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.

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

$$f_{u} = \exp(u - \Psi_{f}(u)) f, \quad \Psi_{f}(u) = \log E_{f}[e^{u}] < \infty$$

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

$$L_{\cosh^{-1}}(f) = \{ u \mid \exists \alpha > 0 \text{ s.t. } E_{f}[\cosh(\alpha u)] < \infty \}$$

$$= \{ u \mid \exists \alpha > 0 \text{ s.t. } E_{f}[e^{\alpha u}] < \infty \text{ and } E_{f}[e^{-\alpha u}] < \infty \}$$

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.
    
    \[
    X_1, \ldots, X_n : \text{i.i.d. } \sim f_{0\mu} \quad X_n = (X_1, \ldots, X_n)
    \]

    MLE: \( \theta \) that maximizes
    \[
    \ell_n(\theta; X_n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{a=1}^{m} \theta^a u_a(X_i) - \Psi(\theta) \right\}
    \]

  - Is MLE extendable to the maximal exponential manifold?
    \[
    \ell_n(u; X_n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ u(X_i) - \Psi_f(u) \right\}
    \]
    
    ➡️ 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)
  - $\Omega$: set. A Hilbert space $\mathcal{H}$ consisting of functions on $\Omega$ is called a (real-valued) reproducing kernel Hilbert space (RKHS) if the evaluation functional
    \[ e_x : \mathcal{H} \to \mathbb{R}, \quad f \mapsto f(x) \]
    is continuous for each $x \in \Omega$.

  - A Hilbert space $\mathcal{H}$ consisting of functions on $\Omega$ is a RKHS if and only if there exists $k(\cdot, x) \in \mathcal{H}$ (reproducing kernel) for each $x \in \Omega$ s.t.
    \[ \langle k(\cdot, x), f \rangle_{\mathcal{H}} = f(x) \quad \forall f \in \mathcal{H}, \ x \in \Omega. \quad \text{[reproducing property]} \]
    (by Riesz’s lemma)
Positive definite kernel and RKHS

Positive definite kernel
A symmetric function $k: \Omega \times \Omega \to \mathbb{R}$ is said to be positive definite, if for any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \Omega$, the matrix $\left(k(x_i, x_j)\right)$ (Gram matrix) is positive semidefinite, i.e.
\[
\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \quad \text{(for any } c_1, \ldots, c_n \in \mathbb{R}).
\]

A reproducing kernel is positive definite.

Positive definite kernel and RKHS

Theorem (Moore-Aronszajn)
If $k: \Omega \times \Omega \to \mathbb{R}$ is positive definite, there uniquely exists a RKHS $\mathcal{H}_k$ consisting of functions on $\Omega$ such that
(1) The linear hull of $\{k(\cdot, x) : \Omega \to \mathbb{R} \mid x \in \Omega\}$ 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 $\mathbb{R}^m$
  \[ k(x, y) = x^T y \]

- Polynomial kernel on $\mathbb{R}^m$
  \[ k(x, y) = (x^T y + c)^d \quad (c \geq 0, d \in \mathbb{N}) \]
  \[ \mathcal{H}_k = \{ \text{polyn. deg } \leq d \} \]

- Gaussian kernel on $\mathbb{R}^m$
  \[ k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right) \quad \text{dim } \mathcal{H}_k = \infty \]

- Laplacian kernel on $[0, 1]$
  \[ k(x, y) = \exp\left(-|x - y|\right) \]
  \[ \mathcal{H}_k = H^1(0,1) = \left\{ u \in L^2[0,1] | \exists u' \in L^2[0,1] \right\} \quad \text{(Sobolev space)} \]
  \[ \|u\|_{\mathcal{H}_k}^2 = \frac{1}{2} \left\{ u(0)^2 + u(1)^2 \right\} + \frac{1}{2} \int_0^1 \left\{ u(x)^2 + u'(x)^2 \right\} dx \]
Some properties of RKHS

- For \( f = \sum_{i=1}^{n} a_i k(\cdot, x_i), \quad g = \sum_{j=1}^{m} b_j k(\cdot, y_j), \)
  \[
  \langle f, g \rangle_{H_k} = \sum_{ij} a_i b_j k(x_i, y_j).
  \]
  In particular,
  \[
  \|k(\cdot, x)\|_{H_k} = \sqrt{k(x, x)}.
  \]

- If a pos. def. kernel \( k \) is of class \( C^d \), so are all the functions in \( H_k \).

  \[
  \therefore \text{ for } C^0, \\
  |f(x) - f(y)| = \langle k(\cdot, x) - k(\cdot, y), f \rangle \leq \|k(\cdot, x) - k(\cdot, y)\|_{H_k} \|f\|_{H_k} \\
  \|k(\cdot, x) - k(\cdot, y)\|_{H_k}^2 = k(x, x) - 2k(x, y) + k(y, y)
  \]
Reproducing Kernel Exponential Manifold
Exponential Manifold by RKHS

Definitions

$\Omega$: topological space. $\mu$: Borel measure on $\Omega$ s.t. support of $\mu = \Omega$.

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

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

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

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

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

Then, for any \( u \in W_f \)
\[
f_u := \exp(u - \Psi_f(u))f \in M_\mu(k).
\]
\[
\therefore \ E_f[u \left( e^{\delta\sqrt{k(X,X)}} \right)] = E_f \left[ e^{\delta\sqrt{k(X,X)}} e^{u(X)-\Psi_f(u)} \right] < \infty.
\]

Define
\[
\xi_f : W_f \rightarrow M_\mu(k), \quad u \mapsto f_u \quad \text{(one-to-one)} \quad \mathcal{E}_f := \xi_f(W_f)
\]
\[
\varphi_f : \mathcal{E}_f \rightarrow W_f, \quad \varphi_f = \xi_f^{-1} \rightarrow \text{works as a local coordinate}
\]

Lemma
(1) \( W_f \) is an open subset of \( T_f \).
(2) \( g \in \mathcal{E}_f \iff \mathcal{E}_f = \mathcal{E}_g \).
Reproducing Kernel Exponential Manifold (RKEM)

Theorem.

The system \( \{(E_f, \varphi_f)\}_{f \in M_\mu(k)} \) is a \( C^\infty \)-atlas of \( M_\mu(k) \), i.e.,

1. \( E_f \cap E_g \neq \emptyset \Rightarrow \varphi_f(E_f \cap E_g) \) is open in \( T_f \).

2. \( \varphi_g \circ \varphi_f^{-1} \mid_{\varphi_f(E_f \cap E_g)} : \varphi_f(E_f \cap E_g) \to \varphi_g(E_f \cap E_g) \) is a \( C^\infty \)-map.
Sketch of the proof

(2). Let \( h \in E_f \cap E_g \) 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 \( C^\infty \).

A structure of \( C^\infty \) 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 $\mathcal{H}_k$.
  - $E_f$ is a connected component in $M_\mu(k)$, and
    $$E_f = \{g \in M_\mu(k) \mid \exists u \in T_f, \ g = \exp(u - \Psi_f(u))f\}$$
  - 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.
- $\Omega = \mathbb{R}, \mu = \mathcal{N}(0,1)$
  $k(x,y) = (xy+1)^2$.  $\Rightarrow \mathcal{H}_k = \{\text{polyn. deg } \leq 2\}$
  $M_\mu(k) = \{\mathcal{N}(m, \sigma) \mid m \in \mathbb{R}, \sigma > 0\}$: exponential family of normal distributions.

- $\Omega = [0,1], \mu = \text{Unif}[0,1]$
  $k(x, y) = \exp(-|x - y|)$  $\Rightarrow \mathcal{H}_k = H^1(0,1)$.
  $M_\mu(k) = \{f : [0,1] \to \mathbb{R} \mid f : \text{continuous}, f > 0, \int_{0}^{1} f(x) dx = 1\}$
  $\therefore k(x,x) = 1 \Rightarrow E_f[e^{\delta \sqrt{k(X,X)}}] < \infty.$
Moments in RKEM

- **Mean parameter**: for any \( f \in M_{\mu}(k) \), there uniquely exists \( m_f \in \mathcal{H}_k \) such that
  \[
  E_f[u(X)] = \langle u, m_f \rangle_{\mathcal{H}_k} 
  \]
  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 \( \Sigma_f \) on \( \mathcal{H}_k \) such that
  \[
  \langle v, \Sigma_f u \rangle_{\mathcal{H}_k} = \text{Cov}_f[v(X), u(X)] 
  \]
  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 \( \Psi_f \) at \( u \) in the direction of \( v_1 \) (and \( v_2 \)) are given by
  \[
  D_u \Psi_f(v_1) = E_g[v_1(X)] = \langle v_1, m_g \rangle_{\mathcal{H}_k}
  \]
  \[
  D_u \Psi_f(v_1, v_2) = \text{Cov}_g[v_1(X), v_2(X)] = \langle v_2, \Sigma_g v_1 \rangle_{\mathcal{H}_k}
  \]
Pseudo-Maximum Likelihood Estimation with RKEM
MLE with RKEM

- **Likelihood equation**

  \[ \mathcal{E}: \text{connected component of } M_\mu(k). \quad f_0 \in \mathcal{E}: \text{fixed.} \]

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

  \[ f_* = f_{u*} : \text{true p.d.f. to give i.i.d. sample } X_1, \ldots, X_n \sim f_*\mu. \]

  **MLE in} \mathcal{E}:**

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

  \[ \Rightarrow \max_{u \in W_0} \langle \hat{m}^{(n)}, u \rangle - \Psi_0(u) \quad \text{where} \quad \hat{m}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \]

  \[ \Rightarrow \langle m_u, v \rangle_{\mathcal{H}_k} = \langle \hat{m}^{(n)}, v \rangle_{\mathcal{H}_k} \quad \text{for all } v \in \mathcal{H}_k. \quad \text{-- Moment matching} \]

  \[ \Rightarrow m_u = \hat{m}^{(n)} \]

  The ML mean parameter should be \( \hat{m}^{(n)} \).

  What is the corresponding p.d.f. element (or natural parameter \( u \)) in the RKEM?
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).
  \[
  \{\text{probability measure on } \Omega\} \rightarrow H_k, \quad P \mapsto m_P \quad \text{is injective.}
  \]
  
  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
  \[
  D^2\Psi_0(u, v) = \langle v, \Sigma_f u \rangle \quad \text{and } \Sigma_f \text{ can have arbitrary small eigenvalues.} 
  \]
Asymptotics of mean parameter

- Theorem ($\sqrt{n}$-consistency of the ML mean parameter)
  \((\Omega, \mathcal{B}, P)\) : probability space.
  \(k\): positive definite kernel on \(\Omega\) s.t. \(E_p[k(X,X)] < \infty\).
  \(X_1, \ldots, X_n\): i.i.d. \(\sim P\).
  \(\hat{m}(n) = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i)\)
  \(\implies \left\| \hat{m}(n) - m_p \right\|_{\mathcal{H}_k} = O_p \left(1/\sqrt{n}\right) \quad (n \to \infty)\)

  \textbf{Proof}\) \(E\left\| \hat{m}(n) - m_p \right\|^2_{\mathcal{H}_k} = \frac{1}{n} \left\{ E[k(X,X)] - E[k(X, \tilde{X})] \right\},\)
  where \(\tilde{X}\) is an independent copy of \(X\). \(q.e.d.\)

- Theorem implies the uniform law of large numbers;
  \[ \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - E_p[f(X)] \right| = O_p \left(1/\sqrt{n}\right).\]

- Convergence in law to a Gaussian process \(\mathcal{G}\) 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 \( \mathcal{H}^{(\ell)} \rightarrow \mathcal{H}^{(\ell+1)} \), \( \mathcal{H}^{(\ell)} \rightarrow \mathcal{H}_k \) are continuous.

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

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

- **Assumptions**

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

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

  (approximation)

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

  \[ \lambda^{(\ell)} : = \inf_{u \in \mathcal{H}_k, \|u - u_*\| \leq \delta} \inf_{v \in T_f^{(\ell)}, \|v\| \leq 1} \langle v, \Sigma f_u v \rangle \] satisfies \( \lim_{n \rightarrow \infty} \sqrt{n} \lambda^{(\ell_n)} = +\infty. \)

  (stability)
Consistency of Pseudo-MLE

Theorem (Fukumizu, IGAIA2005)
\[ \hat{f}_n = \exp(\hat{\mu}^{(n)} - \Psi_0(\hat{\mu}^{(n)})) f_0. \]
Under the assumptions (A-1) and (A-2),
\[ KL(f_* || \hat{f}_n) \to 0 \quad (n \to \infty) \quad \text{in probability.} \]

Sketch of the Proof
\[ KL(f_* || \hat{f}_n) = KL(f_* || f_{u_*^{(\ell_n)}}) + \Psi_0(\hat{\mu}^{(n)}) - \Psi_0(u_*^{(\ell_n)}) - E_{f_*}[\hat{\mu}^{(n)} - u_*^{(\ell_n)}] \]
(i) \[ KL(f_* || f_{u_*^{(\ell_n)}}) \to 0 \quad \text{by (A-1)}. \]
(ii) For the rest terms, it suffices to show \[ \Pr\left(\|\hat{\mu}^{(n)} - u_*^{(\ell_n)}\| \geq \varepsilon\right) \to 0 \]
for an arbitrary \( \varepsilon > 0 \).
\[ \Pr\left(\|\hat{\mu}^{(\ell_n)} - u_*^{(\ell_n)}\| \geq \varepsilon\right) \leq \Pr\left(\sup_{u \in W^{(\ell_n)}, \|u-u_*^{(\ell_n)}\| \geq \varepsilon} \langle u, \hat{m}^{(n)} \rangle - \Psi_0(u) \geq \langle u_*^{(\ell_n)}, \hat{m}^{(n)} \rangle - \Psi_0(u_*^{(\ell_n)})\right) \equiv P_n \]
Consistency of Pseudo-MLE (cont’d)

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

\[
\langle u, \hat{m}^{(n)} \rangle - \langle u_0^{(\ell_n)}, \hat{m}^{(n)} \rangle - \Psi_0(u) + \Psi_0(u_0^{(\ell_n)})
\]

\[
= \langle u - u_0^{(\ell_n)}, \hat{m}^{(n)} - m_f \rangle - \langle u - u_0^{(\ell_n)}, m_f - m_{u_0^{(\ell_n)}} \rangle = 0
\]

\[
+ \langle u - u_0^{(\ell_n)}, m_{u_0^{(\ell_n)}} \rangle - \Psi_0(u) + \Psi_0(u_0^{(\ell_n)})
\]

\[
= \langle u - u_0^{(\ell_n)}, \hat{m}^{(n)} - m_f \rangle - \left\{ \Psi_0(u) - \Psi_0(u_0^{(\ell_n)}) - \langle u - u_0^{(\ell_n)}, m_{u_0^{(\ell_n)}} \rangle \right\} \quad (\ast)
\]

By convexity of \( \Psi_0 \), the supremum can be considered in a neighborhood. By (A-2),

\[
(\ast) \leq \| u - u_0^{(\ell_n)} \| \| \hat{m}^{(n)} - m_f \| - \frac{1}{2} \lambda^{(\ell_n)} \| u - u_0^{(\ell_n)} \|^2
\]

\[
\implies P_n \leq \Pr\left( \| \hat{m}^{(n)} - m_f \| \geq \frac{1}{2} \lambda^{(\ell_n)} \epsilon \right) \to 0.
\]

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 $\{\mathcal{H}^{(\ell)}\}_{\ell=1}^\infty$?

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

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

  \( \Theta \): (finite dimensional) manifold.

  "True" density: \( f_0(x) = f(x; \theta_0) \) \( (\theta_0 \in \Theta) \)

  Maximum likelihood estimator (MLE)

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

  Under some regularity conditions,

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

  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 (cont’d)

- Singular submodel in ordinary exponential family

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

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

Tangent cone:

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

Under some regularity conditions,

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

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

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 \([0,1]\))

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

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

- Beta distribution

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

\( \alpha = 0, \ \beta \) is not identifiable.

\( \Rightarrow \) singularity in the space of probability densities.
Mixture of Beta distributions (cont’d)

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

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

- RKEM with the Sobolev space
  Connected component including $f_0 = \text{Unif}[0,1]$: 
    \[ \mathcal{E}_{f_0} = \{ g \in M_\mu(k) | \exists u \in T_{f_0}, \ g = \exp(u - \Psi_{f_0}(u)) f_0 \} \]

- Submodel of $\mathcal{E}_{f_0}$
  \[ u_{\alpha,\beta}(x) := \log f(x; \alpha, \beta) - E_{f_0} [\log f(x; \alpha, \beta)] \in T_{f_0} \]
  \[ S = \{ f(\cdot; \alpha, \beta) = \exp(u_{\alpha,\beta} - \Psi_f(u_{\alpha,\beta})) f_0 | 0 \leq \alpha < 1, \beta > 3/2 \} \]
  $\Rightarrow f_0$ is a singularity of $S$.

- Tangent cone at $f_0$ is not finite dimensional.
  \[ \frac{\log f(\cdot; \alpha, \beta)}{\alpha} \rightarrow w_\beta := \beta x^{\beta - 1} - 1 \ (\alpha \downarrow 0) \text{ in } H^1(0,1) \]
Asymptotics on singular submodel

- General theory of singular submodel
  \[ M_\mu(k): \text{RKEM. } f \in M_\mu(k), \]

  Submodel \( S \subset E_f \) defined by \( \varphi : K \times [0,1] \rightarrow T_f \)
  \[ S = \{ \exp(u - \Psi_f(u))f \in E_f \mid u \in \varphi(K \times [0,1]) \} \]

  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
     \[ \frac{\partial \varphi}{\partial t}(a,t) \] is continuous on \( K \times [0,1] \)
  4. \( \min_{a \in K} \left\| \frac{\partial \varphi}{\partial t}(a,t) \right\|_{t=0} > 0 \)
Asymptotics on singular submodel (cont’d)

**Lemma** (tangent cone)

\[
C_f S = \mathbb{R} = \left\{ \frac{\partial \phi}{\partial t}(a,t) \big|_{t=0} \bigg| a \in K \right\}
\]

**Theorem**

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

projection of empirical mean parameter

\[
\Rightarrow \frac{1}{2} \sup_{w \in C_f S, E_f |w|^2=1} G_w^2 \quad G_w: \text{Gaussian process}
\]

• 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


Appendix: Relation with maximal exponential manifold

Proposition

Let \( f \in M_\mu(k) \), and \( A_f = \inf\{ \alpha > 0 \mid E_f[\exp(\sqrt{k(X,X)/\alpha})] \leq 2 \} \).

Then,

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

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

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

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

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