Theory on Positive Definite Kernels

Statistical Data Analysis with Positive Definite Kernels

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS
Department of Statistical Science, Graduate University for Advanced Studies

October 6-10, 2008, Kyushu University

Outline

Review: operations that preserve positive definiteness I

Proposition 1

If $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ $(i = 1, 2, \ldots)$ are positive definite kernels, then so are the following:

1. *(positive combination)* $ak_1 + bk_2 \quad (a, b \geq 0)$.
2. *(product)* $k_1 k_2 \quad (k_1(x, y)k_2(x, y))$.
3. *(limit)* $lim_{i\to\infty} k_i(x, y)$, *assuming the limit exists.*

Remark. Proposition 1 says that the set of all positive definite kernels is closed (w.r.t. pointwise convergence) convex cone stable under multiplication.

Example: If $k(x, y)$ is positive definite,

$$e^{k(x,y)} = 1 + k + \frac{1}{2}k^2 + \frac{1}{3!}k^3 + \cdots$$

is also positive definite.

Review: operations that preserve positive definiteness II

Proposition 2

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ be a positive definite kernel and $f : \mathcal{X} \to \mathbb{C}$ be an arbitrary function. Then,

$$\tilde{k}(x, y) = f(x)k(x, y)\overline{f(y)}$$

is positive definite. In particular,

$$f(x)\overline{f(y)}$$

is a positive definite kernel.

Example. Normalization:

$$\tilde{k}(x, y) = \frac{k(x, y)}{\sqrt{k(x, x)k(y, y)}}$$

Negative definite kernel

Definition. A function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is called a negative definite kernel if it is Hermitian i.e. $\psi(y, x) = \overline{\psi(x, y)}$, and

$$\sum_{i,j=1}^{n} c_i \overline{c_j} \psi(x_i, x_j) \leq 0$$

for any $x_1, \ldots, x_n$ ($n \geq 2$) in $\mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{C}$ with $\sum_{i=1}^{n} c_i = 0$.

Note: a negative definite kernel is not necessarily minus pos. def. kernel because of the condition $\sum_{i=1}^{n} c_i = 0$.

Properties of negative definite kernels

Proposition 3

1. *If k is positive definite, $\psi = -k$ is negative definite.*
2. *Constant functions are negative definite.*

(2) $\sum_{i,j=1}^{n} c_i c_j = \sum_{i=1}^{n} c_i \sum_{j=1}^{n} c_j = 0$.

Proposition 4

If $\psi_i : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ $(i = 1, 2, \ldots)$ are negative definite kernels, then so are the following:

1. *(positive combination)* $a\psi_1 + b\psi_2 \qquad (a, b \geq 0)$.
2. *(limit)* $lim_{i\to\infty}\psi_i(x, y)$, *assuming the limit exists.*

- The set of all negative definite kernels is a closed convex cone.
- Multiplication does not preserve negative definiteness.

Example of negative definite kernel

Proposition 5

Let V be an inner product space, and $\phi : \mathcal{X} \to V$. Then,

$$\psi(x, y) = \|\phi(x) - \phi(y)\|^2$$

is a negative definite kernel on $\mathcal{X}$.

Proof. Suppose $\sum_{i=1}^n c_i = 0$.

$$\begin{aligned}
&\textstyle\sum_{i,j=1}^n c_i \overline{c_j} \|\phi(x_i) - \phi(x_j)\|^2 \\
&= \textstyle\sum_{i,j=1}^n c_i \overline{c_j} \left\{ \|\phi(x_i)\|^2 + \|\phi(x_j)\|^2 - (\phi(x_i), \phi(x_j)) - (\phi(x_j), \phi(x_i)) \right\} \\
&= \textstyle\sum_{i=1}^n c_i \|\phi(x_i)\|^2 \sum_{j=1}^n \overline{c_j} + \sum_{j=1}^n c_j \|\phi(x_j)\|^2 \sum_{i=1}^n c_i \\
&\qquad - \left(\textstyle\sum_{i=1}^n c_i \phi(x_i), \sum_{j=1}^n c_j \phi(x_j) \right) - \left(\sum_{j=1}^n \overline{c_j} \phi(x_j), \sum_{i=1}^n \overline{c_i} \phi(x_i) \right) \\
&= -\left\| \textstyle\sum_{i=1}^n c_i \phi(x_i) \right\|^2 - \left\| \sum_{i=1}^n \overline{c_i} \phi(x_i) \right\|^2 \leq 0
\end{aligned}$$

Relation between positive and negative definite kernels

Lemma 6

Let $\psi(x,y)$ be a hermitian kernel on $\mathcal{X}$. Fix $x_0 \in \mathcal{X}$ and define

$$\varphi(x,y) = -\psi(x,y) + \psi(x,x_0) + \psi(x_0,y) - \psi(x_0,x_0).$$

Then, ψ is negative definite if and only if φ is positive definite.

Proof. "If" part is easy (exercise). Suppose ψ is neg. def. Take any $x_i \in \mathcal{X}$ and $c_i \in \mathbb{C}$ $(1 = 1, \ldots, n)$. Define $c_0 = -\sum_{i=1}^n c_i$. Then,

$$\begin{aligned}
0 &\geq \textstyle\sum_{i,j=0}^n c_i \overline{c_j} \psi(x_i, x_j) \qquad\qquad [\text{for } x_0, x_1, \ldots, x_n] \\
&= \textstyle\sum_{i,j=1}^n c_i \overline{c_j} \psi(x_i, x_j) + \overline{c_0} \sum_{i=1}^n c_i \psi(x_i, x_0) + c_0 \sum_{j=1}^n c_i \psi(x_0, x_j) \\
&\qquad + |c_0|^2 \psi(x_0, x_0) \\
&= \textstyle\sum_{i,j=1}^n c_i \overline{c_j} \{\psi(x_i, x_j) - \psi(x_i, x_0) - \psi(x_0, x_j) + \psi(x_0, y_0)\} \\
&= -\textstyle\sum_{i,j=1}^n c_i \overline{c_j} \varphi(x_i, x_j).
\end{aligned}$$

Schoenberg's theorem

Theorem 7 (Schoenberg's theorem)

Let $\mathcal{X}$ be a nonempty set, and $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ be a kernel.
ψ is negative definite if and only if $\exp(-t\psi)$ is positive definite for all $t > 0$.

Proof.
If part:

$$\psi(x, y) = \lim_{t \downarrow 0} \frac{1 - \exp(-t\psi(x, y))}{t}.$$

Only if part: We can prove only for $t = 1$. Take $x_0 \in \mathcal{X}$ and define

$$\varphi(x, y) = -\psi(x, y) + \psi(x, x_0) + \psi(x_0, y) - \psi(x_0, x_0).$$

φ is positive definite (Lemma 6).

$$e^{-\psi(x,y)} = e^{\varphi(x,y)} e^{-\psi(x,x_0)} \overline{e^{-\psi(y,x_0)}} e^{\psi(x_0,x_0)}.$$

This is also positive definite. □

Generating new kernels I

Proposition 8

If $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ *is negative definite and* $\psi(x, x) \geq 0$*. Then, for any* $0 < p \leq 1$*,*

$$\psi(x, y)^p$$

is negative definite.

Proof. Use the following formula.

$$\psi(x, y)^p = \frac{p}{\Gamma(1-p)} \int_0^\infty t^{-p-1}(1 - e^{-t\psi(x,y)})dt$$

The integrand is negative definite for all $t > 0$. □.

- For any $0 < p \leq 2$ and $\alpha > 0$,

  $$\exp(-\alpha \|x - y\|^p)$$

  is positive definite on $\mathbb{R}^n$.
- $\alpha = 2 \Rightarrow$ Gaussian kernel. $\alpha = 1 \Rightarrow$ Laplacian kernels.

Generating new kernels II

Proposition 9

If $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is negative definite and $\mathrm{Re}\psi(x,y) \geq 0$. Then, for any $a > 0$,

$$\frac{1}{\psi(x,y)+a}$$

is positive definite.

Proof.

$$\frac{1}{\psi(x,y)+a} = \int_0^\infty e^{-t(\psi(x,y)+a)}dt.$$

The integrand is positive definite for all $t > 0$. □.

For any $0 < p \leq 2$,

$$\frac{1}{1+|x-y|^p}$$

is positive definite on $\mathbb{R}$.

Positive definite functions

Definition. Let $\phi : \mathbb{R}^n \to \mathbb{C}$ be a function. ϕ is called a positive definite function (or function of positive type) if

$$k(x, y) = \phi(x - y)$$

is a positive definite kernel on $\mathbb{R}^n$, i.e.

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

for any $x_1, \ldots, x_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{C}$.

- A positive definite kernel of the form $\phi(x - y)$ is called shift invariant (or translation invariant).
- Gaussian and Laplacian kernels are examples of shift-invariant positive definite kernels.

Bochner's theorem I

The Bochner's theorem characterizes *all* the continuous shift-invariant kernels on $\mathbb{R}^n$.

Theorem 10 (Bochner)

Let ϕ be a continuous function on $\mathbb{R}^n$. Then, ϕ is positive definite if and only if there is a finite non-negative Borel measure Λ on R^n such that

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

- ϕ is the inverse Fourier (or Fourier-Stieltjes) transform of Λ.
- Roughly speaking, the shift invariant functions are the class that have non-negative Fourier transform.

Bochner's theorem II

- The Fourier kernel $e^{\sqrt{-1}x^T\omega}$ is a positive definite function for all $\omega \in \mathbb{R}^n$.

$$\exp(\sqrt{-1}(x-y)^T\omega) = \exp(\sqrt{-1}x^T\omega)\overline{\exp(\sqrt{-1}y^T\omega)}.$$

- The set of all positive definite functions is a convex cone, which is closed under the pointwise-convergence topology.

- The generator of the convex cone is the Fourier kernels $\{e^{\sqrt{-1}x^T\omega} \mid \omega \in \mathbb{R}^n\}$.

- Example on $\mathbb{R}$: (positive scales are neglected)

$$\exp(-\tfrac{1}{2\sigma^2}x^2) \qquad \exp(-\tfrac{\sigma^2}{2}|\omega|^2)$$

$$\exp(-\alpha|x|) \qquad \frac{1}{\omega^2+\alpha^2}$$

- Bochner's theorem is extended to topological groups and semigroups [BCR84].

Integral characterization of positive definite kernels I

Ω: compact Hausdorff space.
μ: finite Borel measure on Ω.

Proposition 11

Let $K(x,y)$ be a continuous symmetric function on $\Omega \times \Omega$. $K(x,y)$ is a positive definite kernel on Ω if and only if

$$\int_\Omega \int_\Omega K(x,y) f(x) \overline{f(y)} dx dy \geq 0$$

for each function $f \in L^2(\Omega, \mu)$.

c.f. Definition of positive definiteness:

$$\sum_{i,j} K(x_i, x_j) c_i \overline{c_j} \geq 0.$$

Integral characterization of positive definite kernels II

Proof.

(⇒). For a continuous function f, a Riemann sum satisfies

$$\sum_{i,j} K(x_i, x_j) f(x_i)\overline{f(x_j)}\mu(E_i)\mu(E_j) \geq 0.$$

The integral is the limit of such sums, thus non-negative. For $f \in L^2(\Omega, \mu)$, approximate it by a continuous function.

(⇐). Suppose

$$\sum_{i,j=1}^n c_i \overline{c_j} K(x_i, x_j) = -\delta < 0.$$

By continuity of K, there is an open neighborhood U_i of x_i such that

$$\sum_{i,j=1}^n c_i \overline{c_j} K(z_i, z_j) \leq -\delta/2.$$

for all $z_i \in U_i$.
We can approximate $\sum_i \frac{c_i}{\mu(U_i)} I_{U_i}$ by a continuous function f with arbitrary accuracy.

Integral Kernel

$(\Omega, \mathcal{B}, \mu)$: measure space.

$K(x, y)$: measurable function on $\Omega \times \Omega$ such that

$$\int_\Omega \int_\Omega |K(x,y)|^2 dxdy < \infty. \qquad \text{(square integrability)}$$

Define an operator T_K on $L^2(\Omega, \mu)$ by

$$(T_K f)(x) = \int_\Omega K(x,y) f(y) dy \qquad (f \in L^2(\Omega, \mu)).$$

T_K: integral operator with integral kernel K.

Fact: $T_K f \in L^2(\Omega, \mu)$.

$$\begin{aligned} \because) \quad \int |T_K f(x)|^2 dx &= \int \left\{ \int K(x,y) f(y) dy \right\}^2 dx \\ &\le \int \int |K(x,y)|^2 dy \int |f(y)|^2 dy dx \\ &= \int \int |K(x,y)|^2 dxdy \|f\|_{L^2}^2. \end{aligned}$$

Hilbert-Schmidt operator I

$\mathcal{H}$: separable Hilbert space.

Definition. An operator T on $\mathcal{H}$ is called Hilbert-Schmidt if for a CONS $\{\varphi_i\}_{i=1}^{\infty}$

$$\sum_{i=1}^{\infty} \|T\varphi_i\|^2 < \infty.$$

For a Hilbert-Schmidt operator T, the Hilbert-Schmidt norm $\|T\|_{HS}$ is defined by

$$\|T\|_{HS} = \left(\sum_{i=1}^{\infty} \|T\varphi_i\|^2\right)^{1/2}.$$

- $\|T\|_{HS}$ does not depend on the choice of a CONS.

  $\because$) From Parseval's equality, for a CONS $\{\psi_j\}_{j=1}^{\infty}$,

$$\|T\|_{HS}^2 = \sum_{i=1}^{\infty} \|T\varphi_i\|^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |(\psi_j, T\varphi_i)|^2$$
$$= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} |(T^*\psi_j, \varphi_i)|^2 = \sum_{j=1}^{\infty} \|T^*\psi_j\|^2.$$

Hilbert-Schmidt operator II

- Fact: $\|T\| \leq \|T\|_{HS}$.
- Hilbert-Schmidt norm is an extension of Frobenius norm of a matrix:

$$\|T\|_{HS}^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |(\psi_j, T\varphi_i)|^2.$$

$(\psi_j, T\varphi_i)$ is the component of the matrix expression of T with the CONS's $\{\varphi_i\}$ and $\{\psi_j\}$.

Hilbert-Schmidt operator and integral kernel I

Recall

$$(T_K f)(x) = \int_\Omega K(x,y)f(y)dy \qquad (f \in L^2(\Omega,\mu))$$

with square integrable kernel K.

Theorem 12

Assume $L^2(\Omega,\mu)$ is separable. Then, T_K is a Hilbert-Schmidt operator, and

$$\|T_K\|_{HS}^2 = \int\int |K(x,y)|^2 dxdy.$$

Proof. Let $\{\varphi_i\}$ be a CONS. From Parseval's equality,

$$\int |K(x,y)|^2 dy = \sum_i \left|(K(x,\cdot),\varphi_i)_{L^2}\right|^2 = \sum_i \left|\int K(x,y)\overline{\varphi_i(y)}dy\right|^2 = \sum_i |T_K\overline{\varphi_i}(x)|^2.$$

Integrate w.r.t. x, ($\{\overline{\varphi_i}\}$ is also a CONS)

$$\int\int |K(x,y)|^2 dxdy = \sum_i \|T_K\overline{\varphi_i}\|^2 = \|T_K\|_{HS}^2.$$

□

Hilbert-Schmidt operator and integral kernel II

Converse is true!

Theorem 13

Assume $L^2(\Omega, \mu)$ is separable. For any Hilbert-Schmidt operator T on $L^2(\Omega, \mu)$, there is a square integrable kernel $K(x, y)$ such that

$$T\varphi = \int K(x, y)\varphi(y)dy.$$

Outline of the proof.

Fix a CONS $\{\varphi_i\}$. Define

$$K_n(x, y) = \sum_{i=1}^{n} (T\varphi_i)(x)\overline{\varphi_i(y)} \qquad (n = 1, 2, 3, \ldots,).$$

We can show $\{K_n(x, y)\}$ is a Cauchy sequence in $L^2(\Omega \times \Omega, \mu \times \mu)$, and the limit works as K in the statement. □

Integral operator by positive definite kernel

Ω: compact Hausdorff space.
μ: finite Borel measure on Ω.

$K(x,y)$: continuous positive definite kernel on Ω.

$$(T_K f)(x) = \int_\Omega K(x,y)f(y)dy \qquad (f \in L^2(\Omega,\mu))$$

Fact: From Proposition 11

$$(T_K f, f)_{L^2(\Omega,\mu)} \geq 0 \qquad (\forall f \in L^2(\Omega,\mu)).$$

In particular, any eigenvalue of T_K is non-negative.

Mercer's theorem

$K(x, y)$: continuous positive definite kernel on Ω.

$\{\lambda_i\}_{i=1}^{\infty}$, $\{\varphi_i\}_{i=1}^{\infty}$: the positive eigenvalues and eigenfunctions of T_K.

$$\lambda_1 \geq \lambda_2 \geq \cdots > 0, \qquad \lim_{i \to \infty} \lambda_i = 0.$$

$$T_K \varphi_i = \lambda_i \varphi_i, \qquad \int K(x, y) \varphi_i(y) dy = \lambda_i \varphi_i(x).$$

Theorem 14 (Mercer)

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x) \overline{\varphi_i(y)},$$

where the convergence is absolute and uniform over $\Omega \times \Omega$.

Proof is omitted. See [RSN65], Section 98, or [Ito78], Chapter 13.

References I

[BCR84] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel.
Harmonic Analysis on Semigroups.
Springer-Verlag, 1984.

[Ito78] Seizo Ito.
Kansu-Kaiseki III (Iwanami kouza Kiso-suugaku).
Iwanami Shoten, 1978.

[RSN65] Frigyes Riesz and Béla Sz.-Nagy.
Functional Analysis (2nd ed.).
Frederick Ungar Publishing Co, 1965.