STATISTICAL INFERENCE WITH REPRODUCING KERNELS

Kenji Fukumzu

Key words: statistical inference, reproducing kernel Hilbert space, positive definite kernel

AMS Mathematics Subject Classification:. 46E22; 62G05

Abstract. Reproducing kernels has been recently applied to statistical inference problems by using the *kernel mean* expression of probability distributions. Given a random variable X taking values on a measurable space and a reproducing kernel Hilbert space $\mathcal{H}$ on that space with a positive definite kernel k, the *kernel mean* is defined by $E[k(\cdot, X)] \in \mathcal{H}$. This gives a mapping from the probabilities to $\mathcal{H}$. If this mapping is injective, the kernel mean uniquely identifies the probability. This class of kernel is useful for statistical inference and called *characteristic*. This paper gives a brief review on how characteristic kernels can be applied to derive practical methods for statistical inference problems, and discusses conditions that a positive definite kernel is characteristic.

1 Introduction

Statistical inference concerns problems of estimating or testing the relations and properties of distributions of random variables using finite number of data. There are various methods which solve specific tasks of statistical inference problems. Positive kernels or reproducing kernel and reproducing kernel Hilbert spaces have been proved to be useful for statistical data analysis since 1990's [12]. This paper explains a more recent methodology of statistical inference using kernel means.

Let X be a random variable taking values on a measurable space $(\mathcal{X}, \mathcal{B})$, where $\mathcal{B}$ is a σ-algebra on $\mathcal{X}$, and let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS in short) on $\mathcal{X}$ defined by a bounded measurable positive definite kernel k. Define *kernel mean*, m_X^k, of X on $\mathcal{H}$ by

$$m_X^k = E[k(\cdot, X)] \quad \in \mathcal{H}. \tag{1}$$

Note that by reproducing property of k, m_X^k satisfies

$$\langle m_X^k, f \rangle_{\mathcal{H}} = E[f(X)] \tag{2}$$

for any $f \in \mathcal{H}$. If there is no confusion, the kernel mean is simply denoted by m_X by omitting k. Since the kernel mean depends only on the probability distribution of X, it is also denoted by m_P^k or m_P if the distribution is P.

Let $\mathcal{P}$ be the set of probability measures on $(\mathcal{X}, \mathcal{B})$. A bounded measurable kernel k on $(\mathcal{X}, \mathcal{B})$ is called *characteristic* (with respect to $(\mathcal{X}, \mathcal{B})$) if the mapping

$$\mathcal{P} \to \mathcal{H}, \quad P \mapsto m_P^k$$

is injective. Because the kernel mean with a characteristic kernel uniquely identifies the probability, inference problems on the properties of probability distributions can be cast into the inference on the kernel means. For example, as we will see in Section 2, independence of two random variables X and Y can be tested by comparing the kernel mean of the joint variable (X, Y) and the product of kernel means of marginals.

In statistical inference, the kernel mean m_X should be estimated with finite number of data. Suppose $X_1, \ldots, X_n$ is an i.i.d. sample with the same distribution as X. The *empirical kernel mean* $\widehat{m}_X^{(n)}$ is defined by

$$\widehat{m}_X^{(n)} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i).$$

It is known [2] that $\widehat{m}_X^{(n)}$ is a strongly consistent estimator, i.e., $\|\widehat{m}_X^{(n)} - m_X\|$ converges to zero in probability as $n \to \infty$. Moreover, $\sqrt{n}(\widehat{m}_X^{(n)} - m_X)$ converges to a Gaussian process. One of the advantages of using positive definite kernels and RKHS in statistical methods lies in the reproducing property: various useful quantities defined with the RKHS norm of the empirical kernel means can be exactly computed, while they are elements in infinite dimensional functional spaces. We will see some examples in Section 2.

In discussing the relation between two random variables, covariance is an essential notion. Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ be measurable spaces, and (X, Y) be an random variable taking values in $\mathcal{X} \times \mathcal{Y}$. Suppose $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are RKHS's with bounded measurable positive definite kernels $k_\mathcal{X}$ on $\mathcal{X}$ and $k_\mathcal{Y}$ on $\mathcal{Y}$, respectively. The *cross-covariance operator* $\Sigma_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is the operator that satisfies

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = E[g(Y)f(X)] - E[g(Y)]E[f(X)]$$

for any $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$. The cross-covariance operator can be also defined by $\Sigma_{YX} = E[m_{(YX)}^{k_\mathcal{Y} k_\mathcal{X}} - m_Y^{k_\mathcal{Y}} \otimes m_X^{k_\mathcal{X}}]$, where the product space $\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ associated with the product kernel $k_\mathcal{Y} k_\mathcal{X}$ is identified with the space of bounded linear operators from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$ in a standard way. Obviously $\Sigma_{YX}^* = \Sigma_{XY}$, where A^* denotes the adjoint operator of A, and it is easy to see that Σ_{YX} is a Hilbert-Schmidt operator.

When $Y = X$, the self-adjoint operator Σ_{XX} is called *covariance operator*. Σ_{XX} is a self-adjoint, trace class operator.

In a similar manner to the kernel mean, given $(X_1, Y_1), \ldots, (X_n, Y_n)$ the *empirical cross-covariance operator* is defined by

$$\widehat{\Sigma}_{YX}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} k_{\mathcal{Y}}(\cdot, Y_i) \otimes k_{\mathcal{X}}(\cdot, X_i) - \Big(\frac{1}{n}\sum_{i=1}^{n} k_{\mathcal{Y}}(\cdot, Y_i)\Big) \otimes \Big(\frac{1}{n}\sum_{i=1}^{n} k_{\mathcal{X}}(\cdot, X_i)\Big)$$

in the tensor form. $\widehat{\Sigma}_{YX}^{(n)}$ converges to Σ_{YX} in Hilbert-Schmidt norm at the rate of $n^{-1/2}$.

2 Statistical inference with kernel means

This section describes some examples of statistical inference with kernel means and cross-covariance operators.

2.1 Two sample test

Suppose we have two i.i.d. samples $X_1, \ldots, X_\ell$ and $Y_1, \ldots, Y_n$ with law P and Q, respectively. We wish to determine whether $P = Q$ or not. This problem is called two-sample homogeneity test, and has been long studied in statistical literature. In statistical terminology, the null hypothesis is $P = Q$, and the alternative hypothesis is $P \neq Q$. With a characteristic kernel, the problem of comparing two probabilities can be cast into the problem of comparing two kernel means. This motivates us to define the following test statistic:

$$T_{\ell,n} = \Big\| \frac{1}{\ell}\sum_{i=1}^{\ell} k(\cdot, X_i) - \frac{1}{n}\sum_{j=1}^{\ell} k(\cdot, Y_j) \Big\|^2$$

by introducing a positive definite kernel k. Note that this gives an empirical estimator for the squared distance measure of probabilities $\|m_P - m_Q\|^2$. By virtue of the reproducing property, we have

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

A small value of $T_{\ell,n}$ is expected under the null hypothesis $P = Q$, and the null hypothesis is rejected with the error probability α (significance level) if $T_{\ell,n}$ is larger

than some threshold θ_α. The region of rejection is called critical region. For this test statistic, a better statistical property is obtained by debiasing it, which results in

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

It is known ([15], Chap. 12) that $U_{\ell,n}$ is in the class of U-statistics and the asymptotic distribution of $U_{\ell,n}$ under $\ell, n \to \infty$ with constraint $\ell/(\ell+n) \to \gamma \in (0,1)$ is a mixture of χ-square distributions. With this asymptotic distribution, we can determine θ_α for the test. Gretton et al. [8, 9] show some practical applications in comparing with other methods.

2.2 Independence test

Testing independence or dependence of two random variables is an important problem in many situations. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample on $\mathcal{X} \times \mathcal{Y}$ with law P, and consider the statistical test for independence of X_i and Y_i: the null hypothesis is that they are independent, and the alternative hypothesis is that they are not. This problem can be regarded as a special case of two sample test, since we want to compare two probabilities P and $P_X \otimes P_Y$, where P_X and P_Y are marginal probabilities of X_i and Y_i, respectively. Prepare bounded measurable positive definite kernels $k_\mathcal{X}$ on $\mathcal{X}$ and $k_\mathcal{Y}$ for $\mathcal{Y}$ such that $k_\mathcal{X} k_\mathcal{Y}$ is a characteristic kernel on $\mathcal{X} \times \mathcal{Y}$. We can then use the statistic $\|\widehat{m}_{XY} - \widehat{m}_X \otimes \widehat{m}_Y\|^2$ for testing independence of X_i and Y_i. It is easy to see that this is equivalent to using the squared Hilbert-Schmidt norm of the empirical cross-covariance operator $\widehat{\Sigma}^{(n)}_{YX}$. The test statistics is thus given by

$$\begin{aligned}
&\|\widehat{\Sigma}^{(n)}_{YX}\|^2_{HS} \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} k_\mathcal{X}(X_i, X_j) k_\mathcal{Y}(Y_i, Y_j) - \frac{2}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} k_\mathcal{X}(X_i, X_j) \sum_{\ell=1}^{n} k_\mathcal{Y}(Y_i, Y_\ell) \\
&\quad + \frac{1}{n^4} \sum_{i,j=1}^{n} k_\mathcal{X}(X_i, X_j) \sum_{\ell,r=1}^{n} k_\mathcal{Y}(Y_\ell, Y_r).
\end{aligned}$$

In matrix notation,

$$\|\widehat{\Sigma}^{(n)}_{YX}\|^2_{HS} = \frac{1}{n^2} \mathrm{Tr}\big[K_X Q_n K_Y Q_n\big],$$

where K_X and K_Y are Gram matrices given by $(k_{\mathcal{X}}(X_i, X_j))_{ij}$ and $(k_{\mathcal{Y}}(Y_i, Y_j))_{ij}$, respectively, and $Q_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ with $\mathbf{1}_n = (1, \ldots, 1)^T$.

In a similar manner to the two sample test, the asymptotic distribution of the test statistic for $n \to \infty$ is known and can be used for determining the critical region of the independence test at a significance level. Alternatively, we can also use permutation test, which simulates the distribution under the independence assumption by random permutation of either of (X_i) or (Y_i). For the details of this test statistic and numerical examples, see [8, 10].

Another important statistical notion is conditional independence, which is widely used in statistical inference for graphical modeling, causal inference, and Bayesian methods. This paper describes only a sketch of the kernel method for conditional independence, and leaves the details to the original papers [4, 6]. Suppose we have random variables (X, Y, Z) on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, and bounded measurable positive definite kernels $k_{\mathcal{X}}, k_{\mathcal{Y}}, k_{\mathcal{Z}}$ on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, respectively. The respective RKHS are denoted by $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{Z}}$. The *conditional cross-covariance operator* from X to Y given Z is the operator from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$ defined by

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}. \tag{3}$$

Here the operator $\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$ should be rigorously interpreted as $\Sigma_{YY}^{1/2}V_{YZ}V_{ZX}\Sigma_{XX}^{1/2}$, where V_{YZ} is given by the unique operator in the decomposition [1] $\Sigma_{YZ} = \Sigma_{YY}^{1/2}V_{YZ}\Sigma_{ZZ}^{1/2}$ with $|V_{YZ}| \leq 1$, $\mathcal{R}(V_{YZ}) \subset \overline{\mathcal{R}(\Sigma_{YY})}$ and $\mathcal{N}(V_{YZ})^{\perp} \subset \overline{\mathcal{R}(\Sigma_{ZZ})}$. V_{ZX} is given similarly.

The conditional cross-covariance operator is related to the conditional covariance as follows.

Proposition 1 ([3]). *Assume $k_{\mathcal{Z}}$ is characteristic. Then, for any $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$*

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

The above definition is a straightforward extension of the conditional covariance of the Gaussian random variables: for a Gaussian random vector (X, Y, Z) the conditional covariance between X and Y given Z is given by

$$C_{YX|Z} = C_{YX} - C_{YZ}C_{ZZ}^{-1}C_{ZX},$$

where the existence of C_{ZZ}^{-1} is assumed. It is well known that for Gaussian variables X and Y are conditionally independent given Z if and only if $C_{YX|Z} = 0$. As an extension of this fact, we have the following theorem.

Theorem 2 ([3]). *Define $W = (X, Z)$ and use the product kernel $k_{\mathcal{W}} = k_{\mathcal{X}} k_{\mathcal{Z}}$ for W. Assume that $k_{\mathcal{Z}}$ and $k_{\mathcal{Y}} k_{\mathcal{W}}$ are characteristic kernels on $\mathcal{Z}$ and $\mathcal{Y} \times (\mathcal{X} \times \mathcal{Z})$, respectively. Then, X and Y are conditional independent given Z if and only if $\Sigma_{YW|Z} = O$.*

Note that in the above theorem the joint variable $W = (X, Z)$ is used to check the conditional independence. As is shown in Proposition 3, the conditional cross-covariance operator can handle the conditional covariance between $f(X)$ and $g(Y)$ given Z only on average over Z, though conditional independence requires $\mathrm{Cov}[f(X), g(Y)|Z] = 0$ for almost every Z. Intuitively, the joint variable $W = (X, Z)$ in $\Sigma_{WY|Z}$ makes it possible to handle each value of Z.

As $\Sigma_{YW|Z}$ is a Hilbert-Schmidt operator, the squared Hilbert-Schmidt norm $\|\Sigma_{YW|Z}\|_{HS}^2$ can be used for discussing conditional independence or dependence. Further discussions on this statistic can be found in [4], and an application to causal inference is proposed in [14].

3 Characteristic kernels on LCA group

As we have seen in the previous section, the characteristic property of a kernel is important in its statistical applications. This section discusses some conditions of this property. We first start with a general condition.

Proposition 3. *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and k be a measurable bounded positive definite kernel on $\mathcal{X}$ with RKHS $\mathcal{H}_k$. Then, k is characteristic if and only if $\mathcal{H}_k + \mathbb{R}$ is dense in $L^2(P)$ for any probability measure P on $(\mathcal{X}, \mathcal{B})$, where $\mathcal{H}_k + \mathbb{R} = \{f + c \mid f \in \mathcal{H}_k, c \in \mathbb{R}\}$.*

Proof. Suppose $m_P = m_Q$ for different probabilities P and Q while $\mathcal{H}_k + \mathbb{R}$ is dense in $L^2(|P - Q|)$, where $|P - Q|$ is the total variation of $P - Q$. Then, for any $E \in \mathcal{B}$ and $\varepsilon > 0$ there is $f \in \mathcal{H}_k$ and $c \in \mathbb{R}$ such that $\int |f + c - \chi_E| d|P - Q| < \varepsilon$. Here χ_E is the indicator function of E. This means $|(\int f dP - P(E)) - (\int f dQ - Q(E))| < \varepsilon$. It follows from the assumption $m_P = m_Q$ that $\int f dP = \int f dQ$, which implies $|P(E) - Q(E)| < \varepsilon$. As $\varepsilon > 0$ is arbitrary, $P(E) = Q(E)$, which causes contradiction.

Next, suppose $\mathcal{H}_k + \mathbb{R}$ is not dense in $L^2(P)$ for some probability P. Then, there is nonzero $f \in L^2(P)$ such that $\int f g dP = 0$ for any $g \in \mathcal{H}_k$ and $\int f dP = 0$. Define two different probabilities Q_1 and Q_2 by $Q_1(E) = \int_E |f| dP / \|f\|_{L^1(P)}$ and $Q_2(E) = \int_E (|f| - f) dP / \|f\|_{L^1(P)}$. Then, for any $g \in \mathcal{H}_k$, $\int g dQ_1 - \int g dQ_2 = \int f g dP / \|f\|_{L^1(P)} = 0$. This implies $m_{Q_1} = m_{Q_2}$, hence k is not characteristic. □

3.1 Harmonic Analysis on LCA group

This subsection gives a brief review of the harmonic analysis on locally compact group. For the details, see e.g. [11].

A complex-valued Radon measure μ on a locally compact space X is said to be *regular* if $|\mu|$ is outer regular, that is, $|\mu|(E) = \inf\{|\mu|(U) \mid U \text{ is an open set including } E\}$ holds for every Borel set E. The set of regular measures on X is denoted by $M(X)$. For a finite regular measure, there is the largest open set U with $|\mu|(U) = 0$. The complement of U is called the *support* of μ, and denoted by $\mathrm{supp}(\mu)$

Let G be a group. A function $\phi : G \to \mathbb{C}$ is called *positive definite* if $k(x, y) = \phi(y^{-1}x)$ is a positive definite kernel. This type of positive definite kernel is called *shift-invariant*, since $k(zx, zy) = k(x, y)$. There are many examples of shift-invariant positive definite kernels, which are used in practical applications in statistics: Gaussian RBF kernel $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$ and Laplacian kernel $k(x, y) = \exp(-\beta \sum_{i=1}^{n} |x_i - y_i|)$ are famous ones on the additive group $\mathbb{R}^n$.

A particulary interesting class of group in discussing positive definite kernels is locally compact Abelian groups (LCA group, in short), on which famous Bochner's theorem characterizes the continuous positive definite functions. For a LCA group the additive notation $x + y$ is employed for the group operation hereafter.

A function $\gamma : G \to \mathbb{C}$ is called a *character* of a LCA group G if $\gamma(x + y) = \gamma(x)\gamma(y)$ and $|\gamma(x)| = 1$ for all $x, y \in G$. The *dual group* $\widehat{G}$ of G is the set of all continuous characters of G. $\widehat{G}$ is an Abelian group with the value multiplication, which is conventionally denoted by addition, i.e., $(\gamma_1 + \gamma_2)(x) := \gamma_1(x)\gamma_2(x)$. For any $x \in G$, the function $\hat{x}$ on $\widehat{G}$ given by $\hat{x}(\gamma) = \gamma(x)$ $(\gamma \in \widehat{G})$ defines a character of $\widehat{G}$. It is known that $\widehat{G}$ is a LCA group if the weakest topology is introduced so that $\hat{x}$ is continuous for each $x \in G$. As is well known, the Pontryagin duality guarantees that the group homomorphism $G \to \widehat{\widehat{G}}$, $x \mapsto \hat{x}$ is isomorphism and homeomorphic, where $\widehat{\widehat{G}}$ is the dual group of $\widehat{G}$, and thus $\widehat{\widehat{G}}$ can be identified with G. In view of this duality, it is customary to write $(x, \gamma) := \gamma(x)$.

Let $f \in L^1(G)$ and $\mu \in M(G)$, the Fourier transform of f and μ are respectively defined by

$$\hat{f}(\gamma) = \int_G (-x, \gamma) f(x) dx, \qquad \hat{\mu}(\gamma) = \int_G (-x, \gamma) d\mu(x), \qquad (\gamma \in \widehat{G}), \tag{4}$$

where dx is the Haar measure of G. Note that $\hat{f}$ and $\hat{\mu}$ are continuous. For $f \in L^\infty(G)$, $g \in L^1(G)$, and $\mu \in M(G)$, the convolutions are defined respectively

by

$$(g * f)(x) = \int_G f(x-y)g(y)dy, \quad (\mu * f)(x) = \int_G f(x-y)d\mu(y).$$

The convolution $g * f$ is uniformly continuous on G. For any $f, g \in L^1(G)$ and $\mu \in M(G)$, the following relations hold:

$$\widehat{f * g} = \hat{f}\hat{g}, \qquad \widehat{\mu * f} = \widehat{\mu}\widehat{f}. \tag{5}$$

For a LCA group, the continuous positive definite functions are characterized in the following theorem.

Theorem 4 (Bochner's theorem). *A continuous function ϕ on G is positive definite if and only if there is a non-negative measure $\Lambda \in M(\widehat{G})$ such that*

$$\phi(x) = \int_{\widehat{G}} (x, \gamma) d\Lambda(\gamma) \qquad (x \in G). \tag{6}$$

Moreover, such Λ is unique for each ϕ.

Bochner's theorem implies that the continuous positive definite functions form a convex cone with the extreme points given by the dual group $\widehat{G}$.

3.2 Characteristic kernels on LCA group

Let G be a LCA group and k be a shift invariant positive definite kernel on G. We wish to give conditions that k is characteristic. Before going to the formal theorems, we show an intuitive explanation. First note that for a shift invariant kernel k, the kernel mean m_X for a random variable X with law P is given by

$$m_X(x) = \langle m_X, k(\cdot, x) \rangle = \int_G k(x-y)dP(y) = (\phi * P)(x).$$

Thus k is characteristic if and only if $\phi * (P - Q) \neq 0$ for any different probabilities P and Q. By Fourier transforms, this holds if $\widehat{\phi}\widehat{\mu} \neq 0$ for any nontrivial finite signed measure μ. Based on Bochner's theorem, a sufficient condition is easily obtained.

Theorem 5 ([5]). *Let ϕ be a continuous positive definite function on a LCA group G given by Eq. (6) with Λ. If $\mathrm{supp}(\Lambda) = \widehat{G}$, then the positive definite kernel $k(x, y) = \phi(x - y)$ is characteristic.*

Proof. It suffices to prove that if $\mu \in M(G)$ satisfies $\mu * \phi = 0$ then $\mu = 0$. By Fubini's theorem,

$$\int_G (\mu * \phi)(x) d\mu(x) = \int_G \int_G \phi(x-y) d\mu(y) d\mu(x)$$
$$= \int_{\widehat{G}} \int_G (x,\gamma) d\mu(x) \int_G (-y,\gamma) d\mu(y) d\Lambda(\gamma) = \int_{\widehat{G}} |\widehat{\mu}(\gamma)|^2 d\Lambda(\gamma).$$

If $\mu * \phi = 0$, it follows from the continuity of $\widehat{\mu}$ and $\mathrm{supp}(\Lambda) = \widehat{G}$ that $\widehat{\mu} = 0$, which means $\mu = 0$ by the duality. □

In real-valued cases, the condition $\mathrm{supp}(\Lambda) = \widehat{G}$ is almost necessary.

Theorem 6 ([5]). *Let ϕ be a $\mathbb{R}$-valued continuous positive definite function on a LCA group G given by Eq. (6) with Λ. The positive definite kernel $k(x,y) = \phi(x-y)$ is characteristic if and only if either of the following (i) or (ii) holds: (i) G is non-compact and $\mathrm{supp}(\Lambda) = \widehat{G}$, or (ii) G is compact and $\mathrm{supp}(\Lambda) \supset \widehat{G} - \{0\}$.*

Proof. It is obvious that k is characteristic if and only if so is $k(x,y) + 1$. Since $\int_{\widehat{G}} (x,\gamma) d\delta_0$ is a positive constant for compact G, where δ_0 is the Dirac measure at $0 \in \widehat{G}$, we can assume w.l.o.g. that $0 \in \mathrm{supp}(\Lambda)$. The "if" part is thus given by Theorem 5.

For "only if" part, assuming $\mathrm{supp}(\Lambda) \neq \widehat{G}$, we will construct two different probabilities P_1 and P_2 such that $(P_1 - P_2) * \phi = 0$. In the following, for a set A in G or $\widehat{G}$, the notations $-A = \{-x \mid x \in A\}$, $A - A = \{x - y \mid x, y \in A\}$, and $A + x = \{y + x \mid y \in A\}$ are used. Since ϕ is real-valued, $\Lambda(-E) = \Lambda(E)$ for every Borel set E. Thus $U := \widehat{G} \backslash \mathrm{supp}(\Lambda)$ is a non-empty open set with $-U = U$ and $0 \notin U$. Fix $\gamma_0 \in U$. By the continuity of $(\gamma_1, \gamma_2) \mapsto \gamma_1 - \gamma_2$, there exits an open neighborhood W of $0 \in \widehat{G}$ such that $\pm\gamma_0 \notin W - W$, $\mathrm{cl}(W - W) \pm \gamma_0 \subset U$, and $(W - W) + \gamma_0 \cap (W - W) - \gamma_0 = \emptyset$.

Let $g = \chi_W * \chi_{-W}$, where χ_E denotes the indicator function of E. It is easy to see that g is continuous and positive definite. By Bochner's theorem and Pontryagin duality, there is a nonzero, non-negative measure $\mu \in M(G)$ such that

$$g(\gamma) = \int_G (x,\gamma) d\mu(x) \qquad (\gamma \in \widehat{G}).$$

Define a function h on $\widehat{G}$ by

$$h(\gamma) := g(\gamma - \gamma_0) + g(\gamma + \gamma_0) = \int_G (x,\gamma) d((\gamma_0 + \overline{\gamma_0})\mu)(x),$$

where $(\gamma_0 + \overline{\gamma_0})\mu$ is a signed measure defined by $((\gamma_0 + \overline{\gamma_0})\mu)(E) = \int_E (\gamma_0(x) + \overline{\gamma_0}(x))d\mu(x)$. Note $\mathrm{supp}(g) \subset \mathrm{cl}(W-W)$. Since $\mathrm{cl}(W-W)+\gamma_0 \cap \mathrm{cl}(W-W)-\gamma_0 = \emptyset$ and g is nonzero, h is a nonzero function. Also $\mathrm{supp}(h) \subset \mathrm{cl}(W-W)+\gamma_0 \cup \mathrm{cl}(W-W)-\gamma_0 \subset U$, which does not contain 0. Thus, by setting $\gamma = 0$, we have

$$((\gamma_0 + \overline{\gamma_0})\mu)(G) = 0. \tag{7}$$

Let $m = |(\gamma_0 + \overline{\gamma_0})\mu|(G)$ $(\neq 0)$, and define two different probability measures by

$$P_1 = \frac{1}{m}|(\gamma_0 + \overline{\gamma_0})\mu|, \qquad P_2 = \frac{1}{m}\{|(\gamma_0 + \overline{\gamma_0})\mu| - (\gamma_0 + \overline{\gamma_0})\mu\}.$$

From Fubini's theorem,

$$\begin{aligned} m \cdot ((P_1 - P_2) * \phi)(x) &= \int_G \phi(x-y)(\gamma_0(y) + \overline{\gamma_0}(y))d\mu(y) \\ &= \int_{\widehat{G}} (x,\gamma) \int_G \overline{\{(y,\gamma-\gamma_0) + (y,\gamma+\gamma_0)\}} d\mu(y) d\Lambda(\gamma) \\ &= \int_{\widehat{G}} (x,\gamma)\{g(\gamma-\gamma_0) + g(\gamma+\gamma_0)\}d\Lambda(\gamma) = \int_{\widehat{G}} (x,\gamma)h(\gamma)d\Lambda(\gamma). \end{aligned}$$

Since $\mathrm{supp}(h) \subset U = \widehat{G}\backslash\mathrm{supp}(\Lambda)$, we have $(P_1 - P_2) * \phi = 0$. □

Theorems 5 and 6 are generalization of the results in [13]. From Theorem 6, we can see that the characteristic property is stable under the product for real-valued shift-invariant continuous kernels.

Corollary 7 ([5]). *Let $\phi_1(x-y)$ and $\phi_2(x-y)$ be $\mathbb{R}$-valued continuous shift-invariant characteristic kernels on a LCA group G. If (i) G is non-compact, or (ii) G is compact and $2\gamma \neq 0$ for any nonzero $\gamma \in \widehat{G}$. Then $(\phi_1\phi_2)(x-y)$ is characteristic.*

Proof. We show the proof only for (i). Let Λ_1, Λ_2 be the non-negative measures to give ϕ_1 and ϕ_2, respectively, in Eq. (6). By Theorem 6, $\mathrm{supp}(\Lambda_1) = \mathrm{supp}(\Lambda_2) = \widehat{G}$. This means $\mathrm{supp}(\Lambda_1 * \Lambda_2) = \widehat{G}$. The proof is completed because $\Lambda_1 * \Lambda_2$ gives a positive definite function $\phi_1\phi_2$. □

Example 1. $(\mathbb{R}^n, +)$: Gaussian RBF kernel $\exp(-\frac{1}{2\sigma^2}\|x-y\|^2)$ and Laplacian kernel $\exp(-\beta\sum_{i=1}^n |x_i - y_i|)$ are characteristic on $\mathbb{R}^n$, since the corresponding non-negative measures are $\exp(-\frac{\sigma^2}{2}\|\omega\|^2)$ and $\prod_{j=1}^n 1/(1+\omega_j^2)$, respectively, up to positive constant. An example of a positive definite kernel that is *not* characteristic

on $\mathbb{R}^n$ is $\mathrm{sinc}(x-y) = \frac{\sin(x-y)}{x-y}$: the Fourier transform is the indicator function of a bounded interval.

Example 2. $([0, 2\pi), +)$: The addition is made modulo 2π. The dual group is $\{e^{\sqrt{-1}nx} \mid n \in \mathbb{Z}\}$, and expression of the Bochner's theorem is given by Fourier expansion,

$$\phi(x) = \sum_{n=-\infty}^{\infty} a_n e^{\sqrt{-1}nx}, \qquad a_n \geq 0, \quad \sum_{n=-\infty}^{\infty} a_n < \infty.$$

Among these positive definite functions, the characteristic kernels are given by the ones with coefficients $a_0 \geq 0$ and $a_n > 0$ $(n \neq 0)$. The examples of characteristic kernels are $k_1(x, y) = (\pi - (x-y)_{mod\, 2\pi})^2$. $(a_0 = \pi^2/3, a_n = 2/n^2\ (n \neq 0))$, and $k_2(x, y) = 1/(1 - 2\alpha\cos(x-y) + \alpha^2)$ (Poisson kernel) given by $a_n = \alpha^{|n|}$ $(\alpha \in (0, 1))$. Examples of *non*-characteristic kernels on $[0, 2\pi)$ include $\cos(x-y)$, Féjer, and Dirichlet kernel.

The above conditions of characteristic properties can be extended in part to the case of compact groups using the unitary representations [5].

Acknowledgement

This work has been supported in part by JSPS KAKENHI (B) 22300098.

References

[1] C.R. Baker. Joint Measures and Cross-Covariance Operators. *Trans. Amer. Mathe. Soc.* 186, 273–289, 1973

[2] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic Publisher, 2004.

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

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

[5] K. Fukumizu, B.K. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic Kernels on Groups and Semigroups. Advances in Neural Information Processing Systems 21, 473–480, MIT Press, 2009.

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

[7] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems 19*. MIT Press, 2007.

[8] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems 20*, 585–592. MIT Press, 2008.

[9] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A Fast, Consistent Kernel Two-Sample Test. *Advances in Neural Information Processing Systems 22*, 673–681. MIT Press, 2009.

[10] A. Gretton, K. Fukumizu and B. Sriperumbudur. Discussion of: Brownian distance covariance. *Annals of Applied Statistics* 3(4), 1285–1294. 2009.

[11] W. Rudin. *Fourier Analysis on Groups*. Interscience, 1962.

[12] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press. 2002.

[13] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Proc. COLT 2008*, 111–122, 2008.

[14] X. Sun, D. Janzing, B. Schölkopf, and K. Fukumizu. A Kernel-based causal learning algorithm", *Proc. 24th Intern. Conf. Machine Learning (ICML2007)*, 855–862, 2007.

[15] van der Vaart, A.W. *Asymptotic Statistics*. Cambridge University Press, 1998.

Kenji Fukumzu
The Institute of Statistical Mathematics. 10-3 Midoricho, Tachikawa, Tokyo 190-8562 Japan, +81-50-5533-8540, fukumizu@ism.ac.jp