再生核による指数分布族の構成とその統計的推定への応用

福水　健次

統計数理研究所

総合研究大学院大学

2008年6月12日

統計数学セミナー

Outline

- Introduction
- Reproducing kernel Hilbert space and positive definite kernel
- Reproducing kernel exponential manifold (RKEM)
- Application 1: Pseudo-maximum likelihood estimation with RKEM
- Application 2: Statistical asymptotic theory of singular models
- Conclusion

Introduction

Maximal Exponential Manifold

- Maximal exponential manifold (Pistone &Sempi 95)
  - Idea: a Banach manifold is defined so that the cumulant generating function is well-defined on a neighborhood of each probability density.

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

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

  - Orlicz space $L_{\cosh\text{-}1}(f)$

$$\begin{aligned} L_{\cosh-1}(f) &= \{u \mid \exists \alpha > 0 \text{ s.t. } E_f[\cosh(\alpha u)] < \infty\} \\ &= \{u \mid \exists \alpha > 0 \text{ s.t. } E_f[e^{\alpha u}] < \infty \text{ and } E_f[e^{-\alpha u}] < \infty\} \end{aligned}$$

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.

    A small change of u may cause a very different likelihood.

Reproducing Kernel Hilbert Space and Positive Definite Kernel

Reproducing kernel Hilbert space

- Reproducing kernel Hilbert space (RKHS)
  - Ω: set. A Hilbert space $\mathcal{H}$ consisting of functions on Ω is called a (real-valued) 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) for each $x \in \Omega$ s.t.

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

(by Riesz's lemma)

Positive definite kernel and RKHS

- Positive definite kernel

A symmetric function k: Ω x Ω → **R** is said to be *positive definite*, if for any $n \in N$ and $x_1, \dots, x_n \in \Omega$, the matrix $\left(k(x_i, x_j)\right)$ (Gram matrix) is positive semidefinite, i.e.

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \qquad \text{(for any } c_1, \dots, c_n \in \mathbf{R}\text{)}.$$

  - A reproducing kernel is positive definite.

- Positive definite kernel and RKHS

Theorem (Moore-Aronszajn)

If k: Ω x Ω → **R** is positive definite, there uniquely exists a RKHS $\mathcal{H}_k$ consisting of functions on Ω such that

(1) The linear hull of $\{k(\cdot, x) : \Omega \to \mathbf{R} \mid x \in \Omega\}$ is dense in $\mathcal{H}_k$.

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

Example of positive definite kernel

- Euclidean inner product on $\mathbf{R}^m$

$$k(x,y) = x^T y$$

- Polynomial kernel on $\mathbf{R}^m$

$$k(x,y) = (x^T y + c)^d \qquad (c \geq 0,\ d \in \mathbf{N})$$

$$\mathcal{H}_k = \{\text{polyn. deg} \leqq \text{d}\}$$

- Gaussian kernel on $\mathbf{R}^m$

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

- Laplacian kernel on [0 1]

$$k(x,y) = \exp\left(-|x-y|\right)$$

$$\mathcal{H}_k = H^1(0,1) = \left\{u \in L^2[0,1] \mid \exists u' \in L^2[0,1]\right\} \qquad \text{(Sobolev space)}$$

$$\|u\|^2_{\mathcal{H}_k} = \frac{1}{2}\left\{u(0)^2 + u(1)^2\right\} + \frac{1}{2}\int_0^1 \left\{u(x)^2 + u'(x)^2\right\} dx$$

Some properties of RKHS

- For $f = \sum_{i=1}^{n} a_i k(\cdot\,, x_i),\ \ g = \sum_{j=1}^{m} b_j k(\cdot\,, y_j),$

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{ij} a_i b_j k(x_i, y_j).$$

  In particular,
$$\|k(\cdot, x)\|_{\mathcal{H}} = \sqrt{k(x, x)}.$$

- If a pos. def. kernel k is of class C^d, so are all the functions in $\mathcal{H}_k$.

  $\because$) for C^0,

$$|f(x) - f(y)| = |\langle k(\cdot, x) - k(\cdot, y), f \rangle| \leq \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}_k} \|f\|_{\mathcal{H}_k}$$

$$\|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}_k}^2 = k(x, x) - 2k(x, y) + k(y, y)$$

Reproducing Kernel Exponential Manifold

Exponential Manifold by RKHS

- Definitions

Ω: topological space. μ: Borel measure on Ω s.t. support of $\mu = \Omega$.
k : continuous pos. def. kernel on Ω such that $\mathcal{H}_k$ contains 1 (constants).

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

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

Note: If $\| u \| < \delta$, $E_f[e^{u(X)}] = E_f[e^{\langle u, k(\cdot, X) \rangle}] \le E_f[e^{\|u\| \sqrt{k(X,X)}}] < \infty$.
If k is bounded, the condition $E_f[e^{\delta \sqrt{k(x,x)}}] < \infty$ is not needed.

- Tangent space

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

Exponential Manifold by RKHS (cont'd)

- 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 $u \in W_f$

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

$$\left(\because\ E_{f_u}[e^{\delta\sqrt{k(X,X)}}] = E_f[e^{\delta\sqrt{k(X,X)}} e^{u(X)-\Psi_f(u)}] < \infty.\right)$$

Define

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

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

Lemma

(1) W_f is an open subset of T_f.

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

Exponential Manifold by RKHS (cont'd)

- Reproducing Kernel Exponential Manifold (RKEM)

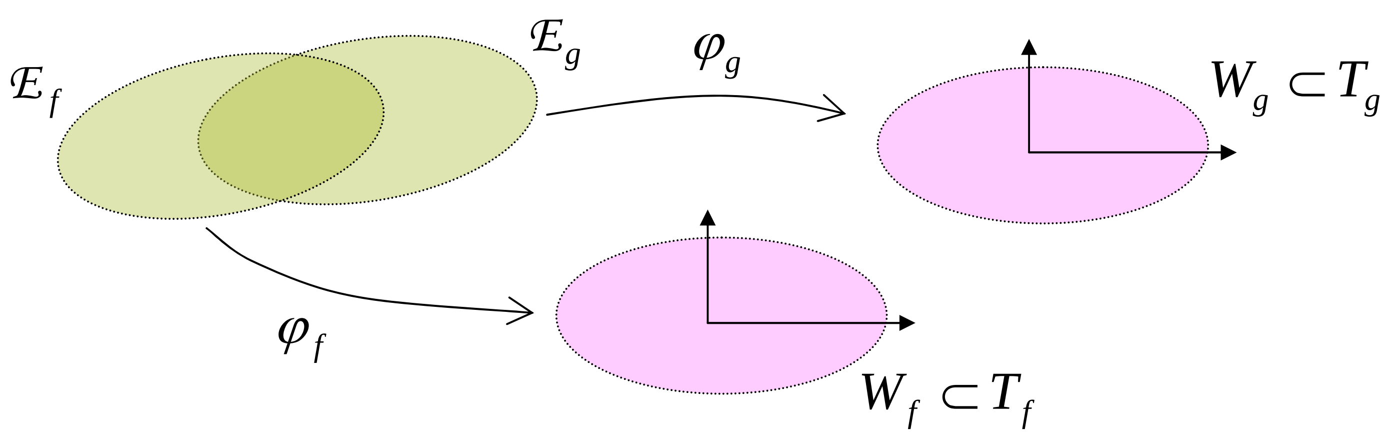
Theorem.

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

(1) $\mathcal{E}_f \cap \mathcal{E}_g \neq \phi \quad \Rightarrow \quad \varphi_f(\mathcal{E}_f \cap \mathcal{E}_g)$ is open in T_f.

(2) $\mathcal{E}_f \cap \mathcal{E}_g \neq \phi$

$\Rightarrow \quad \varphi_g \circ \varphi_f^{-1}|_{\varphi_f(\mathcal{E}_f \cap \mathcal{E}_g)} : \varphi_f(\mathcal{E}_f \cap \mathcal{E}_g) \to \varphi_g(\mathcal{E}_f \cap \mathcal{E}_g)$ is a C^∞-map.



Exponential Manifold by RKHS (cont'd)

- Sketch of the proof

  (2). Let $h \in \mathcal{E}_f \bigcap \mathcal{E}_g$ and $u = \varphi_f(h)$, i.e., $h = \exp(u - \Psi_f(u)) f$.

  Then,

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

$$= u + \log \frac{f}{g} - E_g\left[u + \log \frac{f}{g} \right]$$

  $u \mapsto E_g[u]$ is affine on W_f, thus, of C^∞.

- A structure of C^∞ Hilbert manifold is defined on $M_\mu(k)$.

Properties of RKEM

- Properties of RKEM as a Hilbert manifold
  - The Hilbert manifold $M_\mu(k)$ depends on the choice of a kernel k.
  - The tangent space at $f \in M_\mu(k)$ is identified with T_f, which is codimension one in $\mathcal{H}_k$.
  - $\mathcal{E}_f$ is a connected component in $M_\mu(k)$, and

$$\mathcal{E}_f = \{ g \in M_\mu(k) \mid \exists u \in T_f ,\ g = \exp(u - \Psi_f(u)) f \}$$

  - Log-likelihood $u(X) - \Psi_f(u) + \log f(X)$ is continuous on $M_\mu(k)$.
  - Sufficient statistics = $k(x,y)$

$$\exp(u(x) - \Psi_f(u)) = \exp(\langle u, k(\cdot, x) \rangle - \Psi_f(u))$$

c.f. finite dimensional case: $\exp(\theta \cdot s(x) - \Psi(\theta))$

Examples of RKEM

- RKEM includes any finite dimensional exponential family.
- $\Omega = \mathbf{R}$, $\mu = N(0,1)$
  $k(x,y) = (xy+1)^2$. → $\mathcal{H}_k$ = {polyn. deg ≦ 2}
  $M_\mu(k) = \{N(m, \sigma) \mid m \in \mathbf{R}, \sigma > 0\}$: exponential family of normal distributions.

- $\Omega = [0\ 1]$, $\mu = \mathrm{Unif}[0,1]$

  $k(x,y) = \exp\left(-|x-y|\right)$ → $\mathcal{H}_k = H^1(0,1)$.

  $M_\mu(k) = \left\{ f : [0,1] \to \mathbf{R} \mid f : \text{continuous}, f > 0, \int_0^1 f(x)dx = 1 \right\}$

  $\because)\quad k(x,x) = 1 \quad \Rightarrow \quad E_f[e^{\delta\sqrt{k(X,X)}}] < \infty.$

Moments in RKEM

- Mean parameter: for any $f \in M_\mu(k)$, there uniquely exists $m_f \in \mathcal{H}_k$ such that

$$E_f[u(X)] = \langle u, m_f \rangle_{\mathcal{H}_k} \quad \text{for all } u \in \mathcal{H}_k.$$

$$m_f(y) = E_f[k(y, X)]$$: mean of the sufficient statistics $k(\cdot, x)$

- Covariance operator: for any $f \in M_\mu(k)$, there uniquely exists an operator Σ_f on $\mathcal{H}_k$ such that

$$\langle v, \Sigma_f u \rangle_{\mathcal{H}_k} = \mathrm{Cov}_f[v(X), u(X)] \quad \text{for all } u, v \in \mathcal{H}_k.$$

- Derivatives of cumulant generating function

For $g = e^{u - \Psi_f(u)} f \quad (u \in T_f)$ and $v_1, v_2 \in T_f$,

the derivatives of Ψ_f at u in the direction of v_1 (and v_2) are given by

$$D_u \Psi_f(v_1) = E_g[v_1(X)] = \langle v_1, m_g \rangle_{\mathcal{H}_k}$$

$$D_u \Psi_f(v_1, v_2) = \mathrm{Cov}_g[v_1(X), v_2(X)] = \langle v_2, \Sigma_g v_1 \rangle_{\mathcal{H}_k}$$

Pseudo-Maximum Likelihood Estimation with RKEM

MLE with RKEM

- Likelihood equation

$\mathcal{E}$: connected component of $M_\mu(k)$. $f_0 \in \mathcal{E}$: fixed.

$$\mathcal{E} = \{ f \in M_\mu(k) \mid \exists u \in T_{f_0},\ \ f = f_u = \exp(u - \Psi_{f_0}(u)) f_0 \}$$

$f_* = f_{u_*}$: true p.d.f. to give i.i.d. samle $X_1, \ldots, X_n \sim f_* \mu$.

MLE in $\mathcal{E}$: $\max_{f \in E} \sum_{i=1}^n \log f(X_i) = \max_{u \in W_0} \sum_{i=1}^n u(X_i) - n\Psi_0(u)$

$\Longrightarrow$ $\max_{u \in W_0} \langle \hat{m}^{(n)}, u \rangle - \Psi_0(u)$ where $\hat{m}^{(n)} = \frac{1}{n} \sum_{i=1}^n k(\cdot, X_i)$

$\Longrightarrow$ $\langle m_u, v \rangle_{\mathcal{H}_k} = \langle \hat{m}^{(n)}, v \rangle_{\mathcal{H}_k}$ for all $v \in \mathcal{H}_k$. -- Moment matching

$\Longrightarrow$ $m_u = \hat{m}^{(n)}$

The ML mean parameter should be $\hat{m}^{(n)}$.
What is the corresponding p.d.f. element (or natural parameter u) in the RKEM?

MLE is impossible with RKEM

- Rigorous MLE is impossible for RKEM in general.
  - The mean parameter m_f uniquely determines the probability for a certain class of kernels (characteristic kernel, Fukumizu et al. 08).

    {probability measure on Ω} $\to \mathcal{H}_k, \quad P \mapsto m_P$ is injective.

    e.g.) Gaussian kernel $k(x,y) = \exp\left(-\|x-y\|^2 / \sigma^2\right)$

    - Moment matching with the empirical distribution is impossible.
    - *c.f.* For a finite dimensional exponential family, the moments are given by only the finite number of sufficient statistics.

  - Mean parameter is not a coordinate in general (Pistone & Rogatin 99)

    $u \mapsto m_u = D\Psi_0(u)$ does not have a continuous inverse, because $D^2\Psi_0(u,v) = \langle v, \Sigma_f u \rangle$ and Σ_f can have arbitrary small eigenvalues.

Asymptotics of mean parameter

- Theorem ($\sqrt{n}$-consistency of the ML mean parameter)

  $(\Omega, \mathcal{B}, P)$: probability space.

  k: positive definite kernel on Ω s.t. $E_P[k(X,X)] < \infty$.

  $X_1,\ldots,X_n$: i.i.d. ~ P. $\hat{m}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)$

  $$\Longrightarrow \quad \left\| \hat{m}^{(n)} - m_P \right\|_{\mathcal{H}_k} = O_p\left(1/\sqrt{n}\right) \qquad (n \to \infty)$$

  Proof) $E\left\| \hat{m}^{(n)} - m_P \right\|_{\mathcal{H}_k}^2 = \frac{1}{n}\left\{E[k(X,X)] - E[k(X,\tilde{X})]\right\},$

  where $\tilde{X}$ is an independent copy of X. *q.e.d.*

- Theorem implies the uniform law of large numbers;

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

- Convergence in law to a Gaussian process **G** on $\mathcal{H}$ is also known.

Pseudo-MLE with RKEM

- Pseudo-MLE by regularization

  $\{\mathcal{H}^{(\ell)}\}_{\ell=1}^{\infty}$: sequence of finite dim. subspaces in $\mathcal{H}_k$ such that $\mathcal{H}^{(\ell)} \subset \mathcal{H}^{(\ell+1)}$ and the inclusions $\mathcal{H}^{(\ell)} \to \mathcal{H}^{(\ell+1)}$, $\mathcal{H}^{(\ell)} \to \mathcal{H}_k$ are continuous.

  $$T_f^{(\ell)} = T_f \cap \mathcal{H}^{(\ell)}, \quad W_f^{(\ell)} = W_f \cap \mathcal{H}^{(\ell)}$$

  Pseudo-MLE: $\hat{u}^{(\ell)} := \arg\max_{u \in W_f^{(\ell)}} \left[\langle \hat{m}^{(n)}, u \rangle - \Psi_0(u) \right]$

  - Assumptions

    (A-1) For $u \in W_f$, let $u_*^{(\ell)} := \min_{w \in W_f^{(\ell)}} KL(f_u \| f_w)$. Then,

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

    (A-2) $\exists \delta > 0, \exists (\ell_n)_{n=1}^{\infty} \subset \mathbf{N}$ s.t.

    $\lambda^{(\ell)} := \inf_{u \in \mathcal{H}_k, \|u - u_*\| \le \delta} \ \inf_{v \in T_{f_u}^{(\ell)}, \|v\| \le 1} \langle v, \Sigma_{f_u} v \rangle$ satisfies $\lim_{n \to \infty} \sqrt{n}\, \lambda^{(\ell_n)} = +\infty.$

    (stability)

Consistency of Pseudo-MLE

Theorem (Fukumizu, IGAIA2005)

$\hat{f}_n = \exp(\hat{u}^{(n)} - \Psi_0(\hat{u}^{(n)})) f_0.$

Under the assumptions (A-1) and (A-2),

$$KL(f_* \| \hat{f}_n) \to 0 \quad (n \to \infty) \qquad \text{in probability.}$$

Sketch of the Proof)

$$KL(f_* \| \hat{f}_n) = KL(f_* \| f_{u_*^{(\ell_n)}}) + \Psi_0(\hat{u}^{(n)}) - \Psi_0(u_*^{(\ell_n)}) - E_{f_*}[\hat{u}^{(n)} - u_*^{(\ell_n)}]$$

(i) $KL(f_* \| f_{u_*^{(\ell_n)}}) \to 0$ by (A-1).

(ii) For the rest terms, it suffices to show $\Pr\left(\| \hat{u}^{(n)} - u_*^{(\ell_n)} \| \ge \varepsilon \right) \to 0$ for an arbitrary $\varepsilon > 0$.

$$\Pr\left(\| \hat{u}^{(\ell_n)} - u_*^{(\ell_n)} \| \ge \varepsilon \right)$$
$$\le \Pr\left(\sup_{u \in W^{(\ell_n)},\ \|u - u_*^{(\ell_n)}\| \ge \varepsilon} \left\langle u, \hat{m}^{(n)} \right\rangle - \Psi_0(u) \ge \left\langle u_*^{(\ell_n)}, \hat{m}^{(n)} \right\rangle - \Psi_0(u_*^{(\ell_n)}) \right) \equiv \mathbf{P}_n$$

Consistency of Pseudo-MLE (cont'd)

For any $u \in W^{(\ell_n)}$,

$$\langle u, \hat{m}^{(n)} \rangle - \langle u_*^{(\ell_n)}, \hat{m}^{(n)} \rangle - \Psi_0(u) + \Psi_0(u_*^{(\ell_n)})$$

$$= \langle u - u_*^{(\ell_n)}, \hat{m}^{(n)} - m_{f_*} \rangle - \langle u - u_*^{(\ell_n)}, m_{f_*} - m_{u_*^{(\ell_n)}} \rangle \quad = 0$$

$$+ \langle u - u_*^{(\ell_n)}, m_{u_*^{(\ell_n)}} \rangle - \Psi_0(u) + \Psi_0(u_*^{(\ell_n)})$$

$$= \langle u - u_*^{(\ell_n)}, \hat{m}^{(n)} - m_{f_*} \rangle - \left\{ \Psi_0(u) - \Psi_0(u_*^{(\ell_n)}) - \langle u - u_*^{(\ell_n)}, m_{u_*^{(\ell_n)}} \rangle \right\} \qquad (*)$$

By convexity of Ψ_0, the supremum can be considered in a neighborhood.
By (A-2),

$$(*) \le \| u - u_*^{(\ell_n)} \| \, \| \hat{m}^{(n)} - m_{f_*} \| - \tfrac{1}{2} \lambda^{(\ell_n)} \| u - u_*^{(\ell_n)} \|^2$$

$$\Longrightarrow \qquad \mathbf{P}_n \le \Pr\left(\left\| \hat{m}^{(n)} - m_{f_*} \right\| \ge \tfrac{1}{2} \lambda^{(\ell_n)} \varepsilon \right) \to 0.$$

q.e.d.

Remarks on pseudo-MLE

- Remarks
  - If $\mathcal{H}_k$ is finite dimensional, the Pseudo-MLE is equal to the ordinary MLE.
  - How to construct $\{\mathcal{H}^{(\ell)}\}_{\ell=1}^{\infty}$?

    $$\mathcal{H}^{(\ell)} = span\{k(\cdot, X_1), \ldots, k(\cdot, X_\ell)\}$$

    When does this satisfy the assumptions? → future work.
  - Another way of regularization – Tikhonov regularization. (Canu&Smola06)

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

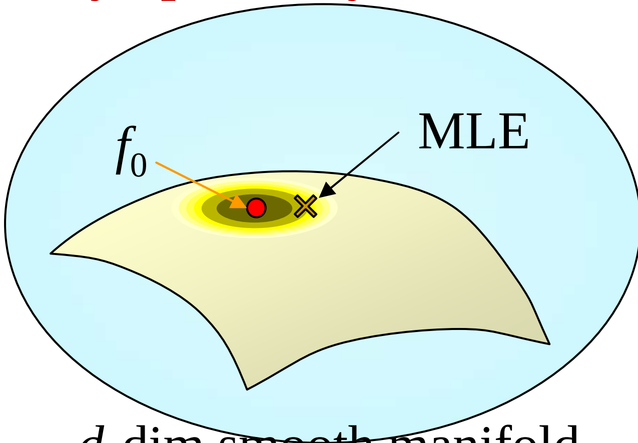
Θ: (finite dimensional) manifold.

"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



d-dim smooth manifold

Under some regularity conditions,

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

Likelihood ratio

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

in law $(n \to \infty)$

Singular Submodel of exponential family (cont'd)

- Singular submodel in ordinary exponential family

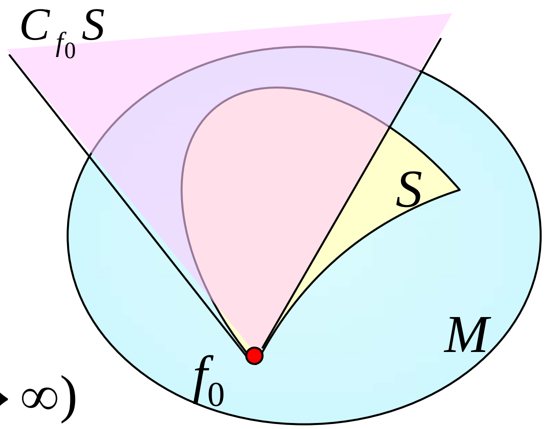
Finite dimensional exponential family $M: \; f(x;\theta) = \exp(\theta^T u(x) - \Psi(\theta))$

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

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
  - The tangent cone of a model defined by a finite number of parameters may not be in a finite dimensional space.
  - Interesting parametric models are
    - not embeddable into a finite dimensional exponential family,
    - but can be embedded into an infinite dimensional RKEM.

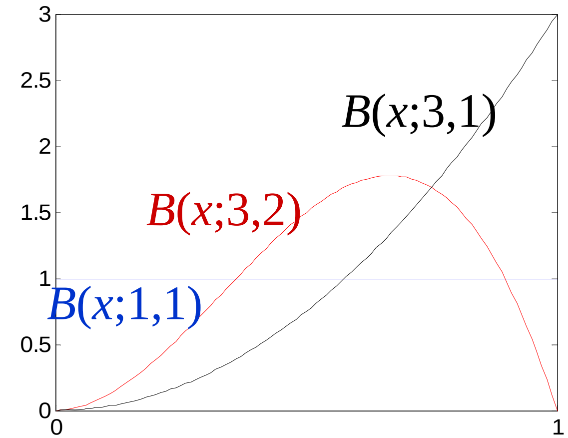
Mixture of Beta distributions

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

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

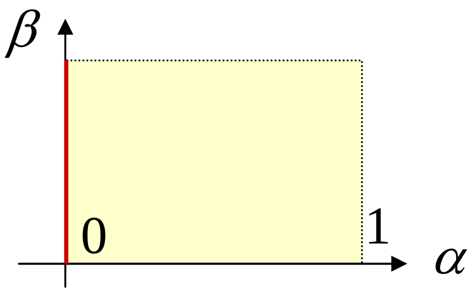
  $$B(x;\beta,\gamma) = \frac{\Gamma(\beta+\gamma)}{\Gamma(\beta)\Gamma(\gamma)} x^{\beta-1}(1-x)^{\gamma-1}$$

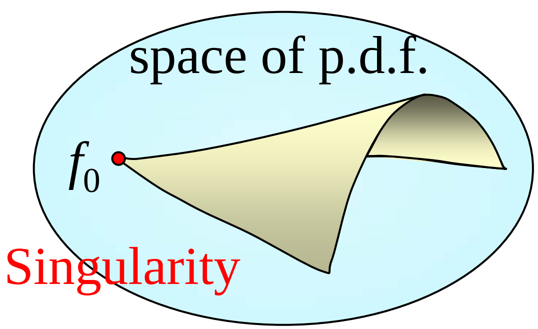
  : Beta distribution



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

  If $\alpha = 0$, β is not identifiable.
  → singularity in the space of probability densities.





Mixture of Beta distributions (cont'd)

- $\mathcal{H}_k$ = Sobolev space $H^1(0,1)$ defined by $k(x,y) = \exp(-|x-y|)$.

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

- RKEM with the Sobolev space
  Connected component including $f_0 = \mathrm{Unif}[0,1]$:

$$\mathcal{E}_{f_0} = \{ g \in M_\mu(k) \mid \exists u \in T_{f_0},\ g = \exp(u - \Psi_{f_0}(u)) f_0 \}$$

- Submodel of $\mathcal{E}_{f_0}$

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

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

  ⇨ f_0 is a singularity of S.

- Tangent cone at f_0 is not finite dimensional.

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

Asymptotics on singular submodel

- General theory of singular submodel

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

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

$$S = \{\exp(u - \Psi_f(u)) f \in E_f \mid u \in \varphi(K \times [0,1])\}$$

such that

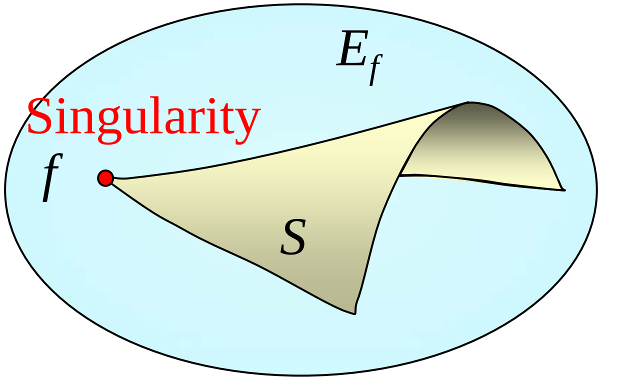
(1) K: compact set

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

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



Asymptotics on singular submodel (cont'd)

Lemma (tangent cone)

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

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 a 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.

Conclusion

- Reproducing kernel exponential manifold are defined as a Hilbert manifold.
  - It is an extension of ordinary finite dimensional exponential family.
  - The model depends on the choice of kernel; the dimension is either finite or infinite.
  - It allows estimation for finite sample, since the likelihood is a continuous functional.

- The pseudo-MLE based on a series of finite dimensional subspaces is proposed, and proved to be consistent.
- It can be used for the asymptotic theory of singular models. The theoretical discussion is easier than general cases.

- Future works:
  - Application to expectation propagation.
  - Dual geometry on reproducing kernel exponential manifolds.

References

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

Fukumizu, K. (2008) Exponential manifold by reproducing kernel Hilbert spaces. to appear.

Fukumizu, K. (2005) Infinite dimensional exponential families by reproducing kernel Hilbert spaces. *Proc. 2nd Intern, Symp. Information Geometry and its Applications (IGAIA2005).* 324-333.

Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B. (2008) Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems 20*. 489-496.

Pistone, G., Rogantin, M. P. (1999). The exponential statistical manifold: mean. parameters, orthogonality and space transformations. *Bernoulli* 5 (4), 721–760.

Appendix: Relation with maximal exponential manifold

Proposition

Let $f \in M_\mu(k)$, and

$$A_f = \inf\left\{\alpha > 0 \,\middle|\, E_f\left[\exp\left(\sqrt{k(X,X)}/\alpha\right)\right] \le 2\right\}.$$

Then,

$$\mathcal{H}_k \subset L_{\cosh -1}(f) \quad \text{and} \quad \|u\|_{L_{\cosh-1}(f)} \le A_f \|u\|_{\mathcal{H}_k} \quad \text{for any } u \in \mathcal{H}_k.$$

Proof.

$$E_f[\cosh(u(X)/\alpha) - 1] = \tfrac{1}{2} E_f[e^{u(X)/\alpha} + e^{-u(X)/\alpha}] - 1$$

$$\le E_f[e^{|u(X)|/\alpha}] - 1 \le E_f\left[\exp\left(\frac{\|u\|_{\mathcal{H}_k}}{\alpha}\sqrt{k(X,X)}\right)\right] - 1.$$

Thus, $\|u\|_{\mathcal{H}_k} / \alpha < 1/A_f \Rightarrow E_f[\cosh(u(X)/\alpha) - 1] \le 1.$ *q.e.d.*

- $M_\mu(k)$ is a subset of the exponential manifold proposed by Pistone and Sempi (1995)