Density Estimation in Infinite Dimensional Exponential Families

Bharath K. Sriperumbudur

Statistical Laboratory, University of Cambridge

Nonparametric Measures of Dependence, 2014
Columbia University

Acknowledgements

- Prof. Kenji Fukumizu : The Institute for Statistical Mathematics, Tokyo, Japan.
- Dr. Arthur Gretton : Gatsby Computational Neuroscience Unit, University College London.
- Revant Kumar : Indian Institute of Technology, Kharagpur.
- Prof. Aapo Hyvärinen : University of Helsinki.

The Exponential Family of Distributions

- Natural form:

$$p_\theta(x) = q_0(x) e^{\theta^T T(x) - A(\theta)}$$

  where
  - $\theta \in \Theta \subset \mathbb{R}^m$ (natural parameter)
  - q_0: probability density defined over $\Omega \subset \mathbb{R}^d$
  - $A(\theta)$: log-partition function

$$A(\theta) = \log \int e^{\theta^T T(x)} q_0(x)\, dx$$

  - $T(x)$: sufficient statistic
- Important class of parametric statistical models
- Includes many commonly used distributions
  - Normal, Binomial, Poisson, Exponential, ...

Many Nice Properties

- Maximum entropy distribution consistent with given constraints on moments
- Natural parameter space, $\{\theta \in \Theta : A(\theta) < \infty\}$ is convex
- $\mathbb{E}_\theta[T_i(X)] = \frac{\partial A(\theta)}{\partial \theta_i},\ 1 \leq i \leq m$
- $\mathrm{Cov}_\theta[T_i(X), T_j(X)] = \frac{\partial^2 A(\theta)}{\partial \theta_i \partial \theta_j},\ 1 \leq i, j \leq m$
- Conjugate prior (useful in Bayesian inference)
- Maximum likelihood estimator (MLE) of θ depends only on the sufficient statistic, $T(x)$ (by Neyman-Fisher factorization theorem).
- MLE can be obtained by solving a convex program
- MLE is a consistent estimator of θ

Infinite Dimensional Generalization

$$\mathcal{P} = \left\{ p_f(x) = e^{f(x) - A(f)} q_0(x),\ x \in \Omega\ :\ f \in \mathcal{F} \right\}$$

where

$$\mathcal{F} = \left\{ f \in \mathcal{H} : A(f) = \log \int e^{f(x)} q_0(x)\, dx < \infty \right\}$$

- Finite dimensional family: $\mathcal{H} = \{f : f = \sum_{i=1}^{m} \theta_i T(\cdot)\}$.
- Separable Hilbert space: $\mathcal{H} = \{\sum_{i \in I} \eta_i \phi_i\}$, $(\eta_i) \subset \mathbb{R}$, (ϕ_i) ONB.
- (Pistone and Sempi, 1995): $\mathcal{H}$ is an Orlicz space.
- (Cule et al., 2010): $\mathcal{H}$ is the space of concave functions $\Rightarrow \mathcal{P}$ is the space of log concave densities on Ω.
- (Canu and Smola, 2005; Fukumizu, 2009): $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is said to be an RKHS with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall\, x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$
  - $\forall x \in \mathcal{X}, \forall\, f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$
- k is the reproducing kernel (r.k.) of $\mathcal{H}$ as

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}},\ x, y \in \mathcal{X}.$$

- Every r.k. is a positive definite function, i.e.,
$\forall\, n \geq 1, \forall\, (a_1, \ldots, a_n) \in \mathbb{C}^n, \forall\, (x_1, \ldots, x_n) \in \mathcal{X}^n,$

$$\sum_{i,j=1}^{n} a_i \bar{a}_j k(x_i, x_j) \geq 0.$$

- For every positive definite function, k on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS, $\mathcal{H}$ with k as its r.k.

Properties of RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$
  - Example: $f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)$ for arbitrary $m \in \mathbb{N}$, $\{\alpha_i\} \subset \mathbb{R}$, $x \in \mathcal{X}$ and $\{x_i\} \subset \mathcal{X}$.
- k is bounded if and only every $f \in \mathcal{H}$ is bounded.
- If $\sqrt{k(x, x)}$ is p-integrable, then $\mathcal{H}$ consists of p-integrable functions.
- Every $f \in \mathcal{H}$ is continuous if and only if $k(\cdot, x)$ is continuous for all $x \in \mathcal{X}$.
- Every $f \in \mathcal{H}$ is m-times continuously differentiable if k is m-times continuously differentiable.

Infinite Dimensional Exponential Family

- $\mathcal{H}$ is an RKHS:

$$\mathcal{P} = \left\{ p_f(x) = e^{\langle f, k(\cdot,x)\rangle_{\mathcal{H}} - A(f)} q_0(x),\ x \in \Omega, f \in \mathcal{F} \right\}$$

where

$$\mathcal{F} = \left\{ f \in \mathcal{H} : A(f) = \log \int e^{f(x)} q_0(x)\, dx < \infty \right\}.$$

The relation to natural parameter exponential family is quite evident.

- Finite dimensional RKHS: one-to-one correspondence between finite dimensional exponential family and RKHS.
- $T(x) \rightsquigarrow k(x,y) = \langle T(x), T(y)\rangle$. Similarly, $k(x,y) = \langle \Phi(x), \Phi(y)\rangle \rightsquigarrow \Phi(x)$.

Examples

Exponential: $\Omega = \mathbb{R}_{++}$, $k(x, y) = xy$.

Normal: $\Omega = \mathbb{R}$, $k(x, y) = xy + x^2 y^2$.

Beta: $\Omega = (0, 1)$, $k(x, y) = \log x \log y + \log(1 - x) \log(1 - y)$.

Gamma: $\Omega = \mathbb{R}_{++}$, $k(x, y) = \log x \log y + xy$.

Inverse Gaussian: $\Omega = \mathbb{R}_{++}$, $k(x, y) = xy + \frac{1}{xy}$.

Poisson: $\Omega = \mathbb{N} \cup \{0\}$, $k(x, y) = xy$, $q_0(x) = (x!\, e)^{-1}$.

Geometric: $\Omega = \mathbb{N} \cup \{0\}$, $k(x, y) = xy$, $q_0(x) = 1$.

Binomial: $\Omega = \{0, \ldots, m\}$, $k(x, y) = xy$, $q_0(x) = 2^{-m} \binom{m}{c}$.

Approximation of Densities by $\mathcal{P}$

Theorem

Define

$$\mathcal{P}_0 := \left\{ \pi_f(x) = \frac{e^{f(x)} q_0(x)}{\int e^{f(x)} q_0(x)\, dx},\ f \in C_0(\Omega) \right\}.$$

Suppose

$(*)$ $k(\cdot, x) \in C_0(\Omega)$ *for all* $x \in \Omega \subset \mathbb{R}^d$;

$(**)$ $\int \int k(x,y) d\mu(x)\, d\mu(y) > 0$ *for all* $\mu \in M_b(\Omega)\backslash\{0\}$;

Then $\mathcal{P}$ *is dense in* $\mathcal{P}_0$ *w.r.t. Kullback-Leibler divergence, total variation and Hellinger distances.*

If $q_0 \in L^1(\Omega) \cap L^r(\Omega)$ *for some* $1 \leq r \leq \infty$, *then* $\mathcal{P}$ *is dense in* $\mathcal{P}_0$ *in* L^r *norm.*

Approximation of Densities by $\mathcal{P}$

Corollary

Define

$$\mathcal{P}_{cc} := \left\{ p \in C(\Omega) : \int_{\Omega} p(x)\, dx = 1, p(x) \geq 0, \forall\, x \in \Omega \right\}$$

where Ω is a non-empty compact subset of $\mathbb{R}^d$. Suppose k is continuous, satisfies $(*)$ and $(**)$ and q_0 is a uniform distribution on Ω.

Then $\mathcal{P}$ is dense in $\mathcal{P}_{cc}$ w.r.t. KL divergence, Hellinger and L^r distances for any for all $1 \leq r \leq \infty$.

Mean Element and Covariance Operator

Analogous to the finite dimensional case, the first (along direction u) and second Fréchet derivatives (along v_1 and v_2) of $A(f)$ yields

- $D_f A(f)(u) = \mathbb{E}_{p_f}(u) = \langle u, m_{p_f} \rangle_{\mathcal{H}}$
- $D_f^2 A(f)(v_1, v_2) = \mathrm{Cov}_{p_f}[v_1(X), v_2(X)] = \langle v_1, \Sigma_{p_f} v_2 \rangle_{\mathcal{H}}$

where

- $m_{p_f} := \int k(\cdot, x)\, p_f(x)\, dx$ is the mean element of k
- Σ_{p_f} is the covariance operator

These objects are particularly useful in applications such as homogeneity and independence testing.

Problem: Given random samples, $X_1, \ldots, X_n$ drawn i.i.d. from an unknown density, $p_0 := p_{f_0} \in \mathcal{P}$, estimate p_0.

Maximum Likelihood Estimation

$$\begin{aligned} f_{ML} &= \arg \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} \log p_f(X_i) \\ &= \arg \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) - n \log \int e^{f(x)} q_0(x)\, dx. \end{aligned}$$

Solving the above yields that f_{ML} satisfies

$$\frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) = \int k(\cdot, x) p_{f_{ML}}(x)\, dx$$

i.e.,

$$\int k(\cdot, x)\, d \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} - \mathbb{P}_{f_{ML}} \right) = 0$$

where $p_{f_{ML}} = \frac{d\mathbb{P}_{f_{ML}}}{dx}$.

Ill-posed!!

Maximum Likelihood Estimation

- Finite dimensional case: Elegant likelihood equations to be solved.
- (Fukumizu, 2009): a sieves method involving pseudo-MLE by restricting $\mathcal{P}$ to a series of finite dimensional submanifolds, which enlarge as the sample size increases.
- The estimator is consistent but cannot be actually computed in practice. Also, the rates are not known.

Kernel density estimation: ignores the structure of $\mathcal{P}$ and is simple to compute, but does not perform well for moderate to large d.

Fisher Divergence

- Score matching (Hyvärinen, 2005)
- Motivated by the simple observation that MLE can be intractable even in finite dimensions if $A(\theta)$ is not easily computable.
- Assuming p_θ to be differentiable (w.r.t. x) and $\int p_0(x) \| \frac{\partial \log p_\theta(x)}{\partial x} \|^2 \, dx < \infty$, $\forall\, \theta \in \Theta$:

$$
\begin{aligned}
D_F(p_0 \| p_\theta) := J(\theta) \quad &:= \quad \frac{1}{2} \int p_0(x) \left\| \frac{\partial \log p_0(x)}{\partial x} - \frac{\partial \log p_\theta(x)}{\partial x} \right\|^2 dx \\
&\overset{(a)}{=} \quad \int p_0(x) \sum_{i=1}^{d} \left(\frac{1}{2} \left(\frac{\partial \log p_\theta(x)}{\partial x_i} \right)^2 + \frac{\partial^2 \log p_\theta(x)}{\partial x_i^2} \right) \\
&\qquad + \frac{1}{2} \int p_0(x) \left\| \frac{\partial \log p_0(x)}{\partial x} \right\|^2 dx,
\end{aligned}
$$

where partial integration is used in (a) under the condition that

$$
p_0(x) \frac{\partial \log p_\theta(x)}{\partial x_i} \to 0 \text{ as } x_i \to \pm\infty, \; \forall\, i = 1, \ldots, d.
$$

Empirical Estimator

$$J_n(\theta) := \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d}\left(\frac{1}{2}\left(\frac{\partial \log p_\theta(X_a)}{\partial x_i}\right)^2 + \frac{\partial^2 \log p_\theta(X_a)}{\partial x_i^2}\right) + \frac{1}{2}\int p_0(x)\left\|\frac{\partial \log p_0(x)}{\partial x}\right\|^2 dx.$$

Since $J_n(\theta)$ is independent of $A(\theta)$,

$$\theta_n^* = \arg\min_{\theta\in\Theta} J_n(\theta)$$

should be easily computable, unlike the MLE.

Interpretation

- KL divergence: $KL(p_0 \| p_\theta) = \int p_0(x) \log \frac{p_0(x)}{p_\theta(x)}\, dx$
- Fisher divergence: $D_F(p_0 \| p_\theta) = \int p_0(x) \left\| \frac{\partial}{\partial x} \log \frac{p_0(x)}{p_\theta(x)} \right\|^2 dx$

Theorem (Lyu, 2011)

Let $p_\sigma = p * N(0,1)$ and $q_\sigma = q * N(0,1)$. Assuming p_σ and q_σ to be smooth and fast decaying, we have

$$\frac{d}{d\sigma} KL(p_\sigma \| q_\sigma) = -D_F(p_\sigma \| q_\sigma)$$

and

$$\frac{d}{d\sigma} KL(p_\sigma \| q_\sigma)\Big|_{\sigma=0} = -D_F(p \| q).$$

Therefore

$$KL(p \| q) = \int_0^\infty D_F(p_\sigma \| q_\sigma)\, d\sigma.$$

Assumptions

(A) $\Omega := \prod_{j=1}^{d}(a_j, b_j) \subset \mathbb{R}^d$ where $a_j < b_j$ and $a_j, b_j \in \mathbb{R} \cup \{\pm\infty\}$ $(j = 1, \ldots, d)$.

(B) k is twice continuously differentiable on $\Omega \times \Omega$.

(C) (Boundary conditions) For any fixed $i \in \{1, \ldots, d\}$ and $\tilde{x}_k \in (a_k, b_k)$ $(k \neq i)$, let $x = (\tilde{x}_1, \ldots, \tilde{x}_{i-1}, x_i, \tilde{x}_{i+1}, \ldots, \tilde{x}_d)$. Then

$$\lim_{x_i \to a_i+ \text{ or } b_i-} \frac{\partial^2 k(x, y)}{\partial x_i \partial y_i}\Big|_{y=x} p_0^2(x) = 0.$$

(D) (Well-definedness of D_F)

$$\left\| \frac{\partial k(\cdot, x)}{\partial x_i} \right\|_{\mathcal{H}} < \infty, \quad \left\| \frac{\partial^2 k(\cdot, x)}{\partial x_i^2} \right\|_{\mathcal{H}} < \infty \text{ and } \left\| \frac{\partial k(\cdot, x)}{\partial x_i} \right\|_{\mathcal{H}} \frac{\partial \log q_0(x)}{\partial x_i} < \infty$$

$\forall\, i = 1, \ldots, d.$

Examples: Gaussian and Matérn kernels.

Density Estimation in $\mathcal{P}$: Our Results

Theorem

Under **(B)**, **(C)** *and* **(D)**, *the following hold.*

(i) For all $f \in \mathcal{F}$,

$$J(f) = \frac{1}{2} \langle f - f_0, C(f - f_0) \rangle_{\mathcal{H}} ,$$

where

$$C := \int p_0(x) \sum_{i=1}^{d} \frac{\partial k(\cdot, x)}{\partial x_i} \otimes \frac{\partial k(\cdot, x)}{\partial x_i} \, dx$$

is a trace-class positive operator (and therefore Hilbert-Schmidt and compact).

Density Estimation in $\mathcal{P}$: Our Results

Theorem (cntd..)

(ii) Alternatively,

$$J(f) = \frac{1}{2}\langle f, Cf\rangle_{\mathcal{H}} + \langle f, \xi\rangle_{\mathcal{H}} + c_0,$$

where

$$\xi := \int p_0(x) \sum_{i=1}^{d} \left(\frac{\partial k(\cdot, x)}{\partial x_i} \frac{\partial \log q_0(x)}{\partial x_i} + \frac{\partial^2 k(\cdot, x)}{\partial x_i^2} \right) dx$$

and c_0 is a constant. In addition, f_0 satisfies

$$Cf_0 = -\xi.$$

Density Estimation in $\mathcal{P}$: Our Results

Theorem (cntd...)

(iii) Given samples $\{X_a\}_{a=1}^n$ *drawn i.i.d. from* p_0, *for any* $\lambda > 0$, *the unique minimizer* $f_{\lambda,n}$ *of*

$$J_n(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

exists and is given by

$$f_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\xi},$$

where J_n *is obtained by replacing* C *and* ξ *by* $\hat{C}$ *and* $\hat{\xi}$ *respectively. Here* $\hat{C}$ *and* $\hat{\xi}$ *are the empirical estimators of* C *and* ξ *respectively.*

Computation of $f_{\lambda,n}$

Theorem

$$f_{\lambda,n} = \alpha\hat{\xi} + \sum_{b=1}^{n}\sum_{j=1}^{d} \beta_{bj} \frac{\partial k(\cdot, X_b)}{\partial x_j},$$

where α *and* (β_{bj}) *are obtained by solving*

$$H \begin{pmatrix} \alpha \\ \beta_{ai} \end{pmatrix} = - \begin{pmatrix} \|\hat{\xi}\|^2 \\ h_j^b \end{pmatrix}$$

with $G_{ij}^{ab} = \frac{\partial^2 k(X_a, X_b)}{\partial x_i \partial y_j}$, $\ell(x) = \log q_0(x)$,

$$H = \begin{pmatrix} \frac{1}{n}\sum_{ai}(h_i^a)^2 + \lambda\|\hat{\xi}\|^2 & \frac{1}{n}\sum_{ai} h_i^a G_{ij}^{ab} + \lambda h_j^b \\ \frac{1}{n}\sum_{ai} h_i^a G_{ij}^{ab} + \lambda h_j^b & \frac{1}{n}\sum_{cm} G_{im}^{ac} G_{mj}^{cb} + \lambda G_{ij}^{ab} \end{pmatrix},$$

and

$$h_j^b = \frac{1}{n}\sum_{ia} \frac{\partial^3 k(X_a, X_b)}{\partial x_i^2 \partial y_j} + \frac{\partial^2 k(X_a, X_b)}{\partial x_i \partial y_j} \frac{\partial \ell(X_a)}{\partial x_i}.$$

Consistency and Rates for $f_{\lambda,n}$

Theorem

Suppose **(B)**, **(C)** *and* **(D)** *hold.*

(i) *(Consistency)* *If* $f_0 \in \overline{\mathcal{R}(C)}$, *then*

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} \to 0 \text{ as } \lambda\sqrt{n} \to \infty,\ \lambda \to 0, \text{ and } n \to \infty.$$

(ii) *(Rates of convergence)* *Suppose* $f_0 \in \mathcal{R}(C^\beta)$ *for some* $\beta > 0$. *Then*

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}\left(n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2(\beta+1)}\right\}}\right)$$

with $\lambda = n^{-\max\left\{\frac{1}{4}, \frac{1}{2(\beta+1)}\right\}}$ *as* $n \to \infty$.

(iii) *(Finite dimension)* *Suppose* $\|C^{-1}\| < \infty$. *Then*

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}\left(n^{-\frac{1}{2}}\right)$$

with $\lambda = n^{-\frac{1}{2}}$ *as* $n \to \infty$.

Consistency and Rates for $p_{f_{\lambda,n}}$

Theorem

Suppose **(B)**, **(C)** *and* **(D)** *hold with* $\sup_{x\in\Omega} k(x,x) < \infty$ *and* $supp(q_0) = \Omega$.

For any $1 \leq r \leq \infty$ *with* $q_0 \in L^r(\Omega) \cap L^1(\Omega)$,

$$\|p_{f_{\lambda,n}} - p_0\|_{L^r(\Omega)} \to 0 \text{ and } KL(p_0 \| p_{f_{\lambda,n}}) \to 0 \text{ as } \lambda\sqrt{n} \to \infty,\ \lambda \to 0$$

and $n \to \infty$. *In addition,*

$$D_F(p_0 \| p_{f_{\lambda,n}}) \to 0 \text{ as } \lambda n \to \infty,\ \lambda \to 0 \text{ and } n \to \infty.$$

Consistency and Rates for $p_{f_{\lambda,n}}$

Theorem (cntd..)

Suppose $f_0 \in \mathcal{R}(C^\beta)$ *for some* $\beta > 0$, *then*

$$\|p_{f_{\lambda,n}} - p_0\|_{L^r(\Omega)} = O_{p_0}(\theta_n), \; KL(p_0 \| p_{f_{\lambda,n}}) = O_{p_0}(\theta_n^2)$$

with $\lambda = n^{-\max\left\{\frac{1}{4}, \frac{1}{2(\beta+1)}\right\}}$ *and* $\theta_n := n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2(\beta+1)}\right\}}$. *Also,*

$$D_F(p_0 \| p_{f_{\lambda,n}}) = O_{p_0}\left(n^{-\min\left\{\frac{2}{3}, \frac{2\beta+1}{2(\beta+1)}\right\}}\right)$$

with $\lambda = n^{-\max\left\{\frac{1}{3}, \frac{1}{2(\beta+1)}\right\}}$. *If* $\|C^{-1}\| < \infty$, *then* $\theta_n = n^{-\frac{1}{2}}$ *and*

$$D_F(p_0 \| p_{f_{\lambda,n}}) = O_{p_0}(n^{-1})$$

with $\lambda = n^{-\frac{1}{2}}$.

Range Space Assumption

$$\mathcal{R}(C^\beta) = \left\{ \sum_{i \in \mathbb{N}} c_i \phi_i \, : \, \sum_{i \in \mathbb{N}} \frac{c_i^2}{\alpha_i^{2\beta}} < \infty \right\}$$

where (α_i) are the positive eigenvalues of C, (ϕ_i) are the corresponding eigenfunctions that form an ONB for $\mathcal{R}(C)$ with $\alpha_i \to 0$ as $i \to \infty$.

- $\mathcal{R}(C^{\beta_1}) \subset \mathcal{R}(C^{\beta_2})$ for $0 < \beta_2 < \beta_1 < \infty$ (Hilbert scales) and $id : \mathcal{R}(C^{\beta_1}) \to \mathcal{R}(C^{\beta_2})$ is continuous.
- Interpolation property

Range Space Assumption

Proposition

- $k(x, y) = \psi(x - y), x, y \in \mathbb{R}^d$ *is the reproducing kernel of* $\mathcal{H}$
- $l(x, y) = \phi(x - y), x, y \in \mathbb{R}^d$ *is the reproducing kernel of* $\mathcal{G}$
- $\psi, \phi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ *are positive definite.*

For $1 \leq r \leq \infty$, *suppose the following hold:*

(i) $\int \|\omega\|^2 \hat{\psi}(\omega)\, d\omega < \infty$;

(ii) $\left\| \frac{\hat{\phi}}{\hat{\psi}} \right\|_\infty < \infty$;

(iii) $\frac{\|\cdot\|^2 \hat{\psi}^2}{\hat{\phi}} \in L^{\frac{r}{2-r}}(\mathbb{R}^d)$;

(iv) $\|q_0\|_{L^r(\mathbb{R}^d)} < \infty$.

Then $f_0 \in \mathcal{R}(C)$ *implies* $f_0 \in \mathcal{G} \subset \mathcal{H}$.

Range Space Assumption

Gaussian RKHS:

$$\mathcal{H}_\sigma = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \int |\hat{f}(\omega)|^2 e^{\frac{\|\omega\|^2}{4\sigma}} \, d\omega < \infty \right\}.$$

$f_0 \in \mathcal{R}(C)$ implies f_0 at least lies in $\mathcal{H}_{\frac{\sigma}{2}+\epsilon}$ for some arbitrary small $\epsilon > 0$.

Sobolev space with $s > 1 + \frac{d}{2}$:

$$\mathcal{H}_2^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \int (1 + \|\omega\|^2)^s |\hat{f}(\omega)|^2 \, d\omega < \infty \right\}.$$

If $f_0 \in \mathcal{R}(C)$, then f_0 should at least lie in $\mathcal{H}_2^{2s-1-\frac{d}{2}-\epsilon}(\mathbb{R}^d)$ for arbitrarily small $\epsilon > 0$. Suppose $f_0 \notin \mathcal{H}_2^\alpha(\mathbb{R}^d)$ for $\alpha \geq 2s - 1 - \frac{d}{2}$. Then the rate of $n^{-1/4}$ is minimax optimal if the RKHS, $\mathcal{H}$ is chosen to be $\mathcal{H}_2^{d+2}(\mathbb{R}^d)$ when d is even and to be $\mathcal{H}_2^{d+\frac{3}{2}}(\mathbb{R}^d)$ when d is odd.

Choice of Regularizer

$$f_0 \in \mathcal{R}(C^\beta) \Leftrightarrow f_0 \in \left\{ g \in \mathcal{H} : \sum_{i=1}^{\infty} \frac{\langle \phi_i, g \rangle_{\mathcal{H}}^2}{\alpha_i^{2\beta}} < \infty \right\}$$

where $(\phi_i)_i$ and $(\alpha_i)_i$ are the eigenfunctions and eigenvalues of C.

- As β increases, $\langle \phi_i, g \rangle_{\mathcal{H}}$ should decrease faster with i, i.e., f_0 becomes increasingly smoother with β.

Not reflected in the rates!!

Better Regularizers: Improved Rates

- $f_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\xi} = -g_\lambda(\hat{C})\,\hat{\xi}$ where

$$g_\lambda : \mathbb{R}_+ \to \mathbb{R}_+ \qquad \alpha \mapsto (\alpha + \lambda)^{-1}.$$

- $f_{\lambda,n}$ is obtained by using the RKHS regularizer, $\frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$.
- Instead, directly construct $f_{\lambda,n} = -g_\lambda(\hat{C})\,\hat{\xi}$ which corresponds to using a data-dependent regularizer,

$$\frac{1}{2}\left\langle f, \left((g_\lambda(\hat{C}))^{-1} - \hat{C}\right) f \right\rangle_{\mathcal{H}}.$$

- With appropriate conditions on g_λ, almost parametric rate of $n^{-1/2}$ can be obtained as $\beta \to \infty$.
- Examples: $g_\lambda(\alpha) = \alpha^{-1}(1 - e^{-\alpha/\lambda})$, $\alpha > 0$; $g_\lambda(\alpha) = \alpha^{-1}\mathbb{1}_{\alpha \geq \lambda}$ (spectral cut-off).

Mis-specified Case

Theorem

Let $p_0, q_0 \in C^1(\Omega)$ *be probability densities such that*

$$\int p_0 \left\| \frac{\partial \log \frac{p_0}{q_0}}{\partial x} \right\|^2 dx < \infty.$$

Assume that **(B)**, **(C)**, **(D)** *hold with* $\sup_{x \in \Omega} k(x,x) < \infty$ *and* $supp(q_0) = \Omega$. *Suppose there exists* $f^* \in \mathcal{F}$ *such that* $D_F(p_0 \| p_{f^*}) = \inf_{p \in \mathcal{P}} D_F(p_0 \| p)$. *Then*

$$D_F(p_0 \| p_{f_{\lambda,n}}) \to \inf_{p \in \mathcal{P}} D_F(p_0 \| p) \text{ as } \lambda \to 0,\ \lambda n \to \infty \text{ and } n \to \infty.$$

In addition, if $f^* \in \mathcal{R}(C^\beta)$ *for some* $\beta \geq 0$, *then*

$$\sqrt{D_F(p_0 \| p_{f_{\lambda,n}})} \leq \sqrt{\inf_{p \in \mathcal{P}} D_F(p_0 \| p)} + O_{p_0}\left(n^{-\min\left\{\frac{1}{3}, \frac{2\beta+1}{4(\beta+1)}\right\}}\right)$$

with $\lambda = n^{-\max\left\{\frac{1}{3}, \frac{1}{2(\beta+1)}\right\}}$. *If* $\|C^{-1}\| < \infty$, *then the rate is* $n^{-1/2}$.

Experiments

Score matching vs. KDE for estimating $N(0, I_d)$:

- $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} + r(x^T y + c)^2$ where $r = 0.1$ and $c = 0.5$.
- σ and λ are chosen by cross-validation of the objective function $J_n(f) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$.
- For the KDE, the Gaussian kernel is used for the smoothing kernel, and the bandwidth parameter is chosen by cross-validation.

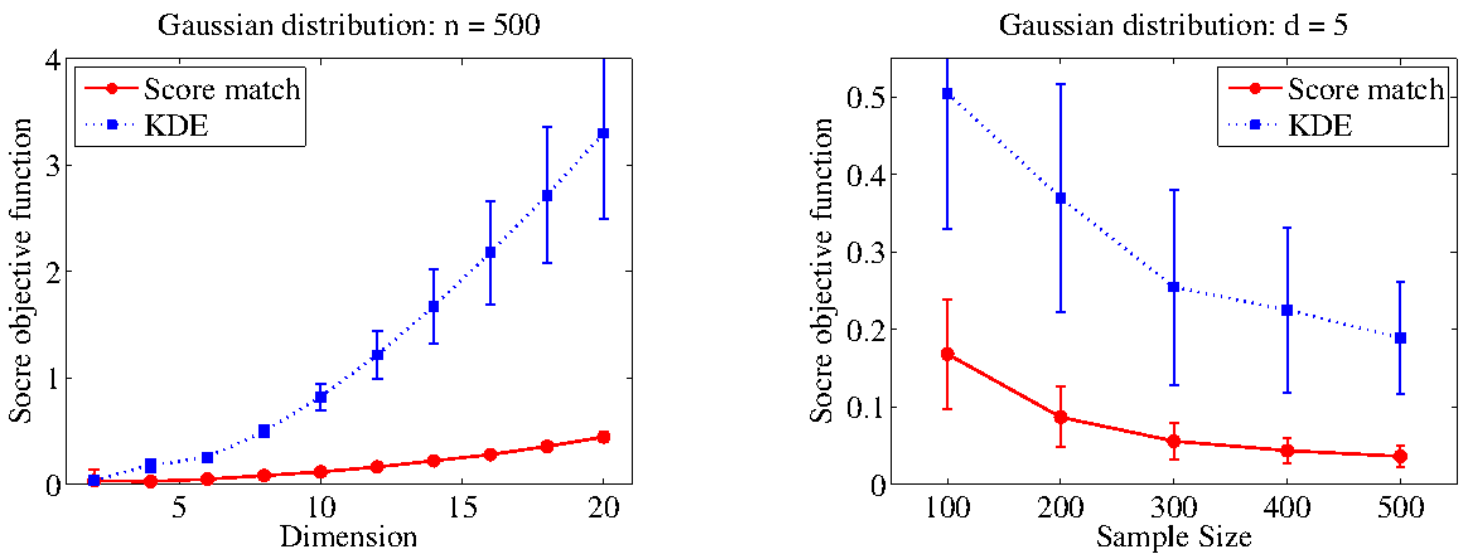

Figure: Estimation errors (Fisher divergence). Left: different dimensions with $n = 500$. Right: different sample size with $d = 5$.

Experiments

Gaussian mixture:

$$p_0(x) = \frac{1}{2}\phi_d(x; 4\mathbf{1}_n, I_d) + \frac{1}{2}\phi_d(x; -4\mathbf{1}_n, I_d)$$

where $\phi_d(x, \mu, \Sigma)$ is the p.d.f. of $N_d(\mu, \Sigma)$.

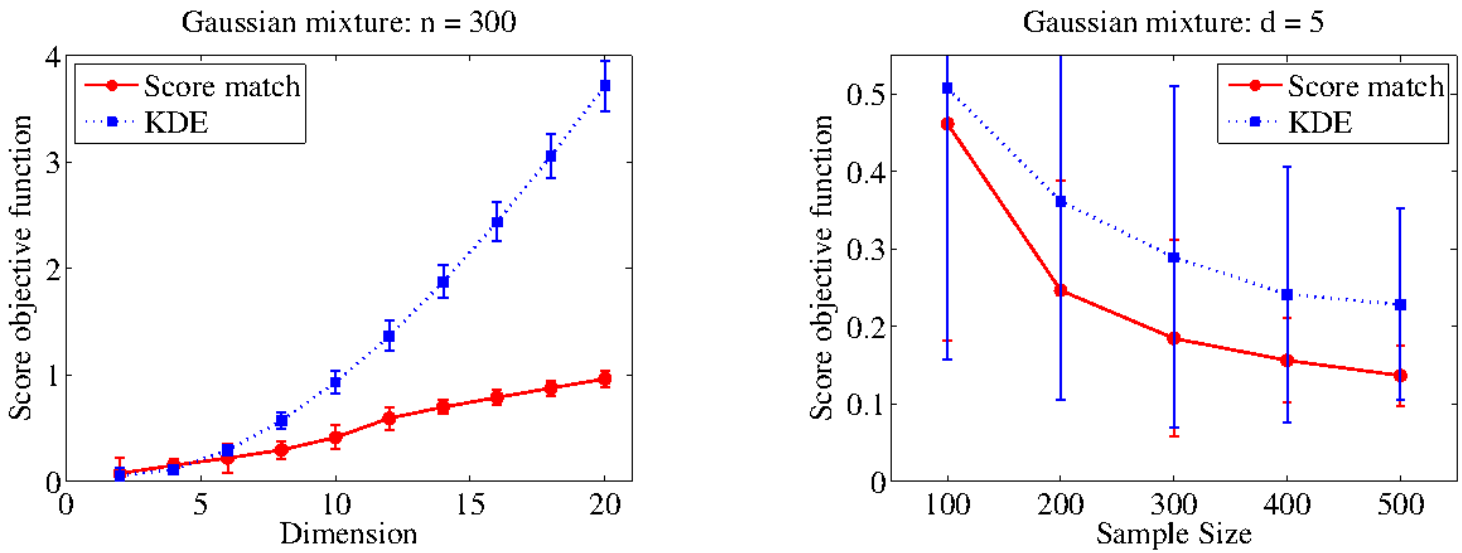

Figure: Estimation errors (Fisher divergence). Left: different dimensions with $n = 300$. Right: different sample size with $d = 5$.

Open Questions

- Minimax optimality of the estimator
- Adaptive estimation

Summary

- RKHS-induced infinite dimensional exponential families
  - Natural generalization of finite dimensional exponential family
  - Approximates a wide class of densities
- Density estimation
  - Maximum likelihood approach is not feasible
  - Kernel density estimation does not perform well in practice
  - Estimation through minimization of Fisher divergence
  - Recovery of parametric rates in finite dimensions
  - Improved rates with an appropriate choice of regularizer

Thank You