヒルベルト空間上で考えることによって，分布の同一性や独立性といった古典的な統計的概念を扱えることが明らかとなり，それに基づいたノンパラメトリックな統計的推論手法が開発されてきた．次章からそのような方法論に関して解説する．

3. 平均による確率分布の特徴づけ

まず再生核ヒルベルト空間上の平均を定義し，それによって確率分布を特徴づけることが可能であることを説明する．

3.1 再生核ヒルベルト空間における平均

$(\mathcal{X},\mathcal{B}_\mathcal{X})$ を可測空間とし，以下 $\mathcal{X}$ 上のカーネル k は常に $\mathcal{X}\times\mathcal{X}$ 上の可測関数であると仮定する．k に対応する再生核ヒルベルト空間を $\mathcal{H}_k$ とし，Borel 集合族によって可測空間と考える．いま，X を $\mathcal{X}$ に値をとる確率変数，すなわち，確率空間 $(\mathcal{M},\mathcal{B},P)$ があって $X:\mathcal{M}\to\mathcal{X}$ は可測写像とする．このときカーネル法の特徴写像 $\Phi:\mathcal{X}\to\mathcal{H}_k$，$x\mapsto k(\cdot,x)$ が可測となることは容易に確認できる．したがって $\Phi(X)$ は再生核ヒルベルト空間 $\mathcal{H}_k$ に値を取る確率変数である．

以下では，確率変数と再生核ヒルベルト空間に対し

$$E[\sqrt{k(X,X)}]<\infty$$

を仮定する．特徴写像 $\Phi(x)=k(\cdot,x)$ に対し $\|\Phi(X)\|^2=k(X,X)$ に注意すると，上の仮定は $E\|\Phi(X)\|<\infty$ を意味する．このとき $\Phi(X)$ の平均 $m_X^k\in\mathcal{H}_k$ が存在して

$$\langle f,m_X^k\rangle=E[\langle f,\Phi(X)\rangle]=E[f(X)]\qquad(\forall f\in\mathcal{H}_k) \tag{3.1}$$

が成り立つ．そこで m_X^k を X の $\mathcal{H}_k$ における平均と呼ぶ．上式から，任意の $f\in\mathcal{H}_k$ に対して期待値 $E[f(X)]$ が f と平均 m_X^k との内積で計算されるので，これは再生性の期待値版と考えられる．

平均 m_X^k の関数としての陽な表示を求めよう．m_X^k は $\mathcal{H}_k$ の元なので，再生性により，任意の $y\in\mathcal{X}$ に対して

$$m_X^k(y)=\langle m_X^k,k(\cdot,y)\rangle=E[k(X,y)] \tag{3.2}$$

である．すなわち，平均 m_X^k はカーネル関数の期待値として与えられる．

$\mathbb{R}$ 上の d 次の多項式カーネル $k(x,y)=(xy+c)^d$ $(c>0)$ が定める再生核ヒルベルト空間 $\mathcal{H}_k$ は，ベクトル空間として d 次以下の多項式全体と一致するので，$\mathbb{R}$ 上の確率変数 X に対し，その r 次モーメント $\mu_r=E[X^r]$ $(0\le r\le d)$ が

$$\mu_r=\langle x^r,m_X^k\rangle_{\mathcal{H}_k}$$

により計算される．これからわかるように，平均 m_X^k は X の分布の高次モーメントの情報を持っている．

次に再生核ヒルベルト空間における平均の推定量を考える．再生核ヒルベルト空間は一般に無限次元の関数空間であるが，以下でみるように，その上で定義された統計量の推定量が容易に構成でき，その統計的性質も比較的容易に調べられる点に長所がある．

$X,X_1,\ldots,X_n$ を P に従う i.i.d. サンプルとするとき，m_X^k の推定量 $\widehat{m}_{(n)}^k$ を

$$\widehat{m}_{(n)}^k=\frac{1}{n}\sum_{i=1}^n k(\cdot,X_i)=\frac{1}{n}\sum_{i=1}^n\Phi(X_i) \tag{3.3}$$

により定義する．これが $m_X^k=E[k(\cdot,X)]=E[\Phi(X)]$ の不偏推定量であることはすぐにわかるが，さらに次のような漸近的性質が導かれる．

定理 1. 上の仮定のもと，

$$E\|\widehat{m}_{(n)}^k-m_X^k\|_{\mathcal{H}_k}^2=\frac{1}{n}\big\{E[k(X,X)]-E[k(X,\tilde{X})]\big\}$$

($\tilde{X}$ は X と独立で同一の分布 P に従う確率変数) が成り立つ．特に

$$\|\widehat{m}_{(n)}^k-m_X^k\|_{\mathcal{H}_k}=O_p(n^{-1/2})\qquad(n\to\infty).$$

証明. $\langle k(\cdot,X_i),m_X^k\rangle_{\mathcal{H}_k}=E_X[k(X,X_i)]$ により

$$\|\widehat{m}_{(n)}^k-m_X^k\|_{\mathcal{H}_k}^2=\left\|\frac{1}{n}\sum_{i=1}^n k(\cdot,X_i)-m_X^k\right\|_{\mathcal{H}_k}^2$$

$$=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(X_j,X_i)-\frac{1}{n}\sum_{i=1}^n E_X[k(X,X_i)]-\frac{1}{n}\sum_{i=1}^n E_X[k(X_i,X)]+E[k(X,\tilde{X})]$$

が成り立つことから第 1 の主張が得られる．第 2 の主張は Chebychev の不等式から従う．□

$\|m^k_{(n)} - m^k_X\| = \sup_{\|f\|\le 1} \left|\langle f, m^k_{(n)} - m^k_X\rangle\right|$ に注意すると，定理 1 の系として，$\mathcal{H}_k$ の単位球に対する一様な大数の法則が得られる．

系 1. 定理 1 と同じ仮定のもと

$$\sup_{f\in\mathcal{H}_k, \|f\|_{\mathcal{H}_k}\le 1} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - E[f(X)] \right| = O_p(n^{-1/2}) \qquad (n\to\infty)$$

が成り立つ．

次に中心極限定理に関して考えよう．$E[k(X,X)] < \infty$ を仮定すると，任意の $f\in\mathcal{H}_k$ に対し $E[f(X)^2] = E|\langle f, k(\cdot,X)\rangle_{\mathcal{H}_k}|^2 \le \|f\|^2 E\|k(\cdot,X)\|^2_{\mathcal{H}_k} = \|f\|^2 E[k(X,X)] < \infty$ により，$f(X)$ は有限の分散 $V(f)$ を持つ．したがって中心極限定理

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - E[f(X)] \right) \Longrightarrow N(0, V(f)) \qquad (n\to\infty)$$

が成り立つ．これを内積によって書き換えると

$$\langle f, \sqrt{n}(m^k_{(n)} - m^k_X)\rangle_{\mathcal{H}_k} \Longrightarrow N(0, V(f)) \qquad (n\to\infty)$$

であるが，このことは $\mathcal{H}_k$ 上の確率変数 $\sqrt{n}(m^k_{(n)} - m^k_X)$ が何らかのガウス確率変数に収束する可能性を示唆している．実際，次の定理が成り立つ．

定理 2. $E[k(X,X)] < \infty$ を仮定する．$G_n = \sqrt{n}(m^k_{(n)} - m^k_X)$ は，$\mathcal{H}_k$ に値をとる確率変数として，$n\to\infty$ のとき $\mathcal{H}_k$ 上のガウス確率変数 G に法則収束する．ここで G は平均 0，共分散関数 $R(f,g) = \mathrm{Cov}[f(X), g(X)]$ により定まる．

証明は省略する．例えば Berlinet and Thomas-Agnan (2004) 第 4 章を見ていただきたい．

3.2 確率分布を特徴づける正定値カーネル

3.1 節で見たように，確率変数を再生核ヒルベルト空間に写像するとその平均は元の確率変数の高次モーメントの情報を含んでいる．直感的に言うと，確率変数に対してすべてのモーメントが表現できればその分布は決まるので，十分広いクラスの関数を含むような再生核ヒルベルト空間における平均を考えれば，確率変数を一意的に定めることが期待できる．本節ではこのような正定値カーネルのクラスを議論する．このクラスは，正定値カーネルを用いた統計的推論において重要な役割を果たす．

$(\mathcal{X}, \mathcal{B}_\mathcal{X})$ を可測空間，$\mathcal{P}$ をその上の確率測度全体とする．$\mathcal{X}$ 上の有界かつ可測な正定値カーネル k が**特性的**(characteristic) であるとは，写像

$$\mathcal{P}\to\mathcal{H}_k, \quad P\mapsto m^k_P$$

が単写であることをいう．ここで m^k_P は分布 P を持つ確率変数の $\mathcal{H}_k$ における平均を表す．正定値カーネルが特性的であるとき，それが定める再生核ヒルベルト空間は特性的であるという．上の定義は，

$$E_{X\sim P}[f(X)] = E_{X\sim Q}[f(X)] \quad (\forall f\in\mathcal{H}_k) \quad\Longrightarrow\quad P = Q$$

と同値であり，特性的な正定値カーネルは，再生核ヒルベルト空間における平均によって $\mathcal{P}$ の確率分布を一意に定める．

$\{k(\cdot,y) \mid y\in\mathcal{X}\}$ の線形結合が $\mathcal{H}_k$ で稠密であることより，条件 $m^k_P = m^k_Q$ は，任意の $y\in\mathcal{X}$ に対し $E_{X\sim P}[k(X,y)] = E_{X\sim Q}[k(X,y)]$ が成り立つことと同値である．したがって特性的な正

定値カーネルは

$$E_{X\sim P}[k(X,y)]=E_{X\sim Q}[k(X,y)] \quad (\forall y\in\mathcal{X}) \quad\Longleftrightarrow\quad P=Q \tag{3.4}$$

を成立させる正定値カーネルである.

後で示すように，ガウス RBF カーネル $k^G_\sigma(x,y)=\exp\{-\|x-y\|^2/(2\sigma^2)\}$ $(\sigma>0)$ やラプラスカーネル $k^L_\lambda(x,y)=\exp\bigl(-\lambda\sum_{i=1}^m|x_i-y_i|\bigr)$ $(\lambda>0)$ は $\mathbb{R}^m$ 上の特性的なカーネルである.

式(3.4)からわかるように，特性的な正定値カーネルは，$\mathbb{R}^m$ 上の確率分布 P に対する特性関数 $E_{X\sim P}[e^{\sqrt{-1}u^TX}]$ と類似性を持つ．よく知られているように，特性関数は確率分布 P を一意に定める．特性的なカーネルは特性関数のこの性質を取り出して定義されている．ただし $e^{\sqrt{-1}x^Ty}$ は $\mathbb{R}^n$ 上の正定値カーネルではない.

次の事実は，特性的な再生核ヒルベルト空間が L^2 の意味で十分広い空間であることを示している.

補題 1. 正定値カーネル k が特性的であるための必要十分条件は，任意の確率分布 $P\in\mathcal{P}$ に対し $\mathcal{H}_k+\mathbb{R}$ が $L^2(P)$ で稠密なことである．ここで，$\mathcal{H}_k+\mathbb{R}$ は再生核ヒルベルト空間としての直和を意味する.

証明. まず十分性を示す．$P,Q\in\mathcal{P}$ に対し，$P\neq Q$ かつ $m^k_P=m^k_Q$ として矛盾を導く．$P-Q$ の全変動を $|P-Q|$ で表すとき，仮定から $\mathcal{H}_k+\mathbb{R}$ は $L^2(|P-Q|)$ で稠密なので，$\mathcal{X}$ の任意の可測集合 A と任意の $\varepsilon>0$ に対し，$\varphi\in\mathcal{H}_k+\mathbb{R}$ があって

$$\int|\varphi(x)-I_A(x)|d(|P-Q|)(x)<\varepsilon$$

が成り立つ．ここで I_A は A の定義関数である．このとき

$$\bigl|(E_{X\sim P}[\varphi(X)]-P(A))-(E_{X\sim Q}[\varphi(X)]-Q(A))\bigr|<\varepsilon$$

である．$m^k_P=m^k_Q$ により $E_{X\sim P}[\varphi(X)]=E_{X\sim Q}[\varphi(X)]$ なので，$|P(A)-Q(A)|<\varepsilon$ であるが，$\varepsilon>0$ は任意なので $P(A)=Q(A)$ となり $P\neq Q$ に反する.

次に必要性を示す．ある $P\in\mathcal{P}$ があって $\mathcal{H}_k+\mathbb{R}$ が $L^2(P)$ で稠密でないと仮定する．このとき，0 でない $f\in L^2(P)$ を $\mathcal{H}_k+\mathbb{R}$ の直交補空間からとると，

$$\int f\varphi dP=0 \quad(\forall\varphi\in\mathcal{H}_k),\qquad \int fdP=0$$

が成立する．$c=1/\|f\|_{L^1(P)}$ とおき，2 つの確率 Q_1，Q_2 を

$$Q_1(E)\equiv c\int_E|f|dP,\qquad Q_2(E)\equiv c\int_E(|f|-f)dP$$

により定義する．$f\neq0$ により $Q_1\neq Q_2$ であるが，一方任意の $\varphi\in\mathcal{H}_k$ に対し

$$E_{X\sim Q_1}[\varphi(X)]-E_{X\sim Q_2}[\varphi(X)]=c\int f\varphi dP=0$$

により $m^k_{Q_1}=m^k_{Q_2}$ である．したがって k は特性的でない．□

$\mathbb{R}^n$ 上の連続で平行移動不変な正定値カーネル($k(x,y)=\phi(x-y)$ と書けるもの)に関しては，ある非負測度 Λ の逆フーリエ変換，すなわち

$$\phi(x-y)=\int e^{\sqrt{-1}(x-y)^T\omega}d\Lambda(\omega)$$

の形に表される(Bochner の定理)．このクラスの正定値カーネルに対しては特性的であるための条件を簡潔に述べることが可能である．この際に重要なのは，平行移動不変な正定値カーネルに対して，確率測度 P の $\mathcal{H}_k$ における平均 m_P が

$$m_P(x) = \int k(x,y)dP(y) = \int \phi(x-y)dP(y) = (\phi * P)(x)$$

と ϕ と P の畳み込みとして表現できる点にある．したがって，特性的であることは，

$$\phi * P = \phi * Q \quad \Longrightarrow \quad P = Q$$

と同値である．ここで，畳み込みの Fourier 変換が Fourier 変換の積で与えられることを用いると，厳密性に多少目を瞑れば，上の条件はさらに

$$\widehat{\phi}\widehat{P} = \widehat{\phi}\widehat{Q} \quad \Longrightarrow \quad P = Q$$

と書き直せる．この条件は $\widehat{\phi} = \Lambda$ が全空間で正であれば成立することが予想されるが，実際以下に見るように，上の議論は厳密化することが可能である．

定理 3.（Sriperumbudur et al., 2010）ϕ を $\mathbb{R}^n$ 上の連続な実正定値関数とし，Λ を Bochner の定理の表示

$$\phi(x) = \int e^{\sqrt{-1}\omega^T x} d\Lambda(\omega)$$

を与える有限非負測度とする．このとき，$\phi(x-y)$ が特性的な正定値カーネルであるための必要十分条件は $\mathrm{Supp}(\Lambda) = \mathbb{R}^n$ である．

ここで $\mathrm{Supp}(\Lambda)$ は測度 Λ の台であり，

$$\mathrm{Supp}(\Lambda) = \{x \in \mathbb{R}^n \mid x \text{ を含む任意の開集合 } U \text{ に対して } \Lambda(U) > 0\}$$

により定義される．定理 3 の証明は省略する．原論文または福水(2010)を参照のこと．

定理 3 を用いると，さまざまな平行移動不変な正定値カーネルが特性的であることがわかる．$\phi^G_\sigma(x,y) = \exp\{-\|x\|^2/(2\sigma^2)\}$ $(\sigma > 0)$ と $\phi^L_\lambda(x) = \exp\left(-\lambda \sum_{i=1}^m |x_i|\right)$ $(\lambda > 0)$ の Fourier 変換は，それぞれ正の定数倍を除いて $\exp\{-\sigma^2\|\omega\|^2/2\}$ および $\prod_{i=1}^m 1/(\lambda + \omega_i^2)$ となり，$\mathbb{R}^m$ 上の特性的なカーネルである．一方，sinc 関数 $\mathrm{sinc}(x) = \sin(x)/x$ の Fourier 変換は(正の定数倍を除いて)区間の定義関数 $I_{[-1,1]}(\omega)$ であるため，正定値関数であるが特性的ではない．これらの例からわかるように，特性的なカーネルは Fourier 変換がすべての周波数で正であり，すべての周波数成分を扱うことができる．一方，特性的でないカーネルは，ある周波数領域を表すことができないため，その周波数成分のみ異なる密度関数をもつ確率を区別できない．

3.3 2 標本問題への応用

特性的な正定値カーネル k を用いると，平均 m^k_X の推定量を用いて 2 標本の均一性検定が行える(Gretton et al., 2007, 2010)．

2 標本の均一性検定とは，2 つのサンプル $(X_1,\ldots,X_\ell)$ と $(Y_1,\ldots,Y_n)$ を発生させた分布が同じかどうかを判定する問題である．以下では $X_1,\ldots,X_\ell$ と $Y_1,\ldots,Y_n$ は可測空間 $(\mathcal{X},\mathcal{B})$ に値をとり，それぞれ独立に確率分布 P および Q に従う i.i.d. サンプルと仮定する．$P = Q$ を帰無仮説，$P \neq Q$ を対立仮説として検定を行う．

k を $\mathcal{X}$ 上の($\mathcal{B}$ に対して)特性的な実正定値カーネルとし，$X \sim P$，$Y \sim Q$ なる独立な変数 X，Y に対して $E[k(X,Y)^2] < \infty$ を満たすとする．P および Q による平均を m^k_P，m^k_Q とするとき，P と Q の距離の 2 乗

$$M^2(P,Q) \equiv \|m^k_P - m^k_Q\|^2_{\mathcal{H}_k}$$

が 0 か否かによって，$P=Q$ であるかどうかを判定することができる．m_P^k および m_Q^k の推定量は，式(3.3)と同様

$$\widehat{m}_P=\frac{1}{\ell}\sum_{i=1}^{\ell}k(\cdot,X_i),\qquad \widehat{m}_Q=\frac{1}{n}\sum_{i=1}^{n}k(\cdot,Y_i) \tag{3.5}$$

で与えられるので，検定統計量として

$$\hat{M}_{\ell,n}=\|\widehat{m}_P-\widehat{m}_Q\|_{\mathcal{H}_k}^2=\frac{1}{\ell^2}\sum_{a,b=1}^{\ell}k(X_a,X_b)+\frac{1}{n^2}\sum_{c,d=1}^{n}k(Y_c,Y_d)-\frac{2}{\ell n}\sum_{a=1}^{\ell}\sum_{c=1}^{n}k(X_a,Y_c)$$

を用いることが可能である．また，これを不偏化して

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

を用いてもよい．$U_{\ell,n}$ は

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

というカーネルによる 2 標本 U-統計量になることが確認できる．

仮説検定を行うためには帰無仮説 $P=Q$ のもとで検定統計量 $U_{\ell,n}$ の分布を知る必要がある．この場合，上の $U_{\ell,n}$ は退化した 2 標本 U 検定統計量であり，その漸近分布は知られている．いま，総データ数を $N=\ell+n$ とおき，

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

を仮定する．N を無限大としたときの漸近分布は以下のよう与えられる(詳しくは福水, 2010 参照)．

定理 4. $P=Q$ の帰無仮説のもと，

$$NU_{\ell,n}\quad\Rightarrow\quad\sum_{i=1}^{\infty}\lambda_i\left(Z_i^2-\frac{1}{\gamma(1-\gamma)}\right)\qquad (n\to\infty) \tag{3.6}$$

と法則収束する．ここで，Z_i は平均 0 分散 $1/\gamma(1-\gamma)$ の正規分布 $N(0,\frac{1}{\gamma(1-\gamma)})$ に従う独立な確率変数であり，$\{\lambda_i\}_{i=1}^{\infty}$ は

$$\tilde{k}(x,y)=k(x,y)-E[k(x,X)]-E[k(X,y)]+E[k(X,\tilde{X})] \tag{3.7}$$

($\tilde{X},X$ は独立に P に従う確率変数)で定まる $L^2(P)$ 上の積分作用素の非零固有値を重複度だけ並べたもの，すなわち，ある単位ベクトル $\phi_i\in L^2(P)$ に対して

$$\int\tilde{k}(x,y)\phi_i(y)dP(y)=\lambda_i\phi_i(x) \tag{3.8}$$

を満たす非負実数 λ_i を重複度だけ考えたものとなる．

一方 k が特性的な場合，対立仮説 $P\neq Q$ のもとでは $M^2(P,Q)\neq 0$ であり，非退化な U 統計量の一般的事実から，$\sqrt{N}(U_{\ell,n}-M^2(P,Q))$ は正の分散を持つ正規分布に法則収束する．したがって $NU_{\ell,n}$ による検定は一致性を持つ．

表 1. 正定値カーネルによる方法と Kolmogorov-Smirnov 検定による均一性検定の結果．有意水準を $\alpha=5\%$，データ数を $N=200,500,1000$ とし，500 回の実験のうち帰無仮説が受容された割合を示した．

	$\hat{M}(P,Q)$					Kolmogorov-Smirnov				
$N\backslash a$	1	0.75	0.5	0.25	0	1	0.75	0.5	0.25	0
200	0.966	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

以上により，漸近的な帰無分布を検定に用いる際には，λ_i $(i=1,2,\ldots)$ が決定できれば棄却域を決定することができる．式(3.7)の積分核は中心化された正定値カーネルに一致していることから，実は，固有値 λ_i の一致推定量が中心化グラム行列

$$\tilde{K}_{ij}=k(X_i,X_j)-\frac{1}{n}\sum_{b=1}^{n}k(X_i,X_b)-\frac{1}{n}\sum_{a=1}^{n}k(X_a,X_j)+\frac{1}{n^2}\sum_{a,b=1}^{n}k(X_a,X_b)$$

の固有値によって与えられることがわかる(Gretton et al., 2010)．そこで，$\tilde{K}$ の固有値 $\hat{\lambda}_1,\ldots,\hat{\lambda}_{n-1}$ 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って，式(3.6)の極限分布の α-%点の近似値を計算機シミュレーションにより求めることができる．

数値実験として，P を正規分布 $N(0,1/3)$，Q_a を 区間 $[-1,1]$ 上の一様分布と $N(0,1/3)$ との混合分布

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

とし，a を変化させて，$\hat{M}^2(P,Q)$ による検定を行った結果を表 1 に示す．P と Q_a は平均と分散が常に一致するため，2 次モーメントまでの情報ではこれらを識別できない．正定値カーネルはガウス RBF カーネルを用い，分散に相当するパラメータ σ には，データ間の距離の中央値を用いた．棄却域は上で述べた方法によって求めた．また比較のために，分布の均一性に対する Kolmogorov-Smirnov 検定を同じサンプルに行った結果も合わせて示している．この例では，カーネル法による 2 標本検定は，ノンパラメトリック検定の標準的方法である Kolmogorov-Smirnov 検定に遜色ない検出力を持っていることがわかる．

4. 正定値カーネルによる依存性・独立性

本章では，確率変数の独立性を正定値カーネルによって扱う方法について述べる．確率変数を再生核ヒルベルト空間に写像しその分散を考えることによって高次の統計量を扱うのが基本的なアイデアである．本章では，ヒルベルト空間の間の作用素 $T:\mathcal{H}_1\to\mathcal{H}_2$ に対し，値域と零空間をそれぞれ $\mathcal{R}(T)=\{Tf\in\mathcal{H}_2\,|\,f\in\mathcal{H}_1\}$ と $\mathcal{N}(T)=\{f\in\mathcal{H}_1\,|\,Tf=0\}$ で表す．

4.1 再生核ヒルベルト空間上の共分散作用素

まず，$\mathbb{R}^n$ に値を取る確率ベクトルに対する通常の共分散行列の一般化として，再生核ヒルベルト空間における共分散作用素を定義する．$(\mathcal{X},\mathcal{B}_\mathcal{X})$，$(\mathcal{Y},\mathcal{B}_\mathcal{Y})$ を可測空間，(X,Y) を $\mathcal{X}\times\mathcal{Y}$ に値をとる確率変数とする．(X,Y) の分布を P_{XY}，X,Y の周辺分布を P_X,P_Y とおく．また，$(\mathcal{H}_\mathcal{X},k_\mathcal{X})$，$(\mathcal{H}_\mathcal{Y},k_\mathcal{Y})$ をそれぞれ $\mathcal{X}$，$\mathcal{Y}$ 上の可測な正定値カーネルと対応する再生核ヒルベルト空間とする．確率変数と正定値カーネルは，常に仮定

(2 乗可積分条件)　　$E[k_\mathcal{X}(X,X)]<\infty,\quad E[k_\mathcal{Y}(Y,Y)]<\infty$

を満たすとする．この仮定は正定値カーネルが有界であれば常に満たされる．

2 乗可積分条件を用いると，$|E[f(X)g(Y)]| \le E|\langle f, k_{\mathcal{X}}(\cdot,X)\rangle\langle g, k_{\mathcal{Y}}(\cdot,Y)\rangle| \le E[\|k_{\mathcal{X}}(\cdot,X)\|\|k_{\mathcal{Y}}(\cdot,Y)\|]\|f\|\|g\| \le \big(E[k_{\mathcal{X}}(X,X)]E[k_{\mathcal{Y}}(Y,Y)]\big)^{1/2}\|f\|\|g\|$ を得るので，双線形写像 $(f,g)\mapsto \mathrm{cov}[f(X), g(Y)]$ は連続写像である．したがって Riesz の定理を用いると，有界線形作用素 $\Sigma_{YX}^{k_{\mathcal{Y}},k_{\mathcal{X}}} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ があって

$$\langle g, \Sigma_{YX}^{k_{\mathcal{Y}},k_{\mathcal{X}}} f\rangle = \mathrm{cov}[f(X), g(Y)] = E[f(X)g(Y)] - E[f(X)]E[g(Y)] \tag{4.1}$$

が成り立つ．この線形作用素 $\Sigma_{YX}^{k_{\mathcal{Y}},k_{\mathcal{X}}}$ を**相互共分散作用素**と呼ぶ（相互共分散作用素の一般的な理論は Baker, 1973 に詳しい）．特徴写像 $\Phi_{\mathcal{X}} : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}, \Phi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ を用いると，$\Sigma_{YX}^{k_{\mathcal{Y}},k_{\mathcal{X}}}$ は

$$E[\Phi_{\mathcal{Y}}(Y) \otimes \Phi_{\mathcal{X}}(X)] - E[\Phi_{\mathcal{Y}}(Y)]E[\Phi_{\mathcal{X}}(X)]$$

とみなすこともできる．以下では特に紛れが無い限り，簡単のため $k_{\mathcal{Y}}, k_{\mathcal{X}}$ を省略して Σ_{YX} と表すことにする．特に $Y = X$ の場合，Σ_{XX} は正値な自己共役作用素であり**共分散作用素**と呼ぶ．

$\mathbb{R}^m, \mathbb{R}^n$ に値をとる通常の確率ベクトル X, Y に対してユークリッド内積による正定値カーネルを用いると，上の相互共分散作用素は，通常の共分散行列の定める線形写像に一致することが容易に確認されるので，通常の共分散行列の拡張となっている．

4.2 共分散作用素による独立性の特徴づけ

2 つの 1 次元確率変数 X, Y の関係を見るために共分散や相関を用いることは基本的であるが，これでは線形な関係しか考慮できない．一方，再生核ヒルベルト空間への特徴写像 $\Phi_{\mathcal{X}}(X), \Phi_{\mathcal{Y}}(Y)$ は高次の情報を含むため，これらの共分散を考えれば，もとの確率変数の高次の依存性を調べられる．以下ではこの考えに基づいて，特性的な正定値カーネルによる相互共分散作用素を用いて確率変数の独立性や依存性を扱う方法を述べる．以下では X と Y が独立であることを $X \perp\!\!\!\perp Y$ で表す．

定理 5.（Fukumizu et al., 2004）$(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ を可測空間，(X,Y) を $\mathcal{X} \times \mathcal{Y}$ 上の確率変数とし，$(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$, $(\mathcal{H}_{\mathcal{Y}}, k_{\mathcal{Y}})$ をそれぞれ可測な正定値カーネルにより定まる $\mathcal{X}$, $\mathcal{Y}$ 上の再生核ヒルベルト空間とし，2 乗可積分条件を仮定する．このとき，積 $k_{\mathcal{X}}k_{\mathcal{Y}}$ が $\mathcal{X} \times \mathcal{Y}$ 上特性的であるならば，

$$X \perp\!\!\!\perp Y \iff \Sigma_{XY} = O \quad (\text{零作用素}) \tag{4.2}$$

の同値関係が成り立つ．

証明. (X,Y) の同時分布を P_{XY}，また X，Y と同じ周辺分布を持ち，互いに独立な確率分布を $P_X \otimes P_Y$ と書く．$\mathcal{H}_{\mathcal{X}}$ から $\mathcal{H}_{\mathcal{Y}}$ への線形作用素をテンソル積 $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ の元とみなすと，Σ_{YX} は $m_{P_{XY}} - m_{P_X} \otimes m_{P_Y}$ で表されるので，$\Sigma_{YX} = O$ は $m_{P_{XY}} = m_{P_X \otimes P_Y}$ と同値である．したがって定理の同値性は $k_{\mathcal{X}}k_{\mathcal{Y}}$ が特性的なことから従う．□

上の定理の特徴づけを言い換えると，

$$X \perp\!\!\!\perp Y \iff E[k_{\mathcal{X}}(x,X)k_{\mathcal{Y}}(y,Y)] = E[k_{\mathcal{X}}(x,X)]E[k_{\mathcal{Y}}(y,Y)] \quad (\forall x \in \mathcal{X}, y \in \mathcal{Y})$$

である．これは累積分布関数や特性関数を用いた通常の確率ベクトルの独立性の特徴づけと類似性を持つ．累積分布関数や特性関数による特徴づけでは，それぞれ $I_{(-\infty,u]}(x)$（区間 $(-\infty,u]$ の定義関数）と $e^{\sqrt{-1}x^T u}$（Fourier カーネル）を一種のテスト関数として P_{XY} と $P_X \otimes P_Y$ を比較しているのに対し，定理 5 では $k_{\mathcal{X}}(x,X)k_{\mathcal{Y}}(y,Y)$ という正定値カーネル関数をテスト関数と

して比較を行っている．これらのアプローチを比べると，累積分布関数や特性関数が通常の確率ベクトルに対してのみ定義できるのに対し，カーネル法による特徴づけは，任意の可測空間に値をとる確率変数に適用可能である．また，以下に見るように，再生性により推定量の構成や解析が容易である．

有限サンプル $(X_1,Y_1),\dots,(X_n,Y_n)$ が与えられた場合の相互共分散作用素の推定量として，**経験相互共分散作用素**を

$$\widehat{\Sigma}_{YX}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\big(k_{\mathcal{Y}}(\cdot,Y_i)-\widehat{m}_Y\big)\big\langle\big(k_{\mathcal{X}}(\cdot,X_i)-\widehat{m}_X\big),\cdot\big\rangle_{\mathcal{H}_{\mathcal{X}}} \tag{4.3}$$

により定義する．ここで

$$\widehat{m}_X = \frac{1}{n}\sum_{i=1}^{n}k_{\mathcal{X}}(\cdot,X_i),\qquad \widehat{m}_Y = \frac{1}{n}\sum_{i=1}^{n}k_{\mathcal{Y}}(\cdot,Y_i)$$

である．簡単のため $\tilde{k}^i_{\mathcal{X}} = k_{\mathcal{X}}(\cdot,X_i)-\widehat{m}_X$，$\tilde{k}^i_{\mathcal{Y}} = k_{\mathcal{Y}}(\cdot,Y_i)-\widehat{m}_Y$ と表すことにすると，

$$\widehat{\Sigma}_{YX}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{k}^i_{\mathcal{Y}}\langle\tilde{k}^i_{\mathcal{X}},\cdot\rangle_{\mathcal{H}_{\mathcal{X}}}$$

とかける．$\sum_{i=1}^n\tilde{k}^i_{\mathcal{X}}=0$ であるので，$\{\tilde{k}^i_{\mathcal{X}}\}_{i=1}^n,\{\tilde{k}^i_{\mathcal{Y}}\}_{i=1}^n$ の張る部分空間をそれぞれ V_X，W_Y とすると，これらは高々 $n-1$ 次元の部分空間である．$\widehat{\Sigma}_{YX}^{(n)}$ は $\mathcal{H}_{\mathcal{X}}$ から $\mathcal{H}_{\mathcal{Y}}$ への作用素であるが，上の表現から明らかなように，$V_X^{\perp}\subset\mathcal{N}(\widehat{\Sigma}_{YX}^{(n)})$ かつ $\mathcal{R}(\widehat{\Sigma}_{YX}^{(n)})\subset W_Y$ であり，特にそのランクは高々 $n-1$ である．$\widehat{\Sigma}_{YX}^{(n)}$ を V_X および W_Y に制約した有限次元作用素をグラム行列を使って表現してみよう．簡単な計算により

$$\widehat{\Sigma}_{YX}^{(n)}\tilde{k}^i_{\mathcal{X}} = \frac{1}{n}\sum_{j=1}^{n}\tilde{k}^j_{\mathcal{Y}}\langle\tilde{k}^j_{\mathcal{X}},\tilde{k}^i_{\mathcal{X}}\rangle_{\mathcal{H}_{\mathcal{X}}} = \frac{1}{n}\sum_{j=1}^{n}\tilde{K}^X_{ji}\tilde{k}^j_{\mathcal{Y}}$$

を得る．ここで $\tilde{K}^X$ は中心化グラム行列

$$\tilde{K}^X_{ij} = k(X_i,X_j)-\frac{1}{n}\sum_{b=1}^{n}k(X_i,X_b)-\frac{1}{n}\sum_{a=1}^{n}k(X_a,X_j)+\frac{1}{n^2}\sum_{a,b=1}^{n}k(X_a,X_b)$$

である．$\{\tilde{k}^i_{\mathcal{X}}\}_{i=1}^n,\{\tilde{k}^i_{\mathcal{Y}}\}_{i=1}^n$ は線形独立ではないため正確には基底とはいえないが，この冗長基底を用いると $\widehat{\Sigma}_{YX}^{(n)}$ の行列表示が $\frac{1}{n}\tilde{K}^X$ により得られることがわかる．

4.3 正定値カーネルによる依存性の尺度

定理 5 に基づいて，相互共分散作用素のノルムを X と Y の独立性・依存性の尺度として用いることが可能である．このとき Σ_{YX} のヒルベルト=シュミットノルムを用いると推定量の構成が容易である．ヒルベルト空間の間の作用素 $T:\mathcal{H}_1\to\mathcal{H}_2$ が**ヒルベルト=シュミット**であるとは，$\mathcal{H}_1$ と $\mathcal{H}_2$ の正規直交基底 $\{\phi_i\}$，$\{\psi_j\}$ に対して $\sum_{i,j}(T\phi_i,\psi_j)^2<\infty$ であることをいう．ヒルベルト=シュミット作用素 T のヒルベルト=シュミットノルムは $\|T\|_{HS}=\big(\sum_{i,j}(T\phi_i,\psi_j)^2\big)^{1/2}$ により定義される．有限次元の場合は行列のフロベニウスノルムと呼ばれることが多い．

定理 6. 2 乗可積分性条件のもと Σ_{YX} はヒルベルト=シュミット作用素であり，

$$\begin{aligned}\|\Sigma_{YX}\|_{HS}^2 = {}& E[k_{\mathcal{X}}(X,\tilde{X})k_{\mathcal{Y}}(Y,\tilde{Y})]-2E\big[E[k_{\mathcal{X}}(X,\tilde{X})|X]E[k_{\mathcal{Y}}(Y,\tilde{Y})|Y]\big]\\ &+E[k_{\mathcal{X}}(X,\tilde{X})]E[k_{\mathcal{Y}}(Y,\tilde{Y})]\end{aligned}$$

(ただし，$(\tilde{X},\tilde{Y})$ と (X,Y) は独立で同一の分布に従う)が成り立つ．また，Σ_{XX} のトレースは有限で

$$\mathrm{Tr}[\Sigma_{XX}] = E[k_{\mathcal{X}}(X,X)] - E[k_{\mathcal{X}}(X,\tilde{X})]$$

が成り立つ．

上式の第 1 式で，第 1，2，3 項はそれぞれ $P_{XY}\otimes P_{\tilde{X}\tilde{Y}}$，$P_X\otimes P_Y\otimes P_{\tilde{X}\tilde{Y}}$，$P_X\otimes P_Y\otimes P_{\tilde{X}}\otimes P_{\tilde{Y}}$ による期待値である．

略証．Σ_{YX} が $\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{Y}}$ の元 $m_{P_{XY}} - m_{P_X}\otimes m_{P_Y}$ とみなせることを用いると，ヒルベルト=シュミットノルムの定義により $\|\Sigma_{YX}\|_{HS}^2 = \|m_{P_{XY}} - m_{P_X}\otimes m_{P_Y}\|_{\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{Y}}}^2$ である．この平均の 2 乗ノルムを展開することにより前半を得る．

後半については，$\{\varphi_i\}_{i=1}^J$ $(J\in\mathbb{N}\cup\{\infty\})$ を $\mathcal{H}_{\mathcal{X}}$ の完全正規直交系とするとき，定義により $\mathrm{Tr}[\Sigma_{XX}] = \sum_{i=1}^J E\big[\langle\varphi_i, k_{\mathcal{X}}(\cdot,X)\rangle_{\mathcal{H}_{\mathcal{X}}}^2\big] - \sum_{i=1}^J \langle\varphi_i, m_X\rangle_{\mathcal{H}_{\mathcal{X}}}^2$ であるが，第 1 項は $E\big[\sum_{i=1}^J \langle\varphi_i, k_{\mathcal{X}}(\cdot,X)\rangle_{\mathcal{H}_{\mathcal{X}}}^2\big] = E\|k(\cdot,X)\|_{\mathcal{H}_{\mathcal{X}}}^2$ に等しく，第 2 項が $\|m_X\|_{\mathcal{H}_{\mathcal{X}}}^2$ に一致することから得られる．□

有限サンプル $(X_1,Y_1),\ldots,(X_n,Y_n)$ が与えられたとき，$\widehat{\Sigma}_{YX}^{(n)}$ のヒルベルト=シュミットノルムは以下のように与えられる．

$$\begin{aligned}\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2 &= \frac{1}{n^2}\sum_{i,j=1}^n k_{\mathcal{X}}(X_i,X_j)k_{\mathcal{Y}}(Y_i,Y_j) - \frac{2}{n^3}\sum_{i=1}^n\sum_{j=1}^n k_{\mathcal{X}}(X_i,X_j)\sum_{\ell=1}^n k_{\mathcal{Y}}(Y_i,Y_\ell)\\ &\quad + \frac{1}{n^4}\sum_{i,j=1}^n k_{\mathcal{X}}(X_i,X_j)\sum_{\ell,r=1}^n k_{\mathcal{Y}}(Y_\ell,Y_r)\\ &= \frac{1}{n^2}\mathrm{Tr}\big[\tilde{K}_X\tilde{K}_Y\big] = \frac{1}{n^2}\mathrm{Tr}\big[K_XQ_nK_YQ_n\big]\end{aligned}$$

ここで $\tilde{K}_X,\tilde{K}_Y$ は中心化グラム行列，$Q_n = I_n - \frac{1}{n}J_n$ (J_n はすべての成分が 1 である $n\times n$ 行列)であり，$(1,\ldots,1)^T$ の直交補空間への正射影行列である．

$\|\Sigma_{YX}\|_{HS}^2 = \|m_{P_{XY}} - m_{P_X}\otimes m_{P_Y}\|_{\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{Y}}}^2$ により，経験平均と平均に関して 3.1 節で示した漸近的議論が，$\widehat{\Sigma}_{YX}^{(n)}$ と Σ_{YX} に対してそのまま成立する．例えば，

$$\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\|_{HS} = O_p(n^{-1/2}) \qquad (n\to\infty) \tag{4.4}$$

が得られる．特に，経験的な独立性尺度 $\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$ は $\|\Sigma_{YX}\|_{HS}^2$ に $n^{-1/2}$ のオーダーで確率収束する．さらに 3.3 節と同様にして，特性的な正定値カーネルを用いて 2 つの確率変数の独立性を検定することが可能である．

以下では独立性検定の統計量として

$$T_n = n\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$$

を用いた数値例を示す．X^1 と Y^1 のデータを図 1 で示したように作成する．ここで左図は X^1，Y^1 が独立になるように作成され，これを回転すると X^1 と Y^1 は独立でない．しかし相関は常に 0 である．さらに $X^2,\ldots,X^m$，$Y^2,\ldots,Y^m$ を X^1，Y^1 に独立なガウス性ノイズとして次元を m 次元に拡張したデータを作成した．このデータに対して，上の T_n を用いた独立性の検定を行った．棄却域は前章の 2 標本検定と同様にして決定することも可能であるが，以下の実験では並べ替え検定を行った．比較としてパワーダイバージェンス(Ku and Fine, 2005)による検定も行った．X と Y が値をとる領域をそれぞれ L 個のセル J_X,J_Y に分割するとき，パワーダイバージェンスは，

$$n\frac{2}{\lambda(\lambda+2)}\sum_{(ij)\in J_X\times J_Y}\hat{p}_{ij}\Big\{\Big(\frac{\hat{p}_{ij}}{\hat{p}_i^X\hat{p}_j^Y}\Big)^\lambda - 1\Big\}$$

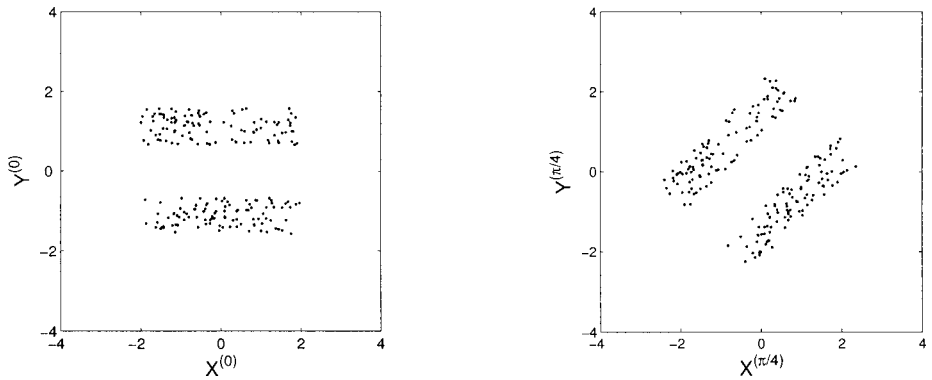

図 1. 独立性・依存性の計算機実験に用いたデータの例．左は独立，右は $\pi/4$ だけ回転したもので，無相関であるが独立ではない．

表 2. 正定値カーネルによる方法とパワーダイバージェンスによる独立性検定の結果．X と Y ともに同じ次元を用いた．有意水準を $\alpha=5\%$，データ数を 100 とし，100 回の実験のうち帰無仮説が受容された割合を示した．並べ替え検定は 500 個のランダムな置換によった．

$n\|\widehat{\Sigma}_{XY}^{(n)}\|_{HS}^2$				
次元 \ 角度	0	$\pi/12$	$\pi/6$	$\pi/4$
2	92	83	33	1
3	92	76	36	1
4	92	82	29	2

Power Divergence ($q=3$)				
次元 \ 角度	0	$\pi/12$	$\pi/6$	$\pi/4$
2	95	84	50	22
3	93	92	90	84
4	98	95	95	93

Power Divergence ($q=4$)				
次元 \ 角度	0	$\pi/12$	$\pi/6$	$\pi/4$
2	94	95	74	53
3	94	94	90	88
4	96	97	97	95

で定義される．ここで $\hat{p}_{ij}$ は (X_ℓ, Y_ℓ) $(\ell=1,\ldots,n)$ がセル $(ij)\in J_X\times J_Y$ に含まれる割合であり，$\hat{p}_i^X$ と $\hat{p}_j^Y$ は，各変数がセル $i\in J_X$ および $j\in J_Y$ に含まれる割合を表す．$\lambda=0$ のとき相互情報量に，$\lambda=2$ のとき χ^2 ダイバージェンスに一致する．この実験では $\lambda=2/3$ とし，各変数のセルは，各次元ごとに最小値から最大値までの区間を q 等分することにより作成した．帰無仮説のもとでパワーダイバージェンスはある自由度の χ^2 分布に法則収束することが知られているが，ここでは並べ替え検定によって棄却域を決定した．表 2 の結果をみると，カーネル法がより高い検出力を持っていることがわかる．特に，セル分割によるパワーダイバージェンスは次元の増加による検出力の劣化が著しいが，カーネル法は次元による影響が少ないことがわかる．

4.4 χ^2 ダイバージェンスの推定量

本節では，相互共分散作用素 Σ_{YX} を正規化した

$$V_{YX}=\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$$

によって，X と Y の依存性尺度を構成する．ここで $\Sigma_{XX}^{1/2}$ は固有展開によって構成される 1/2 乗である．一般にその逆 $\Sigma_{YY}^{-1/2}$ が存在するとは限らないが，実は Σ_{YX} に対して，作用素ノルムが 1 以下の有界作用素 V_{YX} が存在して，

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2} \tag{4.5}$$

かつ，$\mathcal{R}(V_{YX}) \subset \overline{\mathcal{R}(\Sigma_{YY})}$，$\mathcal{N}(V_{YX})^{\perp} \subset \overline{\mathcal{R}(\Sigma_{XX})}$ となるものが一意に存在することが知られている (Baker, 1973, Theorem 1 参照)．上の V_{YX} は正確にはこの意味である．

V_{YX} がヒルベルト=シュミット作用素であるとき，そのヒルベルト=シュミットノルムは興味深い表示を持つ．

定理 7. (Fukumizu et al., 2008) $k_{\mathcal{X}}$ と $k_{\mathcal{Y}}$ は特性的であるとし，(X,Y) の確率分布 P は $\mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$ に対して確率密度 $p_{XY}(x,y)$ を持つとする．上で定義した V_{YX} がヒルベルト=シュミット作用素であるとき，

$$\|V_{YX}\|_{HS}^2 = \int\int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1 \right)^2 p_X(x)p_Y(y) d\mu_{\mathcal{X}} d\mu_{\mathcal{Y}} \tag{4.6}$$

が成り立つ．

証明. Σ_{XX} は自己共役な正値コンパクト作用素なので，関数解析において知られているように，固有ベクトルからなる $\mathcal{H}_{\mathcal{X}}$ の正規直交基底 $\{\phi_i\}$ が存在し，対応する非負固有値 λ_i があって $\Sigma_{XX}\phi_i = \lambda_i \phi_i$ とできる．このとき $\mathcal{R}(\Sigma_{XX})^{\perp} = \mathcal{N}(\Sigma_{XX}) = \{f \in \mathcal{H}_{\mathcal{X}} \mid f = \text{const. a.e.-}P_X\}$ であるので，補題 1 により，$L^2(P_X)$ で $\{\phi_i\}$ の張る部分空間に対する直交補空間は高々定数関数のなす 1 次元部分空間である．したがって，正の固有値の添字集合を $I = \{i \mid \lambda_i > 0\}$ とおくと，$\{\phi_i\}_{i\in I} \cup \{1\}$ は $L^2(P_X)$ の基底をなす．特に，$i \in I$ のとき $E[\phi_i(X)] = 0$ に注意すると $E[\phi_i(X)^2] = \langle \phi_i, \Sigma_{XX}\phi_i \rangle_{\mathcal{H}_{\mathcal{X}}} = \lambda_i$ ゆえ，

$$\tilde{\phi}_i = \frac{1}{\sqrt{\lambda_i}} \phi_i \qquad (i \in I)$$

と定義するとき，$\{\tilde{\phi}_i\}_{i\in I} \cup \{1\}$ は $L^2(P_X)$ の正規直交基底である．同様に $\{\psi_j\}$ を Σ_{YY} の固有ベクトルからなる Σ_{YY} の正規直交基底，$\Sigma_{YY}\psi_j = \nu_j \psi_j$，$J = \{j \mid \nu_j > 0\}$ として，

$$\tilde{\psi}_j = \frac{1}{\sqrt{\nu_j}} \psi_j \qquad (j \in J)$$

と定めると，$\{\tilde{\psi}_j\}_{j\in J} \cup \{1\}$ は $L^2(P_Y)$ の正規直交基底となる．

$\mathcal{R}(V_{YX}) \subset \overline{\mathcal{R}(\Sigma_{YY})}$ と $\mathcal{N}(V_{YX})^{\perp} \subset \overline{\mathcal{R}(\Sigma_{XX})}$ に注意すると $\|V_{YX}\|_{HS}^2 = \sum_{i\in I, j\in J} \langle \psi_j, V_{YX}\phi_i \rangle^2$ となるので，

$$\begin{aligned} \|V_{YX}\|_{HS}^2 &= \sum_{i\in I,\ j\in J} \langle \tilde{\psi}_j,\ \Sigma_{YX}\tilde{\phi}_i \rangle^2 = \sum_{i\in I,\ j\in J} E[\tilde{\psi}_j(Y)\tilde{\phi}_i(X)]^2 \\ &= \sum_{i\in I,\ j\in J} \left(\tilde{\psi}_j\tilde{\phi}_i, \frac{p(x,\ y)}{p_X(x)p_Y(y)} \right)_{L^2(P_Y\otimes P_X)}^2 \end{aligned}$$

が成り立つ．ここで $P_Y \otimes P_X$ は周辺分布が P_Y，P_X で互いに独立な確率分布を意味する．$L^2(P_Y \otimes P_X)$ の正規直交基底は $\{\tilde{\psi}_j\tilde{\phi}_i\}_{i\in I, j\in J} \cup \{\tilde{\phi}_i\}_{i\in I} \cup \{\tilde{\psi}_j\}_{j\in J} \cup \{1\}$ からなるので，Parseval の等式から

$$\|V_{YX}\|_{HS}^2 = \left\| \frac{p(x,y)}{p_X(x)p_Y(y)} \right\|_{L^2(P_Y\otimes P_X)}^2 - \sum_{i\in I} \left(\tilde{\phi}_i, \frac{p(x,y)}{p_X(x)p_Y(y)} \right)_{L^2(P_Y\otimes P_X)}^2$$

$$-\sum_{j\in J}\Big(\tilde{\psi}_j,\frac{p(x,y)}{p_X(x)p_Y(y)}\Big)^2_{L^2(P_Y\otimes P_X)}-\Big(1,\frac{p(x,y)}{p_X(x)p_Y(y)}\Big)^2_{L^2(P_Y\otimes P_X)}$$
$$=\Big\|\frac{p(x,y)}{p_X(x)p_Y(y)}\Big\|^2_{L^2(P_Y\otimes P_X)}-1$$

を得る．末行は定理の主張の右辺に一致する．□

この定理は，正定値カーネルによって定義された量 $\|V_{YX}\|^2_{HS}$ が実は正定値カーネルに依存しないことを意味する．式(4.6)右辺の量は χ^2 ダイバージェンスとしてよく知られており，依存性尺度として用いられる．一方，左辺は

$$\big\|\big(\widehat{\Sigma}^{(n)}_{YY}+\varepsilon_n I_n\big)^{-1/2}\widehat{\Sigma}^{(n)}_{YX}\big(\widehat{\Sigma}^{(n)}_{XX}+\varepsilon_n I_n\big)^{-1/2}\big\|^2_{HS}$$

によって推定することが可能で，これをグラム行列表示することも容易である．ここで $\varepsilon_n>0$ は逆作用素を取るための正則化定数である．したがって，定理 7 は，χ^2 ダイバージェンスの正定値カーネルによる推定量を与えると考えることもできる．ε_n を適当なオーダーで 0 に近づけると，$n\to\infty$ のときにこの推定量は χ^2 ダイバージェンスの一致推定量となることも証明されている．

5. 正定値カーネル法による条件付独立性の特徴づけ

条件付独立性の概念は，グラフィカルモデリングをはじめとする統計的モデリングにおいて重要な概念である．ここでは，正定値カーネルを用いて条件付独立性を議論する方法について概略を述べる．詳細は Fukumizu et al. (2004, 2009)，福水(2010)を参照していただきたい．

5.1 条件付共分散作用素

$(\mathcal{X},\mathcal{B}_\mathcal{X}),(\mathcal{Y},\mathcal{B}_\mathcal{Y}),(\mathcal{Z},\mathcal{B}_\mathcal{Z})$ を可測空間，(X,Y,Z) を $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ に値を取る確率変数，$(\mathcal{H}_\mathcal{X},k_\mathcal{X})$，$(\mathcal{H}_\mathcal{Y},k_\mathcal{Y}),(\mathcal{H}_\mathcal{Z},k_\mathcal{Z})$ をそれぞれ $\mathcal{X},\mathcal{Y},\mathcal{Z}$ 上の可測な正定値カーネルと対応する再生核ヒルベルト空間とする．正定値カーネルと確率変数はすべて 2 乗可積分条件を満たすとする．また，確率変数 X，Y，Z の周辺分布をそれぞれ P_X，P_Y，P_Z で表す．このとき，**条件付相互共分散作用素** $\Sigma_{YX|Z}:\mathcal{H}_\mathcal{X}\to\mathcal{H}_\mathcal{Y}$ を，

$$\Sigma_{YX|Z}=\Sigma_{YX}-\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \tag{5.1}$$

により定義する．Σ_{ZZ}^{-1} は一般に存在するとは限らないので，より正確には，式(4.5)の分解を用いて

$$\Sigma_{YX|Z}=\Sigma_{YX}-\Sigma_{YY}^{1/2}V_{YZ}V_{ZX}\Sigma_{XX}^{1/2} \tag{5.2}$$

として定義される．特に $Y=X$ の場合，$\Sigma_{YY|Z}=\Sigma_{YY}-\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$ を単に**条件付共分散作用素**と呼ぶ．

2 乗可積分条件のもと，$\Sigma_{YX|Z}$ はヒルベルト＝シュミット作用素であり，$\Sigma_{YY|Z}$ は正値な自己共役作用素でそのトレースは有限である．このことは定義から容易に確認できる．

条件付相互共分散作用素は確率変数の条件付共分散と以下のように関係する．

定理 8. 2 乗可積分条件が成り立ち，$k_\mathcal{Z}$ が特性的であるとき，

$$\langle g,\Sigma_{YX|Z}f\rangle_{\mathcal{H}_\mathcal{Y}}=E[\mathrm{Cov}[f(X),g(Y)|Z]]\qquad(\forall f\in\mathcal{H}_\mathcal{X},g\in\mathcal{H}_\mathcal{Y})$$

が成立する．

証明は省略する．上の定理により $\Sigma_{YX|Z}$ が条件付共分散の期待値をあらわしていることがわかる．

次に，条件付共分散作用素と回帰の平均 2 乗誤差との関係について述べる．そのためにガウス確率変数における線形回帰の平均 2 乗誤差に関して復習しておこう．(X,Y) を有限次元ガウス確率ベクトルとし，X を m 次元説明変数として ℓ 次元応答変数 Y を線形回帰により推定することを考える．ここでは確率変数を $\tilde{X}=X-E[X]$，$\tilde{Y}=Y-E[Y]$ と中心化し，

$$\min_{A\in\mathbb{R}^{\ell\times m}} E\|\tilde{Y}-A\tilde{X}\|^2$$

を達成する $\widehat{A}$ によって

$$\hat{\tilde{Y}}=\widehat{A}\tilde{X}$$

という予測を行う．このときよく知られているように $\tilde{Y}$ の予測の残差の分散共分散行列は

$$E(\tilde{Y}-\hat{\tilde{Y}})(\tilde{Y}-\hat{\tilde{Y}})^T=C_{YY}-C_{YX}C_{XX}^{-1}C_{XY}$$

により与えられる．ここで C_{YX} などは通常の分散共分散行列であり，右辺は条件付共分散行列に他ならない．この事実の拡張が次の定理である．

定理 9. 2 乗可積分性を仮定し，$k_{\mathcal{X}}$ は特性的とする．このとき

$$\langle g,\Sigma_{YY|X}g\rangle_{\mathcal{H}_{\mathcal{Y}}}=\inf_{f\in\mathcal{H}_{\mathcal{X}}}E\big|\tilde{g}(Y)-\tilde{f}(X)\big|^2\qquad(\forall g\in\mathcal{H}_{\mathcal{Y}})$$

が成り立つ．ただし $\tilde{f}=f-E[f(X)]$，$\tilde{g}=g-E[g(Y)]$ である．

証明. $\Sigma_{YX}=\Sigma_{YY}^{1/2}V_{YX}\Sigma_{XX}^{1/2}$ を式(4.5)の分解とするとき，

$$\begin{aligned}E\big|\tilde{g}(Y)-\tilde{f}(X)\big|^2&=\mathrm{Var}[g(Y)]-2\mathrm{Cov}[g(Y),f(X)]+\mathrm{Var}[f(X)]\\&=\langle g,\Sigma_{YY}g\rangle_{\mathcal{H}_{\mathcal{Y}}}-2\langle\Sigma_{XY}g,f\rangle_{\mathcal{H}_{\mathcal{X}}}+\langle f,\Sigma_{XX}f\rangle_{\mathcal{H}_{\mathcal{X}}}\\&=\|\Sigma_{XX}^{1/2}f-V_{XY}\Sigma_{YY}^{1/2}g\|_{\mathcal{H}_{\mathcal{X}}}^2+\langle g,\Sigma_{YY}g\rangle_{\mathcal{H}_{\mathcal{Y}}}-\|V_{XY}\Sigma_{YY}^{1/2}g\|_{\mathcal{H}_{\mathcal{X}}}^2\\&=\|\Sigma_{XX}^{1/2}f-V_{XY}\Sigma_{YY}^{1/2}g\|_{\mathcal{H}_{\mathcal{X}}}^2+\langle g,\Sigma_{YY|X}g\rangle_{\mathcal{H}_{\mathcal{Y}}}\end{aligned}$$

であるが，$\overline{\mathcal{R}(V_{XY})}=\overline{\mathcal{R}(\Sigma_{XX})}=\overline{\mathcal{R}(\Sigma_{XX}^{1/2})}$ であることから，第 1 項はいくらでも小さい値を取りえる．これは定理の主張を意味する．□

5.2 条件付独立性の特徴づけ

以下で，条件付(相互)共分散作用素を用いた条件付独立性の特徴づけを 2 通りの方法で行う．以下では Dawid 記法に従い，Z が与えられたもとでの X と Y の条件付独立性を $X\perp\!\!\!\perp Y\,|\,Z$ で表す．

よく知られているように，ユークリッド空間に値を持つ通常のガウス確率ベクトルに対する条件付独立性は

$$X\perp\!\!\!\perp Y\,|\,Z\quad\Longleftrightarrow\quad C_{YX|Z}=O\tag{5.3}$$

によって特徴づけられる．ここで $C_{YX|Z}=C_{YX}-C_{YZ}C_{ZZ}^{-1}C_{ZY}$ は条件付共分散行列である．また，$C_{XX|Z}$ を可逆とするとき，次のような特徴づけも可能である．

$$X\perp\!\!\!\perp Y\,|\,Z\quad\Longleftrightarrow\quad C_{YY|(X,Z)}=C_{YY|Z}\tag{5.4}$$

ここで，$C_{YY|(X,Z)}$ は X,Z を与えたときの Y の条件付分散共分散行列である．$C_{YY|Z}$ は Y を Z によって線形予測したときの平均 2 乗誤差を，$C_{YY|(X,Z)}$ は X と Z を用いて Y を線形予測

したときの平均2乗誤差を表すので，第2の特徴づけは，Z に X を追加しても Y の線形予測が改善しないことを意味している．

以上のガウス変数に対する特徴づけの類似が，一般の確率変数に対しても成り立つ．まず条件付相互共分散作用素 $\Sigma_{YX|Z}$ を考えよう．ガウス確率ベクトルの場合と異なり，一般には $\Sigma_{YX|Z}=O$ と $X \perp\!\!\!\perp Y \,|\, Z$ は同値とは限らない．これは次のことから推察できる．ガウス確率変数の場合には $\mathrm{Cov}[Y,X|Z]$ が Z の値に依存せず $C_{YX|Z}$ に一致するのに対し，一般に $\mathrm{Cov}[Y,X|Z]$ は Z に依存した値を持つ．一方，定理8からわかるように，$\langle g, \Sigma_{YX|Z} f\rangle$ は，Z に関する期待値を取った量しか表せない．したがって Z を固定したときの X と Y の独立性を議論することができない．

しかしながら，X を (X,Z) に置き換え，Z を条件付けの変数としてだけでなく，作用素の働く空間の変数として扱うというトリックを用いると，以下のように条件付独立性が特徴づけられる．

定理 10. 2乗可積分性を仮定し，$k_{\mathcal{Z}}$ は特性的とする．また，$W=(X,Z)$ とし，$\mathcal{X}\times\mathcal{Z}$ 上の正定値カーネルとして $k_{\mathcal{W}}=k_{\mathcal{X}}k_{\mathcal{Z}}$ を用いる．このとき積 $k_{\mathcal{W}}k_{\mathcal{Y}}$ が $(\mathcal{X}\times\mathcal{Z})\times\mathcal{Y}$ 上特性的ならば

$$\Sigma_{YW|Z}=O \qquad \Longleftrightarrow \qquad X \perp\!\!\!\perp Y \,|\, Z \tag{5.5}$$

の同値関係が成立する．

証明. 一般に，任意の可測集合 $A\in\mathcal{B}_{\mathcal{X}}, B\in\mathcal{B}_{\mathcal{Y}}, C\in\mathcal{B}_{\mathcal{Z}}$ に対し，

$$\begin{aligned}
&E[E[\chi_{A\times C}(X,Z)|Z]\,E[\chi_B(Y)|Z]] - E[\chi_{A\times C}(X,Z)\chi_B(Y)] \\
&= E[E[\chi_A(X)|Z]\chi_C(Z)E[\chi_B(Y)|Z]] - E[E[\chi_A(X)\chi_B(Y)|Z]\chi_C(Z)] \\
&= \int_C \{P_{X|Z}(A|z)P_{Y|Z}(B|z) - P_{XY|Z}(A\times B|z)\} dP_Z(z)
\end{aligned}$$

が成り立つ．補題1を用いて集合の定義関数を再生核ヒルベルト空間の関数で近似することにより，$\Sigma_{YW|Z}=O$ という条件が上式第1行が0であることと同値であることがわかる．したがって上式末行の積分が0であることと同値である．これは $P_{X|Z}(A|z)P_{Y|Z}(B|z)-P_{XY|Z}(A\times B|z)=0$ が P_z に関して確率1で成り立つこと，すなわち $X \perp\!\!\!\perp Y \,|\, Z$ と同値である．□

条件付き独立性は X と Y に関して対称な性質なので，上の定理で (X,Z) のかわりに (Y,Z) を用いてもよく，また $(X,Z),(Y,Z)$ をともに用いてもよい．

次に条件付共分散作用素 $\Sigma_{YY|X}$ による特徴づけを見ておこう．

定理 11. 2乗可積分条件を仮定し，$k_{\mathcal{Z}}$ および積 $k_{\mathcal{X}}k_{\mathcal{Z}}$ が特性的であるとするとき，$\mathcal{H}_{\mathcal{Y}}$ 上の自己共役作用素の半順序に関して

$$\Sigma_{YY|Z} \geq \Sigma_{YY|(X,Z)} \tag{5.6}$$

が成立する．ここで $\Sigma_{YY|(X,Z)}$ は (X,Z) に積 $k_{\mathcal{X}}k_{\mathcal{Z}}$ を用いて定義された条件付共分散作用素である．さらに $k_{\mathcal{Y}}$ が特性的であるとき

$$X \perp\!\!\!\perp Y \,|\, Z \quad \Longleftrightarrow \quad \Sigma_{YY|(X,Z)} = \Sigma_{YY|Z}$$

の同値関係が成立する．

この定理の証明は省略する．Fukumizu et al. (2004, 2009) を見ていただきたい．定理9より，式(5.6)は，情報が部分的になれば Y の予測誤差が増加するという自然な事実を意味している．

予測誤差が増加しなければ Z と (X,Z) は Y に関して同じだけの情報量を持つと解釈できるので，定理の同値性は自然で，かつ上で述べたガウス確率ベクトルの場合の拡張となっている．

5.3 条件付相互共分散作用素の推定量

次に，i.i.d. サンプル $(X_1,Y_1,Z_1),\ldots,(X_n,Y_n,Z_n)$ に対して，条件付相互共分散作用素 $\Sigma_{YX|Z}$ の推定量を考えよう．相互共分散作用素の推定量はすでに得ているのでそれを用いて構成すればよいが，$\widehat{\Sigma}_{ZZ}^{(n)}$ はランクが高々 $n-1$ であり逆作用素を持たないので，正則化を用い

$$(\widehat{\Sigma}_{ZZ}^{(n)}+\varepsilon_n I)^{-1}$$

$(\varepsilon_n>0)$ を推定量として用いる．ここで ε_n は正則化のための正定数で $n\to\infty$ のときに $\varepsilon_n\to 0$ となるように定める．これを用いて，条件付相互共分散作用素の推定量 $\widehat{\Sigma}_{YX|Z}^{(n)}$ を

$$\widehat{\Sigma}_{YX|Z}^{(n)}:=\widehat{\Sigma}_{YX}^{(n)}-\widehat{\Sigma}_{YZ}^{(n)}(\widehat{\Sigma}_{ZZ}^{(n)}+\varepsilon_n I)^{-1}\widehat{\Sigma}_{ZX}^{(n)} \tag{5.7}$$

により定める．相互共分散作用素の場合と同様にして，$\widehat{\Sigma}_{YX|Z}^{(n)}$ のヒルベルト=シュミットノルムのグラム行列表示が

$$\begin{aligned}\|\widehat{\Sigma}_{YX|Z}^{(n)}\|_{HS}^2=\frac{1}{n^2}\mathrm{Tr}\big[&\tilde{K}^Y\tilde{K}^X-2\tilde{K}^Y\tilde{K}^Z(\tilde{K}^Z+n\varepsilon_n I_n)^{-1}\tilde{K}^Z\tilde{K}^X\\&+\tilde{K}^Z(\tilde{K}^Z+n\varepsilon_n I_n)^{-1}\tilde{K}^Y\tilde{K}^Z(\tilde{K}^Z+n\varepsilon_n I_n)^{-1}\tilde{K}^Z\tilde{K}^X\big]\end{aligned} \tag{5.8}$$

のように得られる．

下の定理により推定量の一致性が得られる．

定理 12.（Fukumizu et al., 2008）2 乗可積分性を仮定する．正則化定数 ε_n が $\varepsilon_n\to 0$ かつ $\varepsilon_n^2 n\to\infty$ $(n\to\infty)$ を満たすとき

$$\big\|\widehat{\Sigma}_{YX|Z}^{(n)}-\Sigma_{YX|Z}\big\|_{HS}\to 0\qquad(n\to\infty)$$

と確率収束する．

5.4 条件付独立性尺度の応用

本章で紹介した独立性・条件付独立性の尺度は，統計的な問題に多くの応用を持っている．以下ではそのいくつかを簡単に紹介する．その他にも $\|\Sigma_{YX}\|_{HS}^2$ を変数選択に用いた応用など様々な方法が提案されている．

条件付独立性の検定

連続値を取る一般の確率変数に対する条件付独立性の検定は，特に多次元のデータの場合に必ずしも容易ではなく，セル分割やガウス性の仮定などに依る方法が用いられることが多い．これに対し，5.2 節で述べたことから，$\|\widehat{\Sigma}_{Y\dot{X}|Z}^{(n)}\|_{HS}^2$（ここで $\dot{X}=(X,Z)$）を条件付独立性の検定統計量として用いることができる．しかしながら独立性検定に用いた $\|\widehat{\Sigma}_{YX}^{(n)}\|_{HS}^2$ と異なり，帰無仮説のもとでの検定統計量の分布は今のところ知られていない．並べ替え検定を行うことは可能であるが，並べ替えによって条件付独立なサンプルを発生させるためには，条件変数 Z が同一の値をとるサンプル (X_i,Y_i) を並べ替える必要がある．しかし，条件変数 Z が連続値の場合は，ある決まった Z に対して多数の (X_i,Y_i) が存在することを期待するのは難しいので，近い Z の値を持つサンプルをグループ化して，その中で並べ替えをするなどの工夫が必要となる．詳細は Fukumizu et al. (2008) を見ていただきたい．

因果推論への応用

実験を伴わないデータから変数間の統計的な因果関係を推論する問題は，さまざまな分野で重要な課題であるとともに，困難な課題のひとつでもある．因果性推論のための一つのアプローチとして，有向グラフの矢印の向きを因果の向きと考え，真の確率分布が有向非巡回グラフで記述される確率分布に従うと仮定して，データからそのグラフを推定する方法がある(Pearl, 2000; Spirtes et al., 2001)．このアプローチでは，グラフの構造を推定するために，変数の部分集合に対する条件付独立性の判定を利用する．従来よく用いられる因果推論の方法では，連続変数に対する条件付独立性の検定には，変数の離散化を行うか，ガウス性を仮定して χ^2 検定を行う方法が主であった．Sun et al. (2007) では，上で述べたヒルベルト=シュミットノルムによる条件付独立性検定を行い，変数間の因果関係の推論を行う方法を提案している．

カーネル次元削減法

高次元データを扱う際に，データの説明や可視化，予測・決定の精度向上のためのノイズ削減，計算量の軽減などさまざまな目的のために次元削減は重要な方法となっている．ここでは m 次元説明変数 X を用いて従属変数 Y を説明する回帰の問題において，Y に関する情報を保持するような X の低次元部分空間への射影を見つける次元削減の問題を考える．r 次元部分空間への射影を表す $m \times r$ 行列を B ($B^T B = I_r$) とするとき，回帰において Y を説明するのに十分な部分空間の条件は

$$Y \perp\!\!\!\perp X \,|\, U \qquad (U = B^T X) \tag{5.9}$$

という条件付独立性で与えるのが自然である．上式を満たす B の列ベクトルが張る部分空間を推定する問題に対し，正定値カーネルによる条件付独立性の特徴づけが適用可能である．射影行列 B の推定関数として

$$\min_{B: B^T B = I_r} \mathrm{Tr}\big[\widehat{\Sigma}^{(n)}_{YY|B^T X}\big]$$

を用いる方法を，カーネル次元削減法という(Fukumizu et al., 2004, 2009; 福水, 2005)．ただし，$\widehat{\Sigma}^{(n)}_{YY|B^T X}$ を定義する際には，Y，$\mathbb{R}^d$ に対して特性的な正定値カーネル $k_{\mathcal{Y}}, k_d$, を用意し，$B^T X_i$ に対して $k_d(B^T X_i, B^T X_j)$ を用いる．

カーネル次元削減法の導出には，周辺分布，条件付分布および可測集合に関する条件をほとんど必要としない．離散変数などを含んだ幅広い状況に応用可能である．類似の目的に対する従来の方法は，回帰曲線や周辺分布に強い条件が課されることが多く，それに比べてカーネル次元削減法は一般的な状況に適用可能である．

6. おわりに

本論文では，正定値カーネルを用いたデータ解析の新しい流れとして，特徴写像によって再生核ヒルベルト空間に写像したデータの共分散や条件付共分散を考えることにより，もとのデータの独立性や条件付独立性を議論するための方法を紹介した．そのために重要な概念は，3 章で導入した特性的な正定値カーネルであった．

特性的なカーネルと特性関数との類似性からわかるように，正定値カーネルを用いた独立性・条件付き独立性の議論は，特性関数によって独立性・条件付き独立性を扱う方法と類似的であり，ともに確率分布 P を積分変換 $\int k(x,x')dP(x')$ によって表現する．特性関数を用いる場合，$\mathbb{R}^m, \mathbb{R}^\ell$ に値をとる確率ベクトル X, Y の独立性を特徴づけるためには

$$E[e^{\sqrt{-1}(X^T\omega + Y^T\eta)}] - E[e^{\sqrt{-1}X^T\omega}]E[e^{\sqrt{-1}Y^T\eta}] = 0$$

を判定すればよいが，左辺は (ω, η) の関数であるため，何らかの関数空間のノルムを導入して，

その値が 0 かどうかを判定する必要がある．例えば適当な測度 μ_m, μ_ℓ によって定義されるヒルベルト空間 $L^2(\mathbb{R}^{m+\ell}, \mu_m \times \mu_\ell)$ を用いて

$$\int \left| E[e^{\sqrt{-1}(X^T\omega + Y^T\eta)}] - E[e^{\sqrt{-1}X^T\omega}]E[e^{\sqrt{-1}Y^T\eta}] \right|^2 d\mu_m(\omega) d\mu_\ell(\eta)$$

を用いることは可能であるが，この積分を有限サイズのデータから推定することは特別な場合を除いて容易ではない．正定値カーネルの方法では，再生核ヒルベルト空間の特殊性によりこの積分の問題を回避することができる．また，特性関数と異なり任意の測度空間上に値をとる確率変数を考えることができる．

本論文で議論した方法の原理は，再生核ヒルベルト空間で低次(1 次，2 次)の統計量を考えれば，確率変数の性質がすべて把握できるという考えに基づいており，ガウス確率変数に対する方法の拡張と考えることもできる．この立場に立って，再生核ヒルベルト空間によるノンパラメトリック推定の方法論は，さらに幅広い手法に拡張可能であることが期待されており，筆者も含めさまざまな研究者がその研究を進めている．

謝　辞

本研究の一部は科研費・基盤研究(B) 22300098 による助成を受けた．また査読者によるコメントは本稿の改善に非常に有益であった．ここに感謝の意を表したい．

参 考 文 献

Aronszajn, N. (1950). Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68** (3), 337–404.

Baker, C. R. (1973). Joint measures and cross-covariance operators, *Transactions of the American Mathematical Society*, **186**, 273–289.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, Boston, Massachusetts.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers, *Fifth Annual ACM Workshop on Computational Learning Theory* (ed. D. Haussler), 144–152, ACM Press, Pittsburgh, Pennsylvania.

福水健次(2005). 正定値カーネルによる回帰問題における次元削減法，統計数理，**53** (2), 189–200.

福水健次(2010). 『カーネル法入門―正定値カーネルによるデータ解析―』，朝倉書店，東京.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, *Journal of Machine Learning Research*, **5**, 73–99.

Fukumizu, K., Gretton, A., Sun, X. and Schölkopf, B. (2008). Kernel measures of conditional dependence, *Advances in Neural Information Processing Systems 20*, 489–496, MIT Press, Cambridge, Massachusetts.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2009). Kernel dimension reduction in regression, *Annals of Statistics*, **37** (4), 1871–1905.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B. and Smola, A. (2007). A kernel method for the two-sample-problem, *Advances in Neural Information Processing Systems 19* (eds. B. Schölkopf, J. Platt and T. Hoffman), 513–520, MIT Press, Cambridge, Massachusetts.

Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B. (2010). A fast, consistent kernel two-sample test, *Advances in Neural Information Processing Systems 22* (eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta), 673–681, MIT Press, Cambridge,

Massachusetts.

Ku, C. -J. and Fine, T. L. (2005). Testing for stochastic independence: Application to blind source separation, *IEEE Transactions on Signal Processing*, **53**(5), 1815–1826.

Pearl, J. (2000). *Causality*: *Models, Reasoning and Inference*, Cambridge University Press, Cambridge, U.K.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*, MIT Press, Cambridge, Massachusetts.

Spirtes, P., Glymour, C. and Scheines, R. (2001). *Causation, Prediction and Search*, 2nd ed., MIT Press, Cambridge, Massachusetts.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G. (2010). Hilbert space embeddings and metrics on probability measures, *Journal of Machine Learning Research*, **11**, 1517–1561.

Sun, X., Janzing, D., Schölkopf, B. and Fukumizu, K. (2007). A kernel-based causal learning algorithm, *Proceedings of the 24th International Conference on Machine Learning* (*ICML2007*), 855–862.

206 Proceedings of the Institute of Statistical Mathematics Vol. 58, No. 2, 185–206 (2010)

Nonparametric Inference with Positive Definite Kernels

Kenji Fukumizu

The Institute of Statistical Mathematics

The methodology of data analysis with positive definite kernels or reproducing kernel Hilbert spaces is called the "kernel method", which has been developed in the machine learning field. The feature of this methodology is that data are mapped to reproducing kernel Hilbert spaces given by the positive definite kernel, and linear methods of data analysis are applied on the data mapped in the Hilbert spaces. Although the mapped data may be infinite dimensional, the special property of the inner product makes the computation efficient. More recently, it has been revealed that more basic statistics such as mean and covariance considered in reproducing kernel Hilbert spaces are useful in analyzing statistical properties such as homogeneity, independence, and conditional independence of random variables. This paper explains the basic idea of this method for new nonparametric inference, and gives a brief survey of results obtained so far, focusing particularly on nonparametric methods for discussing independence and conditional independence of variables. In discussing the properties of variables, it is important to use a class of kernels that determines a probability uniquely by the mean on the reproducing kernel Hilbert spaces. This class of kernels is called characteristic, and some theoretical analysis is also shown. With a characteristic kernel, the squared distance of the means can be applied to the two-sample test for homogeneity. If the joint probability and product of the marginals are compared, the distance is equal to the Hilbert-Schmidt norm of the covariance operator, which can be used for independence test. It is also shown that the Hilbert-Schmidt norm of the normalized covariance operator is equal to the chi-square divergence, which is a well-known measure of dependence. In the last part of this paper, the method of discussing conditional independence with kernels is briefly surveyed. The method is based on extension of the characteristic of conditional independence of Gaussian random variables to general cases by mapping variables to reproducing kernel Hilbert spaces. Some applications of conditional independence are also shown.

Key words: Positive definite kernel, reproducing kernel, Hilbert space, nonparametric, independence, conditional independence.