2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo
Pages 000–000

INFINITE DIMENSIONAL EXPONENTIAL FAMILIES BY REPRODUCING KERNEL HILBERT SPACES

KENJI FUKUMIZU

1. Introduction

The purpose of this paper is to propose a method of constructing exponential families of Hilbert manifold, on which estimation theory can be built. Although there have been works on infinite dimensional exponential families of Banach manifolds (Pistone and Sempi, 1995; Gibilisco and Pistone, 1998; Pistone and Rogantin, 1999), they are not appropriate to discuss statistical estimation with finite number of samples; the likelihood function with finite samples is not continuous on the manifold.

In this paper we use a reproducing kernel Hilbert space as a functional space for constructing an exponential manifold. A reproducing kernel Hilbert space is defined as a Hilbert space of functions such that evaluation of a function at an arbitrary point is a continuous functional on the Hilbert space. Since we can discuss the value of a function with this space, it is very natural to use a manifold associated with a reproducing kernel Hilbert space as a basis of estimation theory.

We focus on the maximum likelihood estimation (MLE) with the exponential manifold of a reproducing kernel Hilbert space. As in many non-parametric estimation methods, straightforward extension of MLE to an infinite dimensional exponential manifold suffers the problem of ill-posedness caused by the fact that the estimator should be chosen from the infinite dimensional space with only finite number of constraints given by the data. To solve this problem, a pseudo-maximum likelihood method is proposed by restricting the infinite dimensional manifold to a series of finite dimensional submanifolds, which enlarge as the number of samples increases. Some asymptotic results in the limit of infinite samples are shown, including the consistency of the pseudo-MLE.

2. Exponential family associated with a reproducing kernel Hilbert space

2.1. Reproducing kernel Hilbert space.

This subsection provides a brief review of reproducing kernel Hilbert spaces. For the details, see Aronszajn (1950).

Let Ω be a set, and $\mathcal{H}$ be a Hilbert space included in the set of all functions on Ω. The Hilbert space $\mathcal{H}$ is called *reproducing kernel Hilbert space* (RKHS) if the evaluation mapping $e_x : \mathcal{H} \ni f \mapsto f(x) \in \mathbb{R}$ is a continuous linear functional on $\mathcal{H}$ for any $x \in \Omega$.

A function $k : \Omega \times \Omega \to \mathbb{R}$ is said to be *positive definite* if it is symmetric and for any points $x_1, \ldots, x_n$ in Ω the matrix $(k(x_i, x_j))_{i,j}$ is positive semidefinite, i.e., for any real numbers $c_1, \ldots, c_n$ the inequality $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ holds.

If $\mathcal{H}$ is a RKHS on Ω, by Riesz's representation theorem, there exists $\phi_x \in \mathcal{H}$ such that $e_x(f) = f(x) = \langle f, \phi_x \rangle_{\mathcal{H}}$, where $\langle\ ,\ \rangle_{\mathcal{H}}$ is the inner product of $\mathcal{H}$. The function $k(\cdot, x) = \phi_x \in \mathcal{H}$ $(x \in \Omega)$ is called *reproducing kernel*, because it satisfies the reproducing property $\langle k(\cdot, x), f \rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$. It is easy to see that a symmetric function that satisfies the reproducing property is unique. The function $k(x, y)$ is a positive definite kernel, because it is symmetric from

$k(y,x) = \phi_x(y) = \langle \phi_x, \phi_y \rangle_{\mathcal{H}} = \langle \phi_y, \phi_x \rangle_{\mathcal{H}} = \phi_y(x) = k(x,y)$, and positive definite from $\sum_{i,j} c_i c_j k(x_i, x_j) = \| \sum_i c_i \phi_{x_i} \|_{\mathcal{H}}^2 \geq 0$.

It is known that for a positive definite kernel k on Ω there uniquely exists a Hilbert space $\mathcal{H}_k$ such that $\mathcal{H}_k$ consists of functions on Ω, the class of functions $\sum_{i=1}^m a_i k(\cdot, x_i)$ $(m \in \mathbb{N}, x_i \in \Omega, a_i \in \mathbb{R})$ is dense in $\mathcal{H}_k$, and $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$ holds for any $f \in \mathcal{H}_k$ and $x \in \Omega$. The last property means that $\mathcal{H}_k$ is a RKHS with a reproducing kernel k. Conversely, given RKHS $\mathcal{H}$ and its reproducing kernel $k(\cdot, x) = \phi_x$, the Hilbert space $\mathcal{H}_k$ constructed by k coincides with $\mathcal{H}$ because of the uniqueness of the reproducing kernel. Thus, a Hilbert space $\mathcal{H}$ of functions on Ω is a RKHS if and only if $\mathcal{H} = \mathcal{H}_k$ for some positive definite kernel k.

2.2. **Exponential manifold associated with a RKHS.** Let $(\Omega, \mathcal{B}, \mu)$ be a probability space, and $\mathcal{M}_\mu$ be the set of positive probability density functions with respect to μ;

$$\mathcal{M}_\mu = \left\{ f : \Omega \to \mathbb{R} \;\middle|\; f \text{ is measurable, almost everywhere positive, and } \int_\Omega f d\mu = 1 \right\}.$$

Hereafter, the probability given by the density $f \in \mathcal{M}$ is denoted by $f\mu$, and for a measurable function u on Ω the expectation of u with respect to $f\mu$ is denoted by $E_f[u(X)]$.

Let $k : \Omega \times \Omega \to \mathbb{R}$ be a measurable positive definite kernel on Ω. Define a subclass of $\mathcal{M}_\mu$ by

$$\mathcal{M}_\mu(k) = \left\{ f \in \mathcal{M}_\mu \;\middle|\; \text{there exists } \delta > 0 \text{ such that } \int e^{\delta \sqrt{k(x,x)}} f(x) d\mu(x) < \infty \right\}.$$

If the kernel k is bounded, we have $\mathcal{M}_\mu(k) = \mathcal{M}_\mu$.

In this paper, it is assumed that constant functions are included in $\mathcal{H}_k$. This is a mild assumption, because if $\tilde{k}(x,y) = k(x,y) + 1$ is used for a positive definite kernel, the corresponding RKHS $\mathcal{H}_{\tilde{k}}$ is equal to the sum $\mathcal{H}_k + \mathbb{R} = \{ f + c \mid f \in \mathcal{H}_k, c \in \mathbb{R} \}$ (Aronszajn, 1950).

Note $\|k(\cdot, x)\|_{\mathcal{H}_k} = \sqrt{k(x,x)}$ and $|u(x)| = |\langle u, k(\cdot, x) \rangle_{\mathcal{H}_k}| \leq \sqrt{k(x,x)} \|u\|_{\mathcal{H}_k}$ for an arbitrary $u \in \mathcal{H}_k$. For any $f \in \mathcal{M}_\mu(k)$, $E_f[\sqrt{k(X,X)}]$ is finite, because $\delta E_f[\sqrt{k(X,X)}] \leq E_f[e^{\delta \sqrt{k(X,X)}}] < \infty$. Thus, $u \mapsto E_f[u(X)]$ is a bounded functional on $\mathcal{H}_k$. We define a closed subspace T_f of $\mathcal{H}_k$ by

$$T_f := \{ u \in \mathcal{H}_k \mid E_f[u(X)] = 0 \},$$

which works as a tangent space at f, as we will see later.

For $f \in \mathcal{M}_\mu(k)$, let $\mathcal{W}_f$ be a subset of T_f defined by

$$\mathcal{W}_f = \{ u \in T_f \mid \text{there exists } \delta > 0 \text{ such that } E_f[e^{\delta \sqrt{k(X,X)} + u(X)}] < \infty \}.$$

Define the cumulant generating function Ψ_f on $\mathcal{W}_f$ by

$$\Psi_f(u) = \log E_f[e^{u(X)}].$$

For any $u \in \mathcal{W}_f$, the probability density function

$$e^{u - \Psi_f(u)} f$$

belongs to $\mathcal{M}_\mu(k)$. In fact, by the definition of $\mathcal{W}_f$, we can take $\delta > 0$ so that $E_f[e^{\delta \sqrt{k(X,X)} + u(X)}] < \infty$, which derives

$$\int e^{\delta \sqrt{k(x,x)}} e^{u(x) - \Psi_f(u)} f(x) d\mu(x) \leq e^{-\Psi_f(u)} E_f\left[e^{\delta \sqrt{k(X,X)} + u(X)} \right] < \infty.$$

Thus, the following mapping is well defined.

$$\xi_f : \mathcal{W}_f \to \mathcal{M}_\mu(k), \qquad u \mapsto e^{u - \Psi_f(u)} f.$$

The map ξ_f is one-to-one, because $\xi_f(u) = \xi_f(v)$ implies $u - v$ is constant, which is necessarily zero by $E_f[u] = E_f[v] = 0$. Let $\mathcal{S}_f = \xi_f(\mathcal{W}_f)$, and φ_f be the inverse of ξ_f, that is,

$$\varphi_f : \mathcal{S}_f \to \mathcal{W}_f, \qquad g \mapsto \log\frac{g}{f} - E_f\Big[\log\frac{g}{f}\Big].$$

It will be shown that φ_f works as a local coordinate of a Hilbert manifold $\mathcal{M}_\mu(k)$. First, we see the following basic facts;

Lemma 1. *Let f and g be arbitrary elements in $\mathcal{M}_\mu(k)$.*

(i) *$\mathcal{W}_f$ is an open subset of T_f.*

(ii) *$g \in \mathcal{S}_f$ if and only if $\mathcal{S}_g = \mathcal{S}_f$.*

Proof. (i). For an arbitrary $u \in \mathcal{W}_f$, take $\delta > 0$ so that $E_f[e^{u(X)+\delta\sqrt{k(X,X)}}] < \infty$. Define an open subset V_u of T_f by $V_u = \{v \in T_f \mid \|v - u\|_{\mathcal{H}_k} < \delta/2\}$. Then, V_u is an open neighborhood included in $\mathcal{W}_f$, because for arbitrary $v \in V_u$ we have

$$\begin{aligned} E_f\big[e^{(\delta/2)\sqrt{k(X,X)}+v(X)}\big] &= E_f\big[e^{(\delta/2)\sqrt{k(X,X)}+\langle v-u,k(\cdot,X)\rangle_{\mathcal{H}_k}+u(X)}\big] \\ &\le E_f\big[e^{(\delta/2)\sqrt{k(X,X)}+\|v-u\|_{\mathcal{H}_k}\sqrt{k(X,X)}+u(X)}\big] \\ &\le E_f\big[e^{\delta\sqrt{k(X,X)}+u(X)}\big] \quad < \infty. \end{aligned}$$

(ii). "If" part is obvious. For the "only if" part, we first prove $\mathcal{S}_g \subset \mathcal{S}_f$ on condition $g \in \mathcal{S}_f$. Let h be an arbitrary element in $\mathcal{S}_g$, and take $u \in \mathcal{W}_f$ and $v \in \mathcal{W}_g$ so that $g = e^{u-\Psi_f(u)}f$ and $h = e^{v-\Psi_g(v)}g$. From the fact $g \in \mathcal{W}_f$, there is $\delta > 0$ such that $E_g[e^{v(X)+\delta\sqrt{k(X,X)}}] < \infty$. We have $\int e^{v(x)+u(x)+\delta\sqrt{k(x,x)}-\Psi_f(u)}f(x)d\mu(x) < \infty$, which means $v + u - E_f[v] \in \mathcal{W}_f$. From $h = e^{(v+u-E_f[v])-(\Psi_f(u)+\Psi_g(v)-E_f[v])}f$, we have $\Psi_f(v+u-E_f[v]) = \Psi_f(u)+\Psi_g(v)-E_f[v]$ and $h = \xi_f(v+u-E_f[v]) \in \mathcal{S}_f$.

For the opposite inclusion, it suffices to show $f \in \mathcal{S}_g$. Let $\gamma > 0$ be a constant so that $E_f[e^{\gamma\sqrt{k(X,X)}}] < \infty$. From $e^{-u}g = e^{-\Psi_f(u)}f$, we see $\int e^{\gamma\sqrt{k(x,x)}-u(x)}g(x)d\mu(x) < \infty$, which means $-u + E_g[u] \in \mathcal{W}_g$. Thus, the equality $f = e^{-u+\Psi_f(u)}g = e^{(-u+E_g[u])-(-\Psi_f(u)+E_g[u])}g$ shows $f = \xi_g(-u + E_g[u]) \in \mathcal{S}_g$. □

The map φ_f defines a structure of Hilbert Manifold on $\mathcal{M}_\mu(k)$, which we call *reproducing kernel exponential manifold.*

Theorem 1. *The system $\{(\mathcal{S}_f, \varphi_f)\}_{f\in\mathcal{M}_\mu(k)}$ is a C^∞-atlas of $\mathcal{M}_\mu(k)$, that is,*

(i) *If $\mathcal{S}_f \cap \mathcal{S}_g \neq \emptyset$, then $\varphi_f(\mathcal{S}_f \cap \mathcal{S}_g)$ is an open set in T_f.*

(ii) *If $\mathcal{S}_f \cap \mathcal{S}_g \neq \emptyset$, then*

$$\varphi_g \circ \varphi_f^{-1}|_{\varphi_f(\mathcal{S}_f\cap\mathcal{S}_g)} : \varphi_f(\mathcal{S}_f \cap \mathcal{S}_g) \to \varphi_g(\mathcal{S}_f \cap \mathcal{S}_g)$$

is a C^∞ map.

Thus, $\mathcal{M}_\mu(k)$ admits a structure of C^∞-Hilbert manifold.

Proof. The assertion (i) is obvious, because $\mathcal{S}_f \cap \mathcal{S}_g \neq \emptyset$ means $\mathcal{S}_f = \mathcal{S}_g$ from Lemma 1. Suppose $\mathcal{S}_f \cap \mathcal{S}_g \neq \emptyset$, that is, $\mathcal{S}_f = \mathcal{S}_g$. For any $u \in \mathcal{W}_f$, we have

$$\begin{aligned} \varphi_g \circ \varphi_f^{-1}(u) = \varphi_g\big(e^{u-\Psi_f(u)}f\big) &= \log\frac{e^{u-\Psi_f(u)}f}{g} - E_g\Big[\log\frac{e^{u-\Psi_f(u)}f}{g}\Big] \\ &= u + \log(f/g) - E_g\big[u + \log(f/g)\big], \end{aligned}$$

from which the assertion (ii) is obtained, because $u \mapsto E_g[u]$ is of C^∞.

It is known that with the assertions (i) and (ii) a topology is introduced on $\mathcal{M}_\mu(k)$ so that all $\mathcal{S}_f$ are open, and $\mathcal{M}_\mu(k)$ is equipped with the structure of C^∞-Hilbert manifold (see Lang, 1985). □

The open set $\mathcal{S}_f$ is regarded as a local maximal exponential family in $\mathcal{M}_\mu(k)$. In fact, we can prove

$$\mathcal{S}_f = \{g \in \mathcal{M}_\mu(k) \mid \text{ there exists } u \in T_f \text{ such that } g = e^{u-\Psi_f(u)}f\}. \tag{1}$$

To see this, it suffices to show that the right hand side is included in the left hand side, as the opposite inclusion is obvious. Let $g = e^{u-\Psi_f(u)}f$ be in the set of the right hand side. From $g \in \mathcal{M}_\mu(k)$, there exists $\delta > 0$ such that $E_g[e^{\delta\sqrt{k(X,X)}}] < \infty$, which means $E_f[e^{\delta\sqrt{k(X,X)}+u(X)}] < \infty$. Therefore, $u \in \mathcal{W}_f$ and $g = \xi_f(u) \in \mathcal{S}_f$.

From Lemma 1 (ii), we can define an equivalence relation such that f and g are equivalent if and only if they are in the same local maximal exponential family, that is, if and only if $\mathcal{S}_f \cap \mathcal{S}_g \neq \emptyset$. Let $\{\mathcal{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ be the equivalence class. Then, they are equal to the set of connected components.

Theorem 2. *Let $\{\mathcal{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ be the equivalence class of the maximum local exponential families. Then, $\{\mathcal{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ are the connected components of $\mathcal{M}_\mu(k)$. Moreover, each component $\mathcal{S}^{(\lambda)}$ is simply connected.*

Proof. From Lemma 1 and Theorem 1, $\{\mathcal{S}^{(\lambda)}\}_{\lambda\in\Lambda}$ are disjoint open covering of $\mathcal{M}_\mu(k)$. The proof is completed if every $\mathcal{W}_f$ is shown to be convex. Let u_0 and u_1 be arbitrary elements in $\mathcal{W}_f$. Then, there exists $\delta > 0$ such that $E_f\big[e^{\delta\sqrt{k(X,X)}+u_0(X)}\big] < \infty$ and $E_f\big[e^{\delta\sqrt{k(X,X)}+u_1(X)}\big] < \infty$. For $u_t = tu_1 + (1-t)u_0 \in T_f$ $(t \in [0,1])$, we have $e^{u_t(x)} \le te^{u_1(x)} + (1-t)e^{u_0(x)}$ by the convexity of $z \mapsto e^z$. It leads

$$E_f\Big[e^{\delta\sqrt{k(X,X)}+u_t(X)}\Big] \le tE_f\Big[e^{\delta\sqrt{k(X,X)}+u_1(X)}\Big] + (1-t)E_f\Big[e^{\delta\sqrt{k(X,X)}+u_0(X)}\Big] < \infty,$$

which means $u_t \in \mathcal{W}_f$. □

The Hilbert space $\mathcal{H}_k$, which is used for giving manifold structure to $\mathcal{M}_\mu(k)$, has stronger topology than the Orlicz space used for the exponential manifold by Pistone and Sempi (1995). Recall that a function u is an element of the Orlicz space $L^{\cosh-1}(f)$ if and only if there is $\alpha > 0$ such that

$$E_f\Big[\cosh\Big(\frac{u}{\alpha}\Big) - 1\Big] < \infty.$$

The space $u \in L^{\cosh-1}(f)$ is a Banach space with the norm

$$\|u\|_{L^{\cosh-1}(f)} = \inf\Big\{\alpha > 0 \;\Big|\; E_f\Big[\cosh\Big(\frac{u}{\alpha}\Big) - 1\Big] \le 1\Big\}.$$

For the detail of this space, see Pistone and Sempi (1995).

Proposition 1. *For any $f \in \mathcal{M}_\mu(k)$, the RKHS $\mathcal{H}_k$ is continuously included in $L^{\cosh-1}(f)$. Moreover, if a positive number A_f is defined by*

$$A_f = \inf\Big\{\alpha > 0 \;\Big|\; \int e^{\frac{\sqrt{k(x,x)}}{\alpha}} f(x)d\mu(x) \le 2\Big\},$$

then for any $u \in \mathcal{H}_k$

$$\|u\|_{L^{\cosh-1}(f)} \le A_f\|u\|_{\mathcal{H}_k}.$$

Proof. The inequality

$$\begin{aligned} E_f\big[\cosh(u(X)/\alpha) - 1\big] &= \frac{1}{2}E_f[e^{u(X)/\alpha} + e^{-u(X)/\alpha}] - 1 \\ &\le E_f\big[e^{|u(X)|/\alpha}\big] - 1 \\ &\le E_f\Big[e^{\frac{1}{\alpha}\|u\|_{\mathcal{H}_k}\sqrt{k(X,X)}}\Big] - 1 \end{aligned}$$

shows that if $\|u\|_{\mathcal{H}_k}/\alpha < 1/A_f$ then $E_f[\cosh(u/\alpha) - 1] \le 1$. Thus, $A_f\|u\|_{\mathcal{H}_k} \ge \|u\|_{L^{\cosh-1}(f)}$. □

Proposition 1 tells that the manifold $\mathcal{M}_\mu(k)$ is a subset of the maximum exponential manifold. However, the former is not necessarily a submanifold of the latter, because $\mathcal{H}_k$ is not a closed subspace of $L^{\cosh -1}(f)$ in general.

Note also that $L^{\cosh -1}(f)$ is continuously embedded in $L^p(f)$ for all $p \geq 1$. Thus, $E_f|u|^p$ is finite for any $f \in \mathcal{M}_\mu(k)$, $u \in \mathcal{H}_k$, and $p \geq 1$.

The reproducing kernel exponential manifold depends on the underlying RKHS. It may be either finite or infinite dimensional. A different choice of the positive definite kernel results in a different exponential manifold.

2.3. **Properties of reproducing kernel exponential manifolds.** As in the case of finite dimensional exponential families and the exponential manifold by Pistone and Sempi (1995), the derivatives of cumulant generating function provide the cumulant or moments of the random variables given by tangent vectors. Let $f \in \mathcal{M}_\mu(k)$ and $v_1, \ldots, v_d \in T_f$. The d-th derivative of Ψ_f in the directions $v_1, \ldots, v_d$ at $f_u = e^{u-\Psi_f(u)}f$ is denoted by $D_u^d \Psi_f(v_1, \ldots, v_d)$. We have

$$D_u\Psi_f(v) = E_{f_u}[v], \qquad D_u^2\Psi_f(v_1, v_2) = \mathrm{Cov}_{f_u}[v_1(X), v_2(X)],$$

where $\mathrm{Cov}_g[v_1, v_2] = E_g[v_1(X)v_2(X)] - E_g[v_1(X)]E_g[v_2(X)]$ is the covariance of v_1 and v_2 under the probability $g\mu$.

The first and second moments are expressed also by an element and an operator of the Hilbert space. Let P be a probability on Ω such that $E_P[\sqrt{k(X,X)}] < \infty$. Because the functional $\mathcal{H}_k \ni u \mapsto E_P[u(X)]$ is bounded, there exists $m_P \in \mathcal{H}_k$ such that

$$E_P[u(X)] = \langle u, m_P \rangle_{\mathcal{H}_k}$$

for all $u \in \mathcal{H}_k$. We call m_P *mean element* for P. Noticing that the mapping $\mathcal{H}_k \times \mathcal{H}_k \ni (v_1, v_2) \mapsto \mathrm{Cov}_P[v_1(X), v_2(X)]$ is a bounded bilinear form, we see that there exists a bounded operator Σ_P on $\mathcal{H}_k$ such that

$$\mathrm{Cov}_P[v_1(X), v_2(X)] = \langle v_1, \Sigma_P v_2 \rangle_{\mathcal{H}_k}$$

holds for all $v_1, v_2 \in \mathcal{H}_k$. The operator Σ_P is called *covariance operator* for P. For the detail of covariance operator on a RKHS, see (Fukumizu et al., 2005).

When a local coordinate $(\varphi_{f_0}, \mathcal{S}_{f_0})$ in a reproducing kernel exponential manifold $\mathcal{M}_\mu(k)$ is assumed, we use also the notations m_u and Σ_u for the mean element and covariance operator, respectively, with respect to the probability density $f_u = e^{u-\Psi_{f_0}(u)}f_0$. The mapping $\mathcal{W}_f \ni u \mapsto m_u \in \mathcal{H}_k$ is locally one-to-one, because the derivative $\Sigma_u|_{T_{f_0}}$ is injective given that μ is non-degenerate. We call m_u *mean parameter* for the density f_u. We have

$$D_u\Psi_f(v) = \langle m_u, v \rangle_{\mathcal{H}_k}, \qquad D_u^2\Psi_f(v_1, v_2) = \langle v_1, \Sigma_u v_2 \rangle_{\mathcal{H}_k}.$$

Let $f_0 \in \mathcal{M}_\mu(k)$ and $u, v \in \mathcal{W}_{f_0}$. With the local coordinate $(\varphi_{f_0}, \mathcal{S}_{f_0})$, it is easy to see that the Kullback-Leibler divergence from $f_u = e^{u-\Psi_{f_0}(u)}f_0$ to $f_v = e^{v-\Psi_{f_0}(v)}f_0$ is given by

$$KL(f_u||f_v) = \Psi_{f_0}(v) - \Psi_{f_0}(u) - \langle v - u, m_u \rangle_{\mathcal{H}_k}. \tag{2}$$

Let f_u, f_v, and f_w be points in $\mathcal{S}_{f_0}$. It is straightforward to see

$$KL(f_u||f_w) = KL(f_u||f_v) + KL(f_v||f_w) - \langle w - v, m_u - m_v \rangle_{\mathcal{H}_k}.$$

Let U be a closed subspace of T_{f_0}, and $\mathcal{V} = U \cap \mathcal{W}_{f_0}$. The subset $\mathcal{N} = \varphi_{f_0}^{-1}(\mathcal{V})$ is a submanifold of $\mathcal{S}_{f_0}$, which is also an exponential family. Let $f_* = e^{u_* - \Psi_{f_0}(u_*)}$ be a point in $\mathcal{S}_{f_0}$, and consider the minimizer of the KL divergence from f_* to a point in $\mathcal{N}$;

$$u_{opt} = \arg\min_{u \in \mathcal{V}} KL(f_*||f_u).$$

Then, the standard argument derives the orthogonal relation

$$\langle u - u_{opt}, m_{u_*} - m_{u_{opt}} \rangle_{\mathcal{H}_k} = 0. \tag{3}$$

Using the above relations, the Pythagorean equation

$$KL(f_* || f_u) = KL(f_* || f_{u_{opt}}) + KL(f_{u_{opt}} || f_u) \tag{4}$$

holds for any $u \in \mathcal{V}$.

3. Pseudo maximum likelihood estimation with $\mathcal{M}_\mu(k)$

3.1. **Likelihood equation on a reproducing kernel exponential manifold.** Let $\mathcal{S}$ be a connected component of $\mathcal{M}_\mu(k)$. We assume that the true probability density function f_* is an element of $\mathcal{S}$, and i.i.d. samples $X_1, X_2, \ldots, X_n$ are generated by $f_*\mu$. We consider the problem of estimating f_* with the statistical model $\mathcal{S}$ given the finite sample.

From Lemma 1 and Eq. (1), for arbitrary $f_0 \in \mathcal{S}$ the component can be expressed by

$$\mathcal{S} = \{ f \in \mathcal{M}_\mu(k) \mid f = e^{u - \Psi_0(u)} f_0 \text{ for some } u \in T_{f_0} \},$$

where Ψ_0 is an abbreviation of Ψ_{f_0}. For notational simplicity, we use $\mathcal{W}_0 = \mathcal{W}_{f_0}$ and $f_u = e^{u-\Psi_0(u)} f_0$ for $u \in \mathcal{W}_0$.

Consider the maximum likelihood estimation (MLE) with $\mathcal{S}$;

$$\sup_{u \in \mathcal{W}_0} \frac{1}{n} \sum_{i=1}^n u(X_i) - \Psi_0(u).$$

By introducing the empirical mean element

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

the problem of MLE is rewritten by

$$\sup_{u \in \mathcal{W}_0} \langle \widehat{m}^{(n)}, u \rangle_{\mathcal{H}_k} - \Psi_0(u).$$

Taking the partial derivative of the log likelihood function, we have the following likelihood equation;

$$\langle \widehat{m}^{(n)}, v \rangle_{\mathcal{H}_k} = \langle m_u, v \rangle_{\mathcal{H}_k} \qquad (\forall v \in \mathcal{H}_k), \tag{5}$$

where m_u is the mean parameter corresponding to the density f_u. Note that Eq. (5) holds not only for $v \in T_{f_0}$ but for all $v \in \mathcal{H}_k$, since $\langle \widehat{m}^{(n)} - m_u, 1 \rangle_{\mathcal{H}_k}$ always vanishes.

From Eq. (5), the empirical mean element $\widehat{m}^{(n)}$ is regarded as the mean parameter for the maximum likelihood estimation. We call $\widehat{m}^{(n)}$ maximum likelihood mean parameter.

3.2. **$\sqrt{n}$-consistency of maximum likelihood mean parameter.** We establish the $\sqrt{n}$-consistency of the maximal likelihood mean parameter in a general form.

Theorem 3. *Let $(\Omega, \mathcal{B}, P)$ be a probability space, $k : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel so that $E_P[k(X,X)] < \infty$, and $m_P \in \mathcal{H}_k$ be the mean element with respect to P. Suppose $X_1, \ldots, X_n$ are i.i.d. sample from P, and define the empirical mean element $\widehat{m}^{(n)}$ by $\widehat{m}^{(n)} = \frac{1}{n}\sum_{i=1}^n k(\cdot, X_i)$. Then, we have*

$$\|\widehat{m}^{(n)} - m_P\|_{\mathcal{H}_k} = O_p\big(1/\sqrt{n}\big) \quad (n \to \infty).$$

Proof. Let $E_X[\cdot]$ denote the expectation with respect to the random variable X which follows P. Suppose $X, \tilde{X}, X_1, \ldots, X_n$ are i.i.d. We have

$$\begin{aligned} E\|\widehat{m}^{(n)} - m_P\|^2_{\mathcal{H}_k} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E_{X_i}E_{X_j}[k(X_i, X_j)] \\ &\quad - \frac{2}{n}\sum_{i=1}^{n} E_{X_i}E_X[k(X_i, X)] + E_X E_{\tilde{X}}[k(X, \tilde{X})] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} E[k(X_i, X_j)] + \frac{1}{n}E_X[k(X, X)] - E_X E_{\tilde{X}}[k(X, \tilde{X})] \\ &= \frac{1}{n}\{E_X[k(X, X)] - E_X E_{\tilde{X}}[k(X, \tilde{X})]\} \\ &= O(1/n). \end{aligned}$$

The assertion is obtained by Chebyshev's inequality. □

Unlike the finite dimensional exponential family, the likelihood equation Eq. (5) does not necessarily have a solution in the canonical parameter u. The empirical expectation element $\widehat{m}^{(n)}$ works as a mean parameter as in the finite dimensional case. However, as Pistone and Rogantin (1999) point out for their exponential manifold, the inverse mapping from the mean parameter to the canonical parameter u is not bounded in general. In fact, the derivative of the map $u \mapsto m_u$ is the covariance operator Σ_u, which has infinitely small eigenvalues under mild conditions (Fukumizu et al., 2005), if the RKHS is infinite dimensional. Thus, the mean parameter does not give a coordinate system for infinite dimensional manifolds.

Another explanation for the fact that the likelihood equation does not have a solution is given by regarding the equation as moment matching. Given that the positive definite kernel k is continuous, the likelihood equation requires that the empirical distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ and the continuous density function $e^{u-\Psi_0(u)}f_0$ have the same expectation or moment for all functions $w \in \mathcal{H}_k$. If the RKHS $\mathcal{H}_k$ is so large that almost all the moments can be evaluated by the form $E[w(X)]$ for $w \in \mathcal{H}_k$, the probability must be uniquely determined by those moments. Thus, $\mathcal{S}$ cannot include a probability which gives the same moment structure as the empirical distribution.

3.3. **Pseudo maximum likelihood estimation.**

To solve the problem described in the above subsection, we propose a pseudo maximum likelihood estimation using a series of finite dimensional subspaces in $\mathcal{H}_k$ to make the inversion from the mean parameter to the canonical parameter possible. The use of finite dimensional subspaces is a kind of regularization. With an infinite dimensional reproducing kernel exponential manifold, the estimation of the true density with finite number of samples can be considered to be an ill-posed problem, because it attempts to find a function from the infinite dimensional space with only finite number of constraints made by the samples. The regularization, which reduces the degree of freedom, is a popular method of solving the ill-posed problem.

Let $\{\mathcal{H}^{(\ell)}\}_{\ell=1}^{\infty}$ be a series of subspaces of $\mathcal{H}_k$ such that $\mathcal{H}^{(\ell)} \subset \mathcal{H}^{(\ell+1)}$ for all $\ell \in \mathbb{N}$. For any $f \in \mathcal{M}_\mu(k)$, a subspace $T_f^{(\ell)}$ of T_f is defined by $T_f^{(\ell)} = T_f \cap \mathcal{H}^{(\ell)}$, and an open set $\mathcal{W}_f^{(\ell)}$ of $T_f^{(\ell)}$ is defined by $\mathcal{W}_f \cap \mathcal{H}^{(\ell)}$. Also, the notations $\mathcal{W}^{(\ell)}$ and $\mathcal{S}^{(\ell)}$ are used for $\mathcal{W}_{f_0}^{(\ell)}$ and $\{f_u \in \mathcal{S} \mid u \in \mathcal{W}^{(\ell)}\}$, respectively.

For each $\ell \in \mathbb{N}$, the pseudo maximum likelihood estimator $\widehat{u}^{(\ell)}$ in $\mathcal{S}^{(\ell)}$ is defined by

$$\widehat{u}^{(\ell)} = \arg\max_{u\in\mathcal{W}^{(\ell)}} \langle \widehat{m}^{(n)}, u\rangle_{\mathcal{H}_k} - \Psi_0(u).$$

We assume that the maximizer $\widehat{u}^{(\ell)}$ exists in $\mathcal{W}^{(\ell)}$, and further make the following two assumptions;

(A-1) For all $u \in \mathcal{W}_0$, let $u_*^{(\ell)} \in \mathcal{W}^{(\ell)}$ ($\ell \in \mathbb{N}$) be the minimizer of

$$\min_{u^{(\ell)} \in \mathcal{W}^{(\ell)}} KL(f_u || f_{u^{(\ell)}}).$$

Then $\|u - u_*^{(\ell)}\|_{\mathcal{H}_k} \to 0$ ($\ell \to \infty$).

(A-2) For $u \in \mathcal{W}_0$, let $\lambda^{(\ell)}(u)$ be the least eigenvalue of the covariance operator Σ_u restricted on $T_{f_u}^{(\ell)}$, that is,

$$\lambda^{(\ell)}(u) := \inf_{v \in T_{f_u}^{(\ell)},\, \|v\|_{\mathcal{H}_k}=1} \langle v, \Sigma_u v\rangle_{\mathcal{H}_k}.$$

Then, there exists a subsequence $(\ell_n)_{n=1}^\infty$ of $\mathbb{N}$ such that for all $u \in \mathcal{W}_0$ we can find $\delta > 0$ for which

$$\tilde{\lambda}_u^{(\ell)} := \inf_{u' \in \mathcal{W}_0,\, \|u'-u\|_{\mathcal{H}_k} \le \delta} \lambda^{(\ell)}(u')$$

satisfies

$$\lim_{n\to\infty} \sqrt{n}\tilde{\lambda}_u^{(\ell_n)} = +\infty.$$

The assumption (A-1) means $\mathcal{S}^{(\ell)}$ can approximate a function in $\mathcal{S}$ at any precision as ℓ goes to infinity. The assumption (A-2) provides a stable MLE in the submodel $\mathcal{S}^{(\ell)}$ by lower-bounding the lest eigenvalue of the derivative of the map $u \mapsto m_u$.

Theorem 4. *Under the assumptions (A-1) and (A-2),*

$$KL(f_* || f_{\widehat{u}^{(\ell_n)}}) \to 0 \qquad (n \to \infty)$$

in probability.

Moreover, let $u_* \in \mathcal{W}_0$ be the element which gives $f_{u_*} = f_*$, and $u_*^{(\ell)}$ be the element in (A-1) with respect to u_*. If positive constants γ_n and ε_n satisfy

$$\|u_* - u_*^{(\ell)}\|_{\mathcal{H}_k} = o(\gamma_n) \qquad (n \to \infty)$$

and

$$\frac{1}{\sqrt{n}\tilde{\lambda}_{u_*}^{(\ell_n)}} = o(\varepsilon_n) \qquad (n \to \infty),$$

then we have

$$KL(f_* || f_{\widehat{u}^{(\ell_n)}}) = o_p(\max\{\gamma_n, \varepsilon_n\}) \qquad (n \to \infty).$$

Proof. Let m_* and $m_*^{(\ell)}$ be the mean parameter corresponding to u_* and $u_*^{(\ell)}$, respectively. From Eqs. (3) and (4), we have

$$\langle u - u_*^{(\ell)}, m_*^{(\ell)}\rangle_{\mathcal{H}_k} = \langle u - u_*^{(\ell)}, m_*\rangle_{\mathcal{H}_k} \tag{6}$$

for all $u \in \mathcal{W}^{(\ell)}$, and

$$KL(f_* || f_{\widehat{u}^{(\ell_n)}}) = KL(f_* || f_{u_*^{(\ell_n)}}) + KL(f_{u_*^{(\ell_n)}} || f_{\widehat{u}^{(\ell_n)}}).$$

The assumption (A-1) means the convergence

$$KL(f_* || f_{u_*^{(\ell_n)}}) \to 0 \qquad (n \to \infty),$$

since $KL(f_* || f_u)$ is an continuous function on u. Thus, it suffices to show the convergence

$$\Pr\big(\|\widehat{u}^{(\ell_n)} - u_*^{(\ell_n)}\|_{\mathcal{H}_k} \ge \varepsilon_n\big) \to 0, \tag{7}$$

because we have $KL(f_{u_*^{(\ell_n)}}||f_{\widehat{u}^{(\ell_n)}}) = \Psi_0(\widehat{u}^{(\ell_n)}) - \Psi_0(u_*^{(\ell_n)}) - \langle m_*, \widehat{u}^{(\ell_n)} - u_*^{(\ell_n)}\rangle_{\mathcal{H}_k}$ from Eqs. (2) and (6), which implies $KL(f_{u_*^{(\ell_n)}}||f_{\widehat{u}^{(\ell_n)}})$ is also of $o_p(\varepsilon_n)$.

Let $\delta > 0$ be the constant in the assumption (A-2). If the event of the probability in Eq. (7) holds, we have

$$\sup_{\substack{u\in\mathcal{W}^{(\ell)} \\ \|u-u_*^{(\ell)}\|_{\mathcal{H}_k}\geq\varepsilon_n}} \big\{\langle u, \widehat{m}^{(n)}\rangle_{\mathcal{H}_k} - \Psi_0(u)\big\} - \big\{\langle u_*^{(\ell_n)}, \widehat{m}^{(n)}\rangle_{\mathcal{H}_k} - \Psi_0(u_*^{(\ell_n)})\big\} \geq 0.$$

On the other hand, using Eq. (6), the relation

$$\begin{aligned}&\langle u, \widehat{m}^{(n)}\rangle_{\mathcal{H}_k} - \Psi_0(u) - \langle u_*^{(\ell_n)}, \widehat{m}^{(n)}\rangle_{\mathcal{H}_k} + \Psi_0(u_*^{(\ell_n)}) \\ &= \langle u - u_*^{(\ell_n)}, \widehat{m}^{(n)} - m_*\rangle_{\mathcal{H}_k} - \big\{\Psi_0(u) - \Psi_0(u_*^{(\ell_n)}) - \langle u - u_*^{(\ell_n)}, m_*^{(\ell_n)}\rangle_{\mathcal{H}_k}\big\}\end{aligned}$$

is obtained for any $u \in \mathcal{W}^{(\ell)}$. By the definition of $\tilde{\lambda}^{(\ell)}$, for sufficiently large n so that $\|u_*^{(\ell_n)} - u_*\|_{\mathcal{H}_k} \leq \delta$ holds, we obtain

$$\begin{aligned}&\sup_{\substack{u\in\mathcal{W}^{(\ell)} \\ \|u-u_*^{(\ell)}\|_{\mathcal{H}_k}\geq\varepsilon_n}} \langle u, \widehat{m}^{(n)}\rangle_{\mathcal{H}_k} - \Psi_0(u) - \langle u_*^{(\ell_n)}, \widehat{m}^{(n)}\rangle_{\mathcal{H}_k} + \Psi_0(u_*^{(\ell_n)}) \\ \leq &\sup_{\substack{u\in\mathcal{W}^{(\ell)} \\ \|u-u_*^{(\ell)}\|_{\mathcal{H}_k}\geq\varepsilon_n}} \|u - u_*^{(\ell_n)}\|_{\mathcal{H}_k}\|\widehat{m}^{(n)} - m_*\|_{\mathcal{H}_k} - \frac{1}{2}\tilde{\lambda}^{(\ell_n)}\|u - u_*^{(\ell_n)}\|_{\mathcal{H}_k}^2 \\ \leq &\sup_{\substack{u\in\mathcal{W}^{(\ell)} \\ \|u-u_*^{(\ell)}\|_{\mathcal{H}_k}\geq\varepsilon_n}} \|u - u_*^{(\ell)}\|_{\mathcal{H}_k}\Big\{\|\widehat{m}^{(n)} - m_*\|_{\mathcal{H}_k} - \frac{1}{2}\tilde{\lambda}^{(\ell_n)}\varepsilon_n\Big\}.\end{aligned}$$

Thus, the probability in Eq. (7) is upper bounded by

$$\Pr\big(\|\widehat{m}^{(n)} - m_*\|_{\mathcal{H}_k} \geq \tfrac{1}{2}\tilde{\lambda}^{(\ell_n)}\varepsilon_n\big),$$

which converges to zero by Theorem 3 and the condition of ε_n. □

4. Concluding Remarks

This paper has proposed a new family of statistical models, reproducing kernel exponential manifold, which includes infinite dimensional exponential models. The most significant property of this exponential manifold is that the empirical mean parameter is included in the Hilbert space. Thus, estimation of the density function with finite samples can be discussed based on this exponential manifold, while many other formulation of exponential manifold cannot provide basis for estimation with finite samples. Using the reproducing kernel exponential manifold, a method of pseudo maximum likelihood estimation has been proposed with a series of finite dimensional submanifolds, and consistency of the estimator has been shown.

As this paper is the first proposal of estimation theory based on infinite dimensional exponential manifolds, many problems remain unsolved. One of them is a practical method for constructing a sequence of subspaces used for the pseudo maximum likelihood estimation. A possible way of defining the sequence is to use the subspace spanned by $k(\cdot, X_1), \ldots, k(\cdot, X_\ell)$. However, with this construction the subspaces are also random, and the results in this paper should be extended to the case of random subspaces to guarantee the consistency. Another practical issue is how to choose the subsequence ℓ_n so that the assumption (A-2) is satisfied. We need to elucidate the properties of least eigenvalue of the covariance operator restricted on finite dimensional subspaces, which is not necessarily obvious. Also, providing examples of the estimator for specific kernels is practically important. Investigation of these problems will be among our future works.

Acknowledgements

This work has been partially supported by JSPS KAKENHI 15700241, research grant of Inamori Foundation, and Scientific Grant of Mitsubishi Foundation.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 69(3): 337–404, 1950.

K. Fukumizu, F. R. Bach, and A. Gretton. Consistency of kernel canonical correlation analysis. Research Memorandum 942, Institute of Statistical Mathematics, 2005.

P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by orlicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 1(2):325–347, 1998.

S. Lang. *Differential Manifolds*. Springer-Verlag, 1985.

G. Pistone and M.-P. Rogantin. The exponential statistical manifold: Mean parameters, orthogonality, and space transformation. *Bernoulli*, 5:721–760, 1999.

G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, 23(5):1543–1561, 1995.

Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, JAPAN

E-mail address: `fukumizu@ism.ac.jp`

URL: `http://www.ism.ac.jp/~fukumizu/`