Unwrapping the "Exponential Manifold by Reproducing Kernel Hilbert Spaces"

(K. Fukumizu, in *Algebraic and Geometric Methods in Statistics, 2009*)

Dino S.

Gatsby Unit

October 18, 2012

Infinite dimensional exponential family?

- Aim: construct an infinite-dimensional exponential family, on which estimation theory can be built

Infinite dimensional exponential family?

- Aim: construct an infinite-dimensional exponential family, on which estimation theory can be built
- In particular: theory of consistent estimation with a finite sample

Infinite dimensional exponential family?

- Aim: construct an infinite-dimensional exponential family, on which estimation theory can be built
- In particular: theory of consistent estimation with a finite sample
- A subset of an **RKHS** as a functional parameter space

Infinite dimensional exponential family?

- Aim: construct an infinite-dimensional exponential family, on which estimation theory can be built
- In particular: theory of consistent estimation with a finite sample
- A subset of an **RKHS** as a functional parameter space
- Maximum-Likelihood ill-posed

Infinite dimensional exponential family?

- Aim: construct an infinite-dimensional exponential family, on which estimation theory can be built
- In particular: theory of consistent estimation with a finite sample
- A subset of an **RKHS** as a functional parameter space
- Maximum-Likelihood ill-posed
- Pseudo-Maximum-Likelihood: restrict attention to a sequence of finite dimensional submanifolds, where dimensionality increases with the sample size

Introduction

- For simplicity, let $\mathcal{X} = [0,1]$, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded continuous kernel, with RKHS $\mathcal{H}_k$.
- **assumption 0**: $\mathcal{H}_k$ contains the constant functions, $u(x) = c$. If it does not, then consider $k(x,y) + 1$ instead. This does. (This assumption is made so that $\mathcal{H}_k$ is closed under subtracting constants)
- Since k is a bounded kernel on bounded domain, integrals $\int u(x)dx$ and $\int e^{u(x)}dx$ converge $\forall u \in \mathcal{H}_k$.
- Let $T := \left\{u \in \mathcal{H}_k \;:\; \int u(x)dx = 0\right\}$. In other words, consider the uniform distribution 1 on $\mathcal{X}$, and define its kernel embedding $\mu_1 = \int k(\cdot, x) \cdot 1dx \in \mathcal{H}_k$. Then, $T = \mu_1^{\perp}$, as $\int u(x)dx = \langle u, \mu_1 \rangle_{\mathcal{H}_k}$.
- Since $\mathcal{H}_k$ includes constants, $u - \int u(x)dx \in T$, $\forall u \in \mathcal{H}_k$

Densities parametrized by RKHS functions

- Now, pick $u \in T$, and define:

$$\Psi(u) = \log \int e^{u(x)} dx$$

Lemma

$\forall u \in T$, $e^{u-\Psi(u)}$ is a valid probability density function on $\mathcal{X}$, and map $\xi : u \mapsto e^{u-\Psi(u)}$ is one-to-one.

Proof.

$\xi(u) = \xi(v) \implies u(x) - v(x) = \Psi(v) - \Psi(u) = \text{const.} \implies \int u(x)dx - \int v(x)dx = \text{that same constant} \implies \text{that constant must be zero.}$ $\square$

- $S = \xi(T)$ is now a set of probability density functions on $\mathcal{X}$ associated to the kernel k, which inherits the Hilbertian structure of $T \subset \mathcal{H}_k$ (exponential Hilbert manifold).
- Let's write f_u for $\xi(u)$ for short (read: "density with parameter u"). We get the usual stuff with fancier names - we can take Fréchet derivatives of Ψ:

$$\begin{aligned} D_u\Psi(v) &= \mathbb{E}_{X\sim f_u}[v(X)] = \langle v, \mu_u\rangle_{\mathcal{H}_k} \\ D_u^2\Psi(v_1, v_2) &= Cov_{X\sim f_u}[v_1(X), v_2(X)] = \langle v_1, \Sigma_u v_2\rangle_{\mathcal{H}_k} \end{aligned}$$

where:

$$\begin{aligned} \mu_u &:= \mathbb{E}_{f_u}[k(\cdot, X)] \\ \Sigma_u &:= \mathbb{E}_{f_u}[k(\cdot, X) \otimes k(\cdot, X)] - \mathbb{E}_{f_u}[k(\cdot, X)] \otimes \mathbb{E}_{f_u}[k(\cdot, X)], \end{aligned}$$

are the kernel embedding and the kernel covariance operator of density f_u.

- In the exponential family language:

- In the exponential family language:

1. $u \in T$ is the natural parameter of f_u, and $k(\cdot, x)$ is the sufficient statistic, as $f_u \propto e^{\langle u, k(\cdot,x)\rangle} = e^{u(x)}$

- In the exponential family language:

1. $u \in T$ is the natural parameter of f_u, and $k(\cdot, x)$ is the sufficient statistic, as $f_u \propto e^{\langle u, k(\cdot, x) \rangle} = e^{u(x)}$
2. $\mu_u \in \mathcal{H}_k$ is the mean parameter of f_u, as it is the mean of the sufficient statistic

- In the exponential family language:

1. $u \in T$ is the natural parameter of f_u, and $k(\cdot, x)$ is the sufficient statistic, as $f_u \propto e^{\langle u, k(\cdot, x) \rangle} = e^{u(x)}$
2. $\mu_u \in \mathcal{H}_k$ is the mean parameter of f_u, as it is the mean of the sufficient statistic

- For characteristic kernels, the mapping $P \mapsto \mu_P$ is injective. In particular, $f_u \mapsto \mu_u$ is injective.

- In the exponential family language:

1. $u \in T$ is the natural parameter of f_u, and $k(\cdot, x)$ is the sufficient statistic, as $f_u \propto e^{\langle u, k(\cdot,x)\rangle} = e^{u(x)}$
2. $\mu_u \in \mathcal{H}_k$ is the mean parameter of f_u, as it is the mean of the sufficient statistic

- For characteristic kernels, the mapping $P \mapsto \mu_P$ is injective. In particular, $f_u \mapsto \mu_u$ is injective.
- So, there is a reparametrization $T \ni u \mapsto \mu_u \in \mathcal{H}_k$

- In the exponential family language:

1. $u \in T$ is the natural parameter of f_u, and $k(\cdot, x)$ is the sufficient statistic, as $f_u \propto e^{\langle u, k(\cdot,x)\rangle} = e^{u(x)}$
2. $\mu_u \in \mathcal{H}_k$ is the mean parameter of f_u, as it is the mean of the sufficient statistic

- For characteristic kernels, the mapping $P \mapsto \mu_P$ is injective. In particular, $f_u \mapsto \mu_u$ is injective.
- So, there is a reparametrization $T \ni u \mapsto \mu_u \in \mathcal{H}_k$
- Note that there are certainly $\mu_u \notin T$. In particular, $u = 0 \in T \Rightarrow f_u = 1 \Rightarrow \mu_u = \mu_1 \perp T$. Conclusion: the mean parameters and the natural parameters do not have the same domain.

KL in S

Kullback-Leibler divergence in the exponential manifold:

$$
\begin{aligned}
KL\left(f_u \| f_v\right) &= \int f_u(x) \log \frac{f_u(x)}{f_v(x)} dx \\
&= \int f_u(x) \left[u(x) - \Psi(u) - v(x) + \Psi(v)\right] dx \\
&= \Psi(v) - \Psi(u) + \langle u - v, \mu_u \rangle_{\mathcal{H}_k} .
\end{aligned}
$$

KL in S (2)

Theorem (Pythagorean KL-relation)

Consider a closed subspace $V \subset T$, let $f_* \in S$, and set:

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

i.e., $f_{u_{opt}}$ is the KL-nearest density in $\xi(V)$ to f_*. Then $\forall u \in V$:

$$KL\left(f_* \| f_u\right) = KL\left(f_* \left\| f_{u_{opt}}\right.\right) + KL\left(f_{u_{opt}} \| f_u\right).$$

The KL divergence between f_* and f_u that is parametrized by a subspace V of T can be broken down as the KL divergence between f_* and the nearest density in that subspace (***approximation error***) plus the KL divergence between the best approximator and a given density (***estimation error***).

RKHS norm and KL divergence

Lemma

Let $\|u_n - u\|_{\mathcal{H}_k} = o(\alpha_n)$, *where* $\lim_{n\to\infty} \alpha_n = 0$. *Then* $KL\,(f_u \,\|\, f_{u_n}) = o(\alpha_n)$.

RKHS norm and KL divergence

Proof.

Start with:

$$KL\left(f_u \,\|\, f_{u_n}\right) \;\leq\; \left|\Psi(u_n) - \Psi(u)\right| + \left|\langle u_n - u, \mu_u\rangle_{\mathcal{H}_k}\right|.$$

By Taylor-expansion, we obtain:

$$\begin{aligned}
\left|\Psi(u_n) - \Psi(u)\right| &= \left|\langle u_n - u, \mu_u\rangle + \frac{1}{2}\left\langle u_n - u, \Sigma_{\tilde{u}}(u_n - u)\right\rangle\right| \\
&\leq \left\|u_n - u\right\| \left[\left\|\mu_u\right\| + \frac{1}{2}\lambda_{\max}\left\|u_n - u\right\|\right], \text{ and thus:} \\
KL\left(f_u \,\|\, f_{u_n}\right) &\leq \left\|u_n - u\right\| \left[2\left\|\mu_u\right\| + \frac{1}{2}\lambda_{\max}\left\|u_n - u\right\|\right],
\end{aligned}$$

where $\tilde{u}$ is a convex combination of u_n and u and $\lambda_{\max}$ is the largest eigenvalue of $\Sigma_{\tilde{u}}$. □

Maximum Likelihood

Observe data $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} f_{u^*}$. Consider the log-likelihood:

$$
\begin{aligned}
L_n(u) &= \frac{1}{n}\sum_{i=1}^{n} \log p(X_i|u) \\
&= \frac{1}{n}\sum_{i=1}^{n} u(X_i) - \Psi(u) \\
&= \left\langle u, \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i) \right\rangle - \Psi(u) \\
&= \left\langle u, \hat{\mu}^{(n)} \right\rangle - \Psi(u)
\end{aligned}
$$

Maximum Likelihood (2)

- Differentiate:

$$D_u L_n(v) \;=\; \left\langle v, \hat{\mu}^{(n)} - \mu_u \right\rangle$$

Maximum Likelihood (2)

- Differentiate:

$$D_u L_n(v) \quad = \quad \left\langle v, \hat{\mu}^{(n)} - \mu_u \right\rangle$$

- ML solution is trivial! Set the mean parameter to the empirical mean parameter and solve for u:

$$\mu_u \quad = \quad \hat{\mu}^{(n)}$$

Maximum Likelihood (2)

- Differentiate:

$$D_u L_n(v) = \left\langle v, \hat{\mu}^{(n)} - \mu_u \right\rangle$$

- ML solution is trivial! Set the mean parameter to the empirical mean parameter and solve for u:

$$\mu_u = \hat{\mu}^{(n)}$$

- **oops**: $\hat{\mu}^{(n)}$ does not correspond to any natural parameter $u \in T$

Maximum Likelihood (3)

- The inverse mapping from the mean parameter to the natural parameter is not bounded
  - the derivative of map $u \mapsto \mu_u$: since μ_u can be identified as the first derivative of the cumulant generating function Ψ, i.e., $D_u\Psi = \langle \cdot, \mu_u \rangle$, the derivative of this map is Σ_u (kernel covariance operator). This is known to be a trace-class operator, so it has arbitrarily small positive eigenvalues.

Maximum Likelihood (3)

- The inverse mapping from the mean parameter to the natural parameter is not bounded
  - the derivative of map $u \mapsto \mu_u$: since μ_u can be identified as the first derivative of the cumulant generating function Ψ, i.e., $D_u\Psi = \langle \cdot, \mu_u \rangle$, the derivative of this map is Σ_u (kernel covariance operator). This is known to be a trace-class operator, so it has arbitrarily small positive eigenvalues.
- If $\hat{\mu}^{(n)}$ would correspond to some natural parameter $\hat{u} \in T$, then a distribution with the **continuous density** $e^{\hat{u}(x)-\Psi(\hat{u})}$ and the empirical distribution $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ must have the same kernel embedding. Recall that this is impossible because, e.g., for characteristic kernels, the mapping $P \mapsto \mu_P$ is injective!

Pseudo-MLE

First, while we cannot go from the mean parameters to the natural parameters, it is still true that mean parameters are useful, namely:

Theorem ($\sqrt{n}$-consistency of the empirical embedding estimator)

$$\left\|\hat{\mu}^{(n)} - \mu_{u_*}\right\|_{\mathcal{H}_k} = \mathcal{O}_P(1/\sqrt{n}).$$

Pseudo-MLE (2)

1. Define a series of finite-dimensional subspaces of T: $\left\{T^{(n)}\right\}_{l=1}^{\infty}$, and the n-th Pseudo-MLE:

$$\hat{u}^{(n)} = \arg\max_{u \in T^{(n)}} \left\langle u, \hat{\mu}^{(n)} \right\rangle - \Psi(u)$$

(the finite-dimensional MLE problem over $T^{(n)}$ which we can solve for u).

Pseudo-MLE (2)

1. Define a series of finite-dimensional subspaces of T: $\left\{T^{(n)}\right\}_{l=1}^{\infty}$, and the n-th Pseudo-MLE:

$$\hat{u}^{(n)} = \arg\max_{u\in T^{(n)}} \left\langle u, \hat{\mu}^{(n)}\right\rangle - \Psi(u)$$

(the finite-dimensional MLE problem over $T^{(n)}$ which we can solve for u).

2. In addition, introduce:

$$\begin{aligned} u_*^{(n)} &= \arg\min_{u\in V} KL\left(f_* \,\|\, f_u\right) \\ &= \arg\max_{u\in T^{(n)}} \langle u, \mu_{u_*}\rangle - \Psi(u) \end{aligned}$$

(the best approximator to the true u_* in $T^{(n)}$).

Assumptions

- **assumption 1**: $\forall u_*$

$$\left\| u_* - u_*^{(n)} \right\|_{\mathcal{H}_k} = o(\gamma_n),\ \gamma_n \to 0,$$

which means that $T^{(n)}$ approximates T with a sub-γ_n rate as $n \to \infty$

Assumptions

- **assumption 1**: $\forall u_*$

$$\left\| u_* - u_*^{(n)} \right\|_{\mathcal{H}_k} = o(\gamma_n), \; \gamma_n \to 0,$$

which means that $T^{(n)}$ approximates T with a sub-γ_n rate as $n \to \infty$

- **assumption 2**: the sequence of subspaces is chosen so that the smallest positive eigenvalues $\lambda^{(n)}$ of Σ_u restricted to $T^{(n)}$ decrease slowly enough (slower than $1/\sqrt{n}$):

$$\frac{1}{\sqrt{n}\lambda^{(n)}} = o(\epsilon_n), \; \epsilon_n \to 0.$$

Consistency of Pseudo-MLE

Theorem

$$KL\left(f_* \,\middle\|\, f_{\hat{u}^{(n)}}\right) = o_p\left(\max\left\{\gamma_n, \epsilon_n\right\}\right)$$

Proof sketch

- Break up KL in two parts:

$$KL\left(f_* \,\|\, f_{\hat{u}^{(n)}}\right) \;=\; KL\left(f_* \,\middle\|\, f_{u_*^{(n)}}\right) + KL\left(f_{u_*^{(n)}} \,\|\, f_{\hat{u}^{(n)}}\right)$$

Proof sketch

- Break up KL in two parts:

$$KL\left(f_* \| f_{\hat{u}^{(n)}}\right) = KL\left(f_* \middle\| f_{u_*^{(n)}}\right) + KL\left(f_{u_*^{(n)}} \| f_{\hat{u}^{(n)}}\right)$$

- The first one is $o(\gamma_n)$ by assumption 1, and Lemma connecting RKHS norm and KL divergence

Proof sketch

- Break up KL in two parts:

$$KL\left(f_* \| f_{\hat{u}^{(n)}}\right) = KL\left(f_* \middle\| f_{u_*^{(n)}}\right) + KL\left(f_{u_*^{(n)}} \| f_{\hat{u}^{(n)}}\right)$$

- The first one is $o(\gamma_n)$ by assumption 1, and Lemma connecting RKHS norm and KL divergence
- For the second term, we will need to show that the estimation error is $o_p(\epsilon_n)$, i.e.,

$$\mathbb{P}\left[\left\|\hat{u}^{(n)} - u_*^{(n)}\right\| \geq \epsilon_n\right] \rightarrow 0$$

Proof sketch (2)

- $\left\|\hat{u}^{(n)} - u_*^{(n)}\right\| \geq \epsilon_n$ implies that we have found a maximizer $\hat{u}^{(n)}$ of $L_n(u) > L_n(u_*^{(n)})$ outside the ϵ_n- ball centered at $u_*^{(n)}$ in $T^{(n)}$. Consider the Taylor expansion of L_n around $u_*^{(n)}$:

$$
\begin{aligned}
& L_n(\hat{u}^{(n)}) - L_n(u_*^{(n)}) = \\
& D_u L_n \Big|_{u=u_*^{(n)}} [\hat{u}^{(n)} - u_*^{(n)}] + \frac{1}{2} D_u^2 L_n \Big|_{u=u_*^{(n)}} [\hat{u}^{(n)} - u_*^{(n)}, \hat{u}^{(n)} - u_*^{(n)}] \\
= \; & \left\langle \hat{u}^{(n)} - u_*^{(n)}, \hat{\mu}^{(n)} - \mu_{u_*^{(n)}} \right\rangle - \frac{1}{2} \left\langle \hat{u}^{(n)} - u_*^{(n)}, \Sigma_{\tilde{u}} \left(\hat{u}^{(n)} - u_*^{(n)} \right) \right\rangle \\
= \; & \left\langle \hat{u}^{(n)} - u_*^{(n)}, \hat{\mu}^{(n)} - \mu_{u_*} \right\rangle - \frac{1}{2} \left\langle \hat{u}^{(n)} - u_*^{(n)}, \Sigma_{\tilde{u}} \left(\hat{u}^{(n)} - u_*^{(n)} \right) \right\rangle \\
\leq \; & \left\| \hat{u}^{(n)} - u_*^{(n)} \right\| \left[\left\| \hat{\mu}^{(n)} - \mu_{u_*} \right\| - \frac{1}{2} \lambda^{(n)} \epsilon_n \right].
\end{aligned}
$$

Proof sketch (3)

$$\begin{aligned}\mathbb{P}\left[\left\|\hat{u}^{(n)} - u_*^{(n)}\right\| \geq \epsilon_n\right] &\leq \mathbb{P}\left[\left\|\hat{\mu}^{(n)} - \mu_{u_*}\right\| \geq \frac{1}{2}\lambda^{(n)}\epsilon_n\right] \\ &\leq \mathbb{P}\left[\left\|\hat{\mu}^{(n)} - \mu_{u_*}\right\| \geq \frac{1}{\sqrt{n}}\right] \to 0,\end{aligned}$$

by the $\sqrt{n}$-consistency of the empirical embedding estimator.

Extensions and Discussion

- domain not bounded: $dx \to \phi(x)dx$, for some density ϕ, and $u \mapsto e^{u-\Psi(u)}\phi$, and restrict attention to an open subset of T for which densities are well defined.
  - e.g., for $\mathcal{X} = \mathbb{R}$, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and $k(x,y) = (1+xy)^2$, we recover Gaussian densities.

Extensions and Discussion

- domain not bounded: $dx \to \phi(x)dx$, for some density ϕ, and $u \mapsto e^{u-\Psi(u)}\phi$, and restrict attention to an open subset of T for which densities are well defined.
  - e.g., for $\mathcal{X} = \mathbb{R}$, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and $k(x, y) = (1 + xy)^2$, we recover Gaussian densities.
- kernel not a bounded function: more u's to discard

Extensions and Discussion

- domain not bounded: $dx \to \phi(x)dx$, for some density ϕ, and $u \mapsto e^{u-\Psi(u)}\phi$, and restrict attention to an open subset of T for which densities are well defined.
  - e.g., for $\mathcal{X} = \mathbb{R}$, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and $k(x, y) = (1 + xy)^2$, we recover Gaussian densities.
- kernel not a bounded function: more u's to discard
- Pseudo MLE is consistent, provided that a sequence of subspaces is chosen in a particular way
  - How to construct such sequences of subspaces?
  - Do they exist for any kernels - i.e., ensuring that the approximation error goes to zero quickly enough while the smallest eigenvalues decay slowly enough?