Reproducing Kernel Exponential Manifold: Estimation and Geometry

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS

Graduate University of Advanced Studies

Mathematical Explorations in Contemporary Statistics

May 19-20, 2008. Sestri Levante, Italy

Outline

- Introduction
- Reproducing kernel exponential manifold (RKEM)
- Statistical asymptotic theory of singular models
- Concluding remarks

Introduction

Maximal Exponential Manifold

- Maximal exponential manifold (P&S'95)
  - A Banach manifold is defined so that the cumulant generating function is well-defined on a neighborhood of each probability density.

$$f_u = \exp(u - \Psi_f(u)) f, \qquad \Psi_f(u) = \log E_f[e^u] < \infty$$

  - Orlicz space $L_{\cosh\text{-}1}(f) = \left\{ u \mid \exists \alpha > 0 \text{ s.t. } E_f[e^{\alpha u}] < \infty \text{ and } E_f[e^{-\alpha u}] < \infty \right\}$

    This space is (perhaps) the most general to guarantee the finiteness of the cumulant generating functions around a point.

Estimation with Data

- Estimation with a finite sample
  - A finite dimensional exponential family is suitable for the maximum likelihood estimation (MLE) with a finite sample.

$X_1, \ldots, X_n$: i.i.d. $\sim \; f_0 \mu$ $\qquad\qquad \mathrm{X}_n = (X_1, \ldots, X_n)$

MLE: θ that maximizes

$$\ell_n(\theta; \mathrm{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{a=1}^{m} \theta^a u_a(X_i) - \Psi(\theta) \right\}$$

  - Is MLE extendable to the maximal exponential manifold?

$$\ell_n(u; \mathrm{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ u(X_i) - \Psi_f(u) \right\}$$

⇨ But, the function value $u(X_i)$ is not a continuous functional on u in the exponential manifold.

Reproducing kernel exponential manifold

Reproducing Kernel Hilbert Space

- Reproducing kernel Hilbert space (RKHS)
  - Ω: set. A Hilbert space $\mathcal{H}$ consisting of functions on Ω is called a reproducing kernel Hilbert space (RKHS) if the evaluation functional

$$e_x : \mathcal{H} \to \mathbf{R}, \quad f \mapsto f(x)$$

    is continuous for each $x \in \Omega$.

  - A Hilbert space $\mathcal{H}$ consisting of functions on Ω is a RKHS if and only if there exists $k(\cdot\,, x) \in \mathcal{H}$ (reproducing kernel) such that

$$\langle k(\cdot\,, x), f \rangle_{\mathcal{H}} = f(x) \qquad \forall f \in \mathcal{H},\ x \in \Omega.$$

    (by Riesz's lemma)

Reproducing Kernel Hilbert Space II

- Positive definite kernel and RKHS

  A symmetric kernel $k: \Omega \times \Omega \to \mathbf{R}$ is said to be positive definite, if for any $x_1,\ldots,x_n \in \Omega$ and $c_1,\ldots,c_n \in \mathbf{R}$,

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

Theorem (construction of RKHS)

  If $k: \Omega \times \Omega \to \mathbf{R}$ is positive definite, there uniquely exists a RKHS $\mathcal{H}_k$ on Ω such that

  (1) $k(\cdot, x) \in \mathcal{H}$ for all $x \in \Omega$,

  (2) the linear hull of $\{k(\cdot, x) \mid x \in \Omega\}$ is dense in $\mathcal{H}_k$,

  (3) $k(\cdot, x)$ is a reproducing kernel of $\mathcal{H}_k$, i.e.,

$$\langle k(\cdot, x), f \rangle_{\mathcal{H}_k} = f(x) \qquad \forall f \in \mathcal{H}_k, \ x \in \Omega.$$

Reproducing Kernel Hilbert Space III

- Some properties
  - If the pos. def. kernel k is of C^r, so is every function in $\mathcal{H}_k$.
  - If the pos. def. kernel k is bounded, so is every function in $\mathcal{H}_k$.

- Examples: positive definite kernels on $\mathbf{R}^m$
  - Euclidean inner product

$$k(x,y) = x^T y \qquad \mathcal{H}_k \cong \mathbf{R}^m$$

  - Gaussian RBF kernel

$$k(x,y) = \exp\left(-\|x-y\|^2 / \sigma^2\right) \qquad \dim \mathcal{H}_k = \infty$$

  - Polynomial kernel

$$k(x,y) = (x^T y + c)^d \quad (c \geq 0,\ d \in \mathbf{N}) \qquad \mathcal{H}_k = \{\text{polyn. deg} \leqq \text{d}\}$$

Exponential Manifold by RKHS

- Definitions

Ω: topological space. μ: Borel probability measure on Ω s.t. supp$\mu = \Omega$.

k : continuous pos. def. kernel on Ω such that $\mathcal{H}_k$ contains 1 (constants).

$$M_\mu(k) := \left\{ f : \Omega \to \mathbf{R} \mid f : \text{continuous}, \ f(x) > 0 \ (\forall x \in \Omega), \int f d\mu = 1, \ \exists \delta > 0, \ \int e^{\delta \sqrt{k(x,x)}} f(x) d\mu(x) < \infty \right\}$$

$M_\mu(k)$ is provided with a Hilbert manifold structure.

Note: If $\| u \| < \delta$, $\quad E_f[e^{u(X)}] = E_f[e^{\langle u, k(\cdot, X)\rangle}] \le E_f[e^{\|u\| \sqrt{k(X,X)}}] < \infty$.

  - Tangent space

$$T_f := \left\{ u \in \mathcal{H}_k \mid E_f[u(X)] = 0 \right\}$$ closed subspace of $\mathcal{H}_k$

Exponential Manifold by RKHS II

- Local coordinate

For $f \in M_\mu(k)$, $\quad W_f := \left\{ u \in T_f \mid \exists \delta > 0,\ E_f[e^{u(X)+\delta\sqrt{k(X,X)}}] < \infty \right\} \subset T_f$

Then, for any $u \in W_f$

$$f_u := \exp(u - \Psi_f(u)) f \quad \in M_\mu(k).$$

Define

$\xi_f : W_f \to M_\mu(k), \quad u \mapsto f_u$ (one-to-one) $\mathcal{E}_f := \xi_f(W_f)$

$\varphi_f : S_f \to W_f, \quad \varphi_f = \xi_f^{-1}$ → works as a local coordinate

Lemma

(1) W_f is an open subset of T_f.

(2) $g \in \mathcal{E}_f \quad \Leftrightarrow \quad \mathcal{E}_f = \mathcal{E}_g$.

Exponential Manifold by RKHS III

- Reproducing Kernel Exponential Manifold (RKEM)

<u>Theorem</u>. The system $\{(\mathcal{E}_f, \varphi_f)\}_{f \in M_\mu(k)}$ is a C^∞-atlas of $M_\mu(k)$.

coordinate transform

$$\varphi_g \circ \varphi_f^{-1}(u) = \log \frac{\exp(u - \Psi_f(u)) f}{g} - E_g\left[\log \frac{\exp(u - \Psi_f(u)) f}{g} \right]$$
$$= u + \log \frac{f}{g} - E_g\left[u + \log \frac{f}{g} \right]$$

- A structure of Hilbert manifold is defined on $M_\mu(k)$ with Riemannian metric $E_f[uv]$.
- Likelihood functional is continuous.
- The function $u(x)$ is decoupled in the inner product $\langle u, k(\cdot, x) \rangle$

  u: natural coordinate, $k(\cdot\,, x)$: sufficient statistics
- The manifold depends on the choice of k.

  e.g. $\Omega = \mathbf{R}$, $\mu = N(0,1)$, $k(x,y) = (xy+1)^2$. → $\mathcal{H}_k$ = {polyn. deg ≦ 2}

  $M_\mu(k) = \{ N(m, \sigma) \mid m \in \mathbf{R},\ \sigma > 0 \}$: the normal distributions.

Mean parameter of RKEM

- Mean parameter
  - For any $f \in M_\mu(k)$, there uniquely exists $m_f \in \mathcal{H}_k$ such that
    $$E_f[u(X)] = \langle u, m_f \rangle_{\mathcal{H}_k} \quad \text{for all } u \in \mathcal{H}_k.$$
  - The mean parameter does not necessarily give a coordinate, as in the case of the maximal exponential manifold.

- Empirical mean parameter
  - $X_1, \ldots, X_n$: i.i.d. sample ~ $f\mu$.

    Empirical mean parameter: $\hat{m}_n := \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)$

Fact 1. $\langle \hat{m}_n, f \rangle = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \qquad (\forall f \in \mathcal{H}_k)$

Fact 2. $\left\| \hat{m}_n - m_f \right\|_{\mathcal{H}_k} = O_p\left(1/\sqrt{n}\right) \qquad (n \to \infty)$

Applications of RKEM

- Maximum likelihood estimation (IGAIA2005)
  - Maximum likelihood estimation with regularization is possible.
  - The consistency of the estimator is proved.

- Statistical asymptotic theory of singular models
  - There are examples of statistical model which is a submodel of an infinite dimensional exponential family, but not embeddable into a finite dimensional exponential family.
  - For a submodel of RKEM, developing asymptotic theory of the maximum likelihood estimator is easy.

- Geometry of RKEM
  - Dual (± 1) connections can be introduced on the tangent bundle in some cases.

Statistical asymptotic theory of singular models

Singular Submodel of exponential family

- Standard asymptotic theory

Statistical model $\{ f(x;\theta) \mid \theta \in \Theta\}$ on a measure space $(\Omega, \mathcal{B}, \mu)$.

Θ: (finite dimensional) manifold.

"True" density: $f_0(x) = f(x;\theta_0) \quad (\theta_0 \in \Theta)$ $\qquad X_1, \ldots, X_n$: i.i.d. $\sim f_0\mu$

Maximum likelihood estimator (MLE)

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f(X_i;\theta)$$

Asymptotically normal

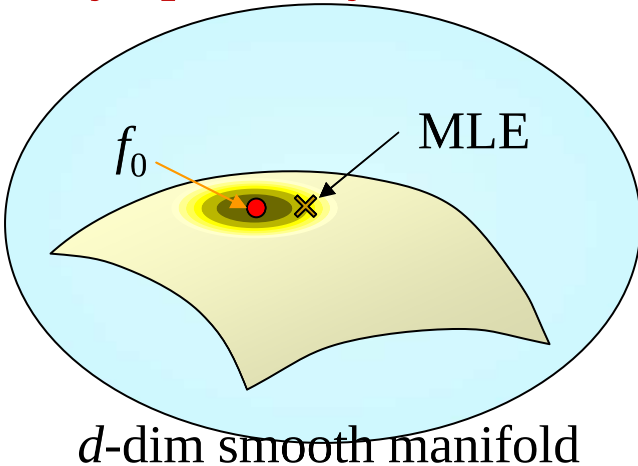
Under some regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, I(\theta_0)^{-1}) \text{ in law } \quad (n \to \infty)$$

Likelihood ratio

$$2\ell_n(\hat{\theta}_n) = 2\sum_{i=1}^{n} \log \frac{f(X_i;\hat{\theta}_n)}{f(X_i;\theta_0)} \Rightarrow \chi^2_d$$

in law $(n \to \infty)$



Singular Submodel of exponential family II

- Singular submodel in ordinary exponential family

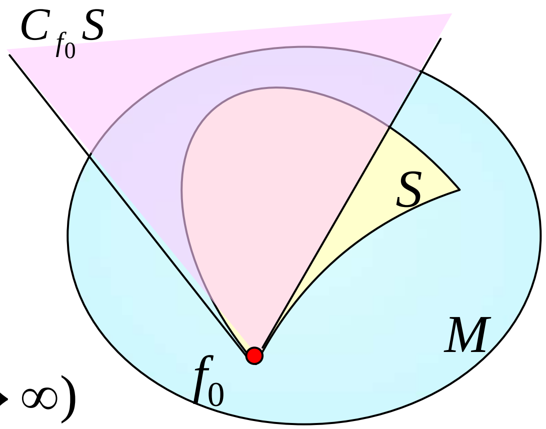
Finite dimensional exponential family M : $f(x;\theta)=\exp(\theta^T u(x)-\Psi(\theta))$

Submodel $S=\{f(x;\theta)\in M \mid \theta\in\Theta_S\}$ $(\theta\in\Theta)$

Tangent cone:

$$C_{f_0}S=\{\xi^T u(x)\in T_{f_0}M \mid \exists\{\theta_n\}\subset\Theta_S, \exists\lambda_n>0 \text{ s.t. } \lambda_n(\theta_n-\theta_0)\to\xi \quad (n\to\infty)\}$$

Under some regularity conditions,



$$\ell_n(\hat{\theta}_n)=\sum_{i=1}^n \log\frac{f(X_i;\hat{\theta}_n)}{f(X_i;\theta_0)}$$

$$=\frac{1}{2}\sup_{\xi^T u\in C_{f_0}S,\ E_{f_0}|\xi^T u|^2=1}\left\{\xi^T\left(\frac{1}{n}\sum_{i=1}^n u(X_i)\right)\right\}^2+o_p(1) \quad (n\to\infty)$$

projection of empirical mean parameter

More explicit formula can be derived in some cases.

Singular submodel in RKEM

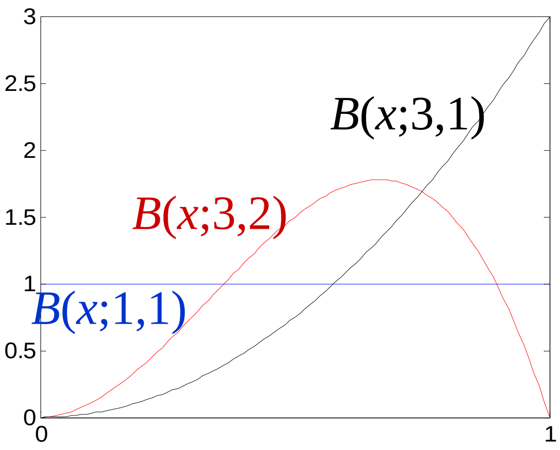
- Submodel of an infinite dimensional exponential family
  - There are some models, which are not embeddable into a finite dimensional exponential family, but can be embedded into an infinite dimensional RKEM.

- Example:

Mixture of Beta distributions (on $[0,1]$)

$$f(x;\alpha,\beta) = \alpha B(x;\beta,1) + (1-\alpha)B(x;1,1),$$

where $B(x;\beta,\gamma) = \frac{\Gamma(\beta+\gamma)}{\Gamma(\beta)\Gamma(\gamma)} x^{\beta-1}(1-x)^{\gamma-1}$



  - Singularity at $f_0(x) = f(x;0,\beta) = B(x;1,1)$

    β is not identifiable.

Singular submodel in RKEM II

- $\mathcal{H}_k$ = Sobolev space $H^1(0,1)$

$$k(x,y) = \exp(-|x-y|), \quad \|u\|_{H_k}^2 = \frac{1}{2}\left(u(0)^2 + u(1)^2\right) + \frac{1}{2}\int_0^1 \left(|u'(x)|^2 + |u(x)|^2\right)dx$$

Fact: $\log f(x;\alpha,\beta) \in H^1(0,1)$ for $0 \le \alpha < 1,\ \beta > 3/2.$

- Submodel of $\mathcal{E}_{f0}$

$$u_{\alpha,\beta}(x) := \log f(x;\alpha,\beta) - E_{f_0}[\log f(x;\alpha,\beta)]$$

$$S = \{ f(\cdot;\alpha,\beta) = \exp(u_{\alpha,\beta} - \Psi_f(u_{\alpha,\beta})) f_0 \mid 0 \le \alpha < 1, \beta > 3/2\}$$

⇨ S is a submodel of $\mathcal{E}_{f0}$, and f_0 is a singularity of S.

- Tangent cone at f_0 is not finite dimensional.

$$\frac{\log f(\cdot;\alpha,\beta)}{\alpha} \to w_\beta := \beta x^{\beta-1} - 1 \quad (\alpha \downarrow 0) \text{ in } H^1(0,1)$$

Singular submodel in RKEM III

- General theory of singular submodel

$M_\mu(k)$: RKEM. $f \in M_\mu(k)$,

Submodel $S \subset E_f$ defined by $\varphi : K \times [0,1] \to T_f$

such that

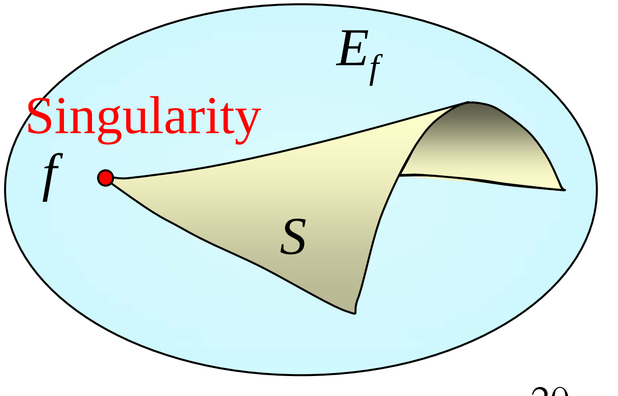
(1) K: compact set

(2) $\varphi(a,t) = 0 \Leftrightarrow t = 0$

(3) $\varphi(a,t)$: Frechet differentiable w.r.t. t and

$$\frac{\partial \varphi}{\partial t}(a,t) \quad \text{is continuous on } K \times [0,1]$$

(4) $\min_{a \in K} \left\| \frac{\partial \varphi}{\partial t}(a,t)\Big|_{t=0} \right\| > 0$



Singular submodel in RKEM IV

Lemma (tangent cone)

$$C_f S = \mathbf{R}_{\geq}\left\{ \frac{\partial \varphi}{\partial t}(a,t)\big|_{t=0} \;\middle|\; a \in K \right\}$$

Theorem

$$\sup_{g \in S} \sum_{i=1}^{n} \log \frac{g(X_i)}{f(X_i)} = \frac{1}{2} \sup_{w \in C_f S, E_f |w|^2 = 1} \underline{\langle w, \hat{m}_n \rangle}^2 + o_p(1) \qquad (n \to \infty)$$

projection of empirical mean parameter

$$\underset{\text{in law}}{\Rightarrow} \frac{1}{2} \sup_{w \in C_f S, E_f |w|^2 = 1} G_w^{\,2}$$

G_w: Gaussian process

- Analogue to the asymptotic theory on submodel in a finite dimensional exponential family.
- The same assertion holds without assuming exponential family, but the sufficient conditions and the proof are much more involved.

Summary

- Exponential Hilbert manifolds, which can be infinite dimensional, is defined using reproducing kernel Hilbert spaces.
- From the estimation viewpoint, an interesting class is submodels of infinite dimensional exponential manifolds, which are not embeddable into a finite dimensional exponential family.
- The asymptotic behavior of MLE is analyzed for singular submodels of infinite dimensional exponential manifolds.