Estimation with Infinite Dimensional Kernel Exponential Families

Kenji Fukumizu
The Institute of Statistical Mathematics

Joint work with Bharath Sriperumbudur (Penn State U), Arthur Gretton (UCL), Aapo Hyvarinen (U Helsinki), Revant Kumar (Georgia Tech)

IGAIA IV.

June 12-17, 2016. Liblice, Czech Republic

Introduction

Infinite dimensional exponential family

- (Finite dim.) exponential family

$$p_\theta(x) = \exp\left(\sum_{j=1}^{m} \theta_j T_j(x) - A(\theta)\right) q_0(x)$$

- Infinite dimensional extension?

$$p_f(x) = \exp\big(f(x) - A(f)\big)\, q_0(x) \quad \text{where } A(f) := \log \int e^{f(x)} q_0(x) dx$$

f is a natural parameter in an infinite dimensional function class.

– Maximal exponential model (Pistone & Sempi AoS 1995):
  - Orlicz space (Banach sp.) is used.
  - Estimation is not at all obvious.

    "Empirical" mean parameter cannot be defined.

Kernel exponential manifold (Fukumizu 2009; Canu & Smola 2005)

Reproducing kernel Hilbert space is used.

- $p_f(x) = \exp\left(\langle \underline{f}, \underline{k(\cdot, x)} \rangle - A(f) \right) q_0(x)$

  Parameter — f

  Infinite dimensional sufficient statistics — $k(\cdot, x)$

- Empirical estimation is possible
  - Mean parameter: $m_f = E_{p_f}[k(\cdot, X)]$
  - Maximum likelihood estimator: $\widehat{m}_f = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i)$

- Manifold structure can be defined (Fukumizu 2009)

Problems in estimation

- Normalization constant / partition function
  - Even in finite dim. cases
    $$A(\theta) := \log \int e^{\sum_{j=1}^{m} \theta_j T_j(x)} q_0(x) dx$$
    is not easy to compute.
  - MLE: "Mean parameter → natural parameter" needs to solve
    $$\frac{\partial A(\theta)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} T(X_i).$$
  - Even more difficult for an infinite dimensional exponential family

- This talk → score matching (Hyvarinen, JMLR 2005)
  - Estimation method without normalization constants.
  - Introducing a new method for (unnormalized) density estimation.

Score Matching

Score matching for exponential family

(Hyvärinen, JMLR2005)

Fisher divergence

p, q: two p.d.f.'s on $\Omega = \prod_{a=1}^{d}(s_a, t_a) \subset (\mathbf{R} \cup \{\pm\infty\})^d$.

$$J(p||q) := \frac{1}{2}\int \sum_{a=1}^{d} \left| \frac{\partial \log p(x)}{\partial x_a} - \frac{\partial \log q(x)}{\partial x_a} \right|^2 p(x)dx$$

– $J(p||q) \geq 0$. Equality holds iff $p = q$ (under mild conditions).

– Derivative w.r.t. x, not parameter.

- For location parameter $p(x) = f(x - \theta)$,

$$\frac{\partial \log p(x)}{\partial x_a} = -\frac{\partial \log f_\theta(x)}{\partial \theta_a}$$

$J(p||q)$ = squared L^2-distance of Fisher scores.

Set $p = p_0$ (true), and $q = p_\theta$ to be estimated.

$$J(\theta) := J(p_0 || p_\theta)$$

$$= \frac{1}{2}\int \sum_{a=1}^{d} \left(\frac{\partial \log p_\theta(x)}{\partial x_a} - \frac{\partial \log p_0(x)}{\partial x_a} \right)^2 p_0(x)dx$$

$$= \frac{1}{2}\int \sum_{a=1}^{d} \left(\frac{\partial \log p_\theta(x)}{\partial x_a} \right)^2 p_0(x)dx + \int \sum_{a=1}^{d} \frac{\partial^2 \log p_\theta(x)}{\partial x_a^2} p_0(x)dx \equiv \tilde{J}(\theta)$$

+ const.

- Assume $\lim_{x_a \to s_a \text{ or } t_a} p_0(x) \frac{\partial \log p_\theta(x)}{\partial x_a} = 0$, and use partial integral

$$\int \frac{\partial \log p_\theta(x)}{\partial x_a} \underbrace{\frac{\partial \log p_0(x)}{\partial x_a} p_0(x)}_{\frac{\partial p_0(x)}{\partial x_a}} dx = \underbrace{\left[p_0(x) \frac{\partial \log p_\theta(x)}{\partial x_a} \right]_{s_a}^{t_a}}_{0} - \int \frac{\partial^2 \log p_\theta(x)}{\partial x_a^2} p_0(x)dx$$

Empirical estimation

$$\tilde{J}(\theta) = \frac{1}{2}\int \sum_{a=1}^{d} \left(\frac{\partial \log p_\theta(x)}{\partial x_a} \right)^2 p_0(x)dx + \int \sum_{a=1}^{d} \frac{\partial^2 \log p_\theta(x)}{\partial x_a^2} p_0(x)dx$$

$X_1, \ldots, X_n$: i.i.d. sample $\sim p_0$.

$$\tilde{J}_n(\theta) = \frac{1}{n} \sum_{a=1}^{d} \sum_{i=1}^{n} \left\{ \frac{1}{2} \left(\frac{\partial \log p_\theta(X_i)}{\partial x_a} \right)^2 + \frac{\partial^2 \log p_\theta(X_i)}{\partial x_a^2} \right\}$$

$\hat{\theta} = \arg\min \tilde{J}_n(\theta)$: Score matching estimator

Score matching for exponential family

– For exponential family $p_\theta(x) = \exp\left(\sum_j \theta_j T_j(x) - A(\theta)\right) q_0(x)$,

$$\tilde{J}_n(\theta)$$
$$= \sum_{i=1}^{n} \sum_{a=1}^{d} \frac{1}{2} \left(\sum_{j=1}^{m} \theta_j \frac{\partial T_j(X_i)}{\partial x_a} + \frac{\partial \log q_0(X_i)}{\partial x_a} \right)^2 + \sum_{j=1}^{m} \theta_j \frac{\partial^2 T_j(X_i)}{\partial x_a^2} + \frac{\partial^2 \log q_0(X_i)}{\partial x_a^2}$$

- No need of $A(\theta)$! (derivative w.r.t. x)

- Quadratic form w.r.t. θ → Solvable!
- In the Gaussian case, $\hat{\theta}$ is the same as MLE.

Kernel Exponential Family

Reproducing kernel Hilbert space

– Def. Ω: set. H: Hilbert space consisting of functions on Ω.

H: reproducing kernel Hilbert space (RKHS), if for any $x \in \Omega$ there is $k_x \in H$ s.t.

$$\langle f, k_x \rangle = f(x) \quad \text{for } \forall f \in H \quad \text{[reproducing property]}$$

– $k(x, y) := k_x(y)$. k is a positive definite kernel, i.e., $k(x, y) = k(y, x)$ and the Gram matrix $\left(k(x_i, x_j)\right)_{ij}$ is positive semidefinite for any $x_1, \ldots, x_n$.

– Moore-Aronszajn theorem: for any positive definite kernel on Ω, there uniquely exists an RKHS s.t. its reproducing kernel is $k(\cdot, x)$.
(One-to-one correspondence between p.d. kernel and RKHS)

– Example of pos. def. kernel on $\mathbf{R}^d$: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$.

Kernel exponential family

Def. k: pos. def. kernel on $\Omega = \prod_{a=1}^{d}(s_a, t_a) \subset (\mathbf{R} \cup \{\pm\infty\})^d$.

H_k: RKHS. q_0: p.d.f. on Ω with $\mathrm{supp}(q_0) = \Omega$.

$F_k \coloneqq \{f \in H_k \mid \int e^{f(x)} q_0(x) dx < \infty\}$ (functional) parameter space

$P_k \coloneqq \{p_f : \Omega \to (0, \infty) \mid$

$$p_f(x) = e^{f(x) - A(f)} q_0(x), \;\; f \in F_k\}$$

where $A(f) \coloneqq \int e^{f(x)} q_0(x) dx$

P_k: kernel exponential family (KEF)

- With finite dimensional H_k, KEF is reduced to a finite dim. exponential family.

  e.g. $k(x, y) = (1 + x^T y)^2$ → Gaussian distributions.

Score matching for KEF

Assume k is of class C^2 ($\partial^{a+b}k(x,y)/\partial^a x \partial^b y$ exists and is continuous for $a+b \le 2$) and

$$\lim_{x_a \to s_a \text{ or } t_a} \frac{\partial^2 k(x,y)}{\partial x_a \partial y_a}\bigg|_{y=x} p_0(x) = 0$$ (for partial integral).

– Score matching objective function

$$\tilde{J}_n(f) := \sum_{i=1}^{n} \sum_{a=1}^{d} \frac{1}{2}\left(\frac{\partial f(X_i)}{\partial x_a} + \frac{\partial \log q_0(X_i)}{\partial x_a}\right)^2 + \frac{\partial^2 f(X_i)}{\partial x_a^2} + \frac{\partial^2 \log q_0(X_i)}{\partial x_a^2}$$

Note $f(X_i) = \langle f, k(\cdot, X_i)\rangle$, $\frac{\partial f(X_i)}{\partial x_a} = \langle f, \frac{\partial k(\cdot,X_i)}{\partial x_a}\rangle$, $\frac{\partial^2 f(X_i)}{\partial x_a^2} = \langle f, \frac{\partial^2 k(\cdot,X_i)}{\partial x_a^2}\rangle$.

$\tilde{J}_n(f)$ is a quadratic form w.r.t. $f \in H$.

– Estimation

$$\hat{C}_n f = \xi_n$$

where

$$\hat{C}_n := \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{d} \frac{\partial k(\cdot, X_i)}{\partial x_a}\langle\frac{\partial k(\cdot, X_i)}{\partial x_a}, *\rangle \ : H_k \to H_k$$

$$\hat{\xi}_n := \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{d}\left\{\frac{\partial k(\cdot, X_i)}{\partial x_a}\frac{\partial \log q_0(X_i)}{\partial x_a} + \frac{\partial^2 k(\cdot, X_i)}{\partial x_a^2}\right\} \in H_k$$

– Regularized estimator

$$\widehat{f_n} = \left(\hat{C}_n + \lambda_n I\right)^{-1}\hat{\xi}_n$$

i.e.,

$$\widehat{f_n} = \mathrm{argmin}_f \ \tilde{J}_n(f) + \lambda_n \|f\|_{H_k}^2$$

Explicit Solution

– Estimator: (from representer theorem)

$$\hat{f}_n = \alpha \hat{\xi}_n + \sum_{j=1}^{n} \sum_{b=1}^{d} \beta_{jb} \frac{\partial k(\cdot, X_j)}{\partial x_b}$$

where

$$\begin{bmatrix} \frac{1}{n}\sum_{a,i}(h_i^a)^2 + \lambda \left\|\hat{\xi}_n\right\|^2 & \frac{1}{n}\sum_{a,i} h_i^a \, G_{ij}^{ab} + \lambda \, h_j^b \\ \frac{1}{n}\sum_{a,i} h_i^a \, G_{ij}^{ab} + \lambda \, h_j^b & \frac{1}{n}\sum_{c,m} G_{im}^{ac} G_{mj}^{bc} + \lambda \, G_{ij}^{ab} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_{ia} \end{bmatrix} = - \begin{bmatrix} \left\|\hat{\xi}_n\right\|^2 \\ h_j^b \end{bmatrix}$$

$$h_j^b = \frac{1}{n}\sum_{i,a} \frac{\partial^3 k(X_i, X_j)}{\partial x_a^2 \partial y_b} + \frac{\partial^2 k(X_i, X_j)}{\partial x_a \partial y_b} \frac{\partial \ell(X_i)}{\partial x_a}, \qquad \left(\frac{\partial \ell(X_i)}{\partial x_a} = \frac{\partial \log q_0(X_i)}{\partial x_a} \right)$$

$$G_{ij}^{ab} = \frac{\partial^2 k(X_i, X_j)}{\partial x_a \partial y_b}, \quad \left\|\hat{\xi}_n\right\|^2 = \frac{1}{n^2}\sum_{ij,ab} \frac{\partial^4 k(X_i, X_j)}{\partial x_a^2 \partial y_b^2} + 2 \frac{\partial^3 k(X_i, X_j)}{\partial x_a^2 \partial y_b} \frac{\partial \ell(X_j)}{\partial x_b} + \frac{\partial^2 k(X_i, X_j)}{\partial x_a \partial y_b} \frac{\partial \ell(X_i)}{\partial x_a} \frac{\partial \ell(X_j)}{\partial x_b}$$

- $\hat{f}_n$ can be taken in Span$\left\{ \frac{\partial k(\cdot, X_j)}{\partial x_b}, \hat{\xi}_n \right\}$.
- Estimator is simply given by solving $(1 + nd)$-dimensional linear equation.

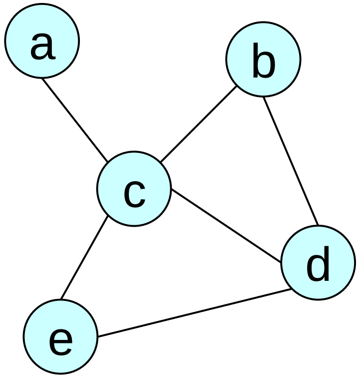
Unnormalized p.d.f.

– Score matching for KEF gives only $f(x)$ or $e^{f(x)}$, unnormalized p.d.f.
  - Estimation of $A(f) \coloneqq \int e^{f(x)} q_0(x) dx$ is yet nontrivial.

– There are interesting applications.
  1) Nonparametric structure learning for graphical model given data (Sun, Kolar, Xu NIPS2015)

$$p(X) \propto \prod_{ij \in E} p_{ij}(X_i, X_j), \qquad G = (V, E)$$

p_{ij} is estimated nonparametrically with KEF (with sparse edges).



2) Hamiltonian Monte Carlo with intractable gradient

(Strathmann et al. NIPS 2015)

Estimate $\frac{\partial \log \pi(x)}{\partial x}$ with EKF, assuming it does not allow a closed form expression (intractable cases).

- Hamiltonian Monte Carlo (Neal 2012)

  Goal: sample from π

  $U(x) = -\log \pi(x)$

  $K(z)$: auxiliary momentum, e.g. $-z^2/\tau^2$

  Hamiltonian $H(z, x) \coloneqq U(x) + K(z)$

  Hamiltonian flow:

$$\frac{dx}{dt} = \frac{\partial H}{dz} = \frac{\partial K}{\partial z},$$

$$\frac{dz}{dt} = -\frac{\partial H}{dx} = \frac{\partial \log \pi(x)}{\partial x}$$

  This flow is used in proposal of MCMC

Convergence

Misspecification

True parameter f_* is taken from a wider space than H_k.

Extended parameter space

$$W_2^0(p_0) := \left\{ f \in C^1(\Omega) \mid \frac{\partial f(x)}{\partial x_a} \in L^2(\Omega; p_0), a = 1, \dots, d \right\} / \sim$$

$$\text{where } f \sim g \Leftrightarrow \sum_{a=1}^{d} \| \partial f / \partial x_a - \partial g / \partial x_a \|_{L^2(p_0)}^2 = 0$$

$$([f],[g])_{W_2^0(p_0)} := \sum_{a=1}^{d} \int \frac{\partial f(x)}{\partial x_a} \frac{\partial g(x)}{\partial x_a} p_0(x) dx.$$

$W_2(p_0)$: completion of the pre-Hilbert space $W_2^0(p_0)$.

- With k is of class C^2 (and other technical conditions), the canonical map

$$I_k : H_k \to W_2(p_0), \qquad f \mapsto [f]$$

defines a (up to constant) embedding of H_k.

Theorem (convergence rate)

Under some assumptions,

(i) If $f_* \coloneqq \log(p_0/q_0) \in \overline{R(I_k I_k^*)}$, with $\lambda_n \to 0, n\lambda_n \to \infty$

$$J(p_0||p_{\hat{f}_n}) \to 0 \;\; (n \to \infty).$$

(ii) If $f_* \in R((I_k I_k^*)^\beta)$ $(0 < \beta \leq 1)$, then with $\lambda_n = n^{-\max\left\{\frac{1}{3}, \frac{1}{2\beta+1}\right\}}$,

$$J(p_0||p_{\hat{f}_n}) = O_p\left(n^{-\min\left\{\frac{2}{3}, \frac{2\beta}{2\beta+1}\right\}}\right).$$

$I_k I_k^*$: operator on $W_2(p_0)$, given by

$$I_k I_k^*[f] = \left[\int \sum_{a=1}^{d} \frac{\partial k(\cdot, x)}{\partial x_a} \frac{\partial f(x)}{\partial x_a} p_0(x) dx\right]$$

Hyperparameter selection

– Hyperparameters
  - Kernel / kernel parameter $\left(k(x,y) = \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)\right)$
  - regularization coefficient

– Cross-validation is possible with the objective function

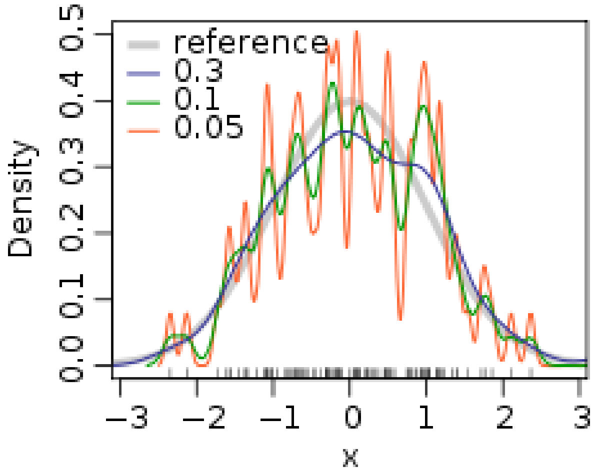
$$\tilde{J}_n(f) \coloneqq \sum_{i=1}^{n} \sum_{a=1}^{d} \frac{1}{2}\left(\frac{\partial f(X_i)}{\partial x_a} + \frac{\partial \log q_0(X_i)}{\partial x_a}\right)^2 + \frac{\partial^2 f(X_i)}{\partial x_a^2} + \frac{\partial^2 \log q_0(X_i)}{\partial x_a^2}.$$

Experiments

Kernel Density Estimation

– KDE: standard nonparametric method for estimating p.d.f.
  Given i.i.d. sample $X_1, \dots, X_n \sim P$

$$\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$



$K(x)$: p.d.f.

e.g. $K(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|x\|^2}{2}\right)$

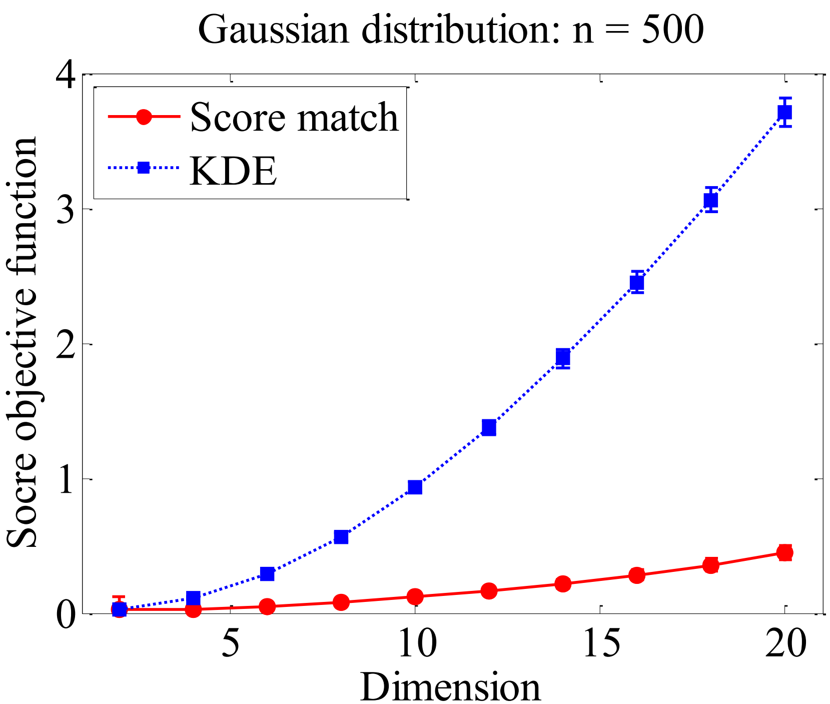
- KDE works well for one-dimensional cases, but weak for high (say, 10) dimensional cases.
- Sensitive to the choice of h_n, (though CV and other methods are applicable).

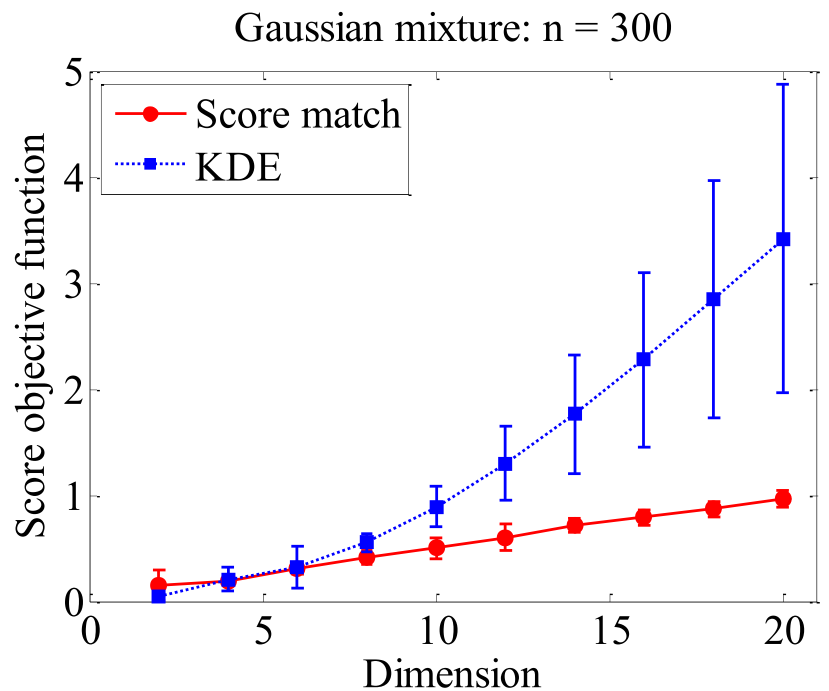
Comparison: EKF vs KDE

- Evaluated by score objective function J

kernel: $k(x,y) = \exp(-\|x-y\|^2/2\sigma^2) + 0.1(x^T y + 0.5)^2$

- Gaussian $p_0 = \phi_d(x; 0, I_d)$
- Gaussian Mixture
  $p_0 = 0.5\phi_d(x; 4\mathbf{1}_d, I_d) + 0.5\phi_d(x; -4\mathbf{1}_d, I_d)$





Evaluated by correlation

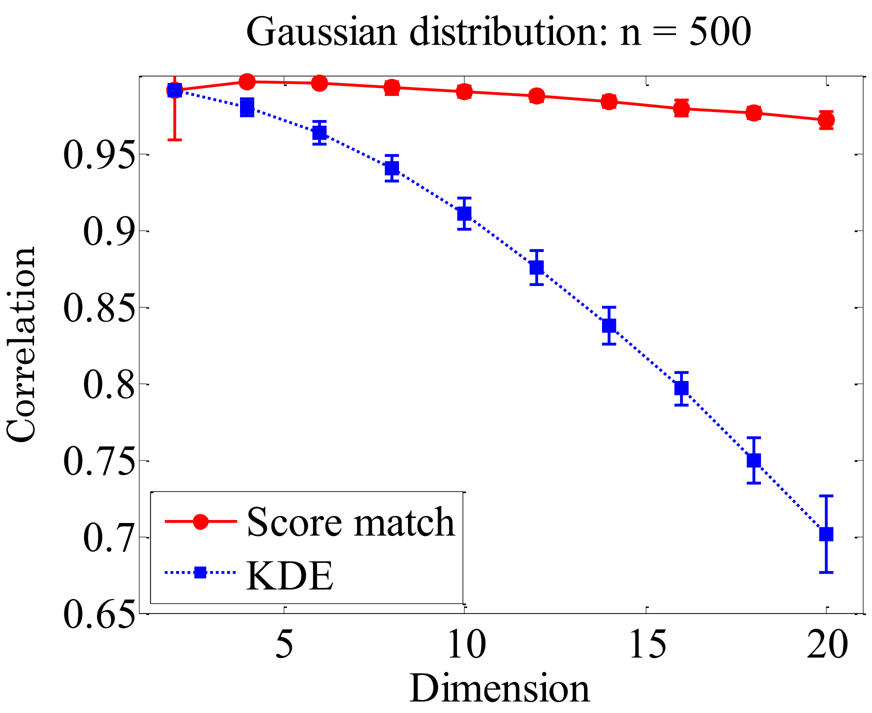
$$Cor(p, p_0) := \frac{E_R[p(Z)p_0(Z)]}{\sqrt{E_R[p(Z)^2]E_R[p_0(Z)^2]}}, \quad Z \sim \frac{1}{10^4}\sum_{i=1}^{10^4} \delta_{X_i}, \quad X_i \underset{i.i.d.}{\sim} p_0 dx$$

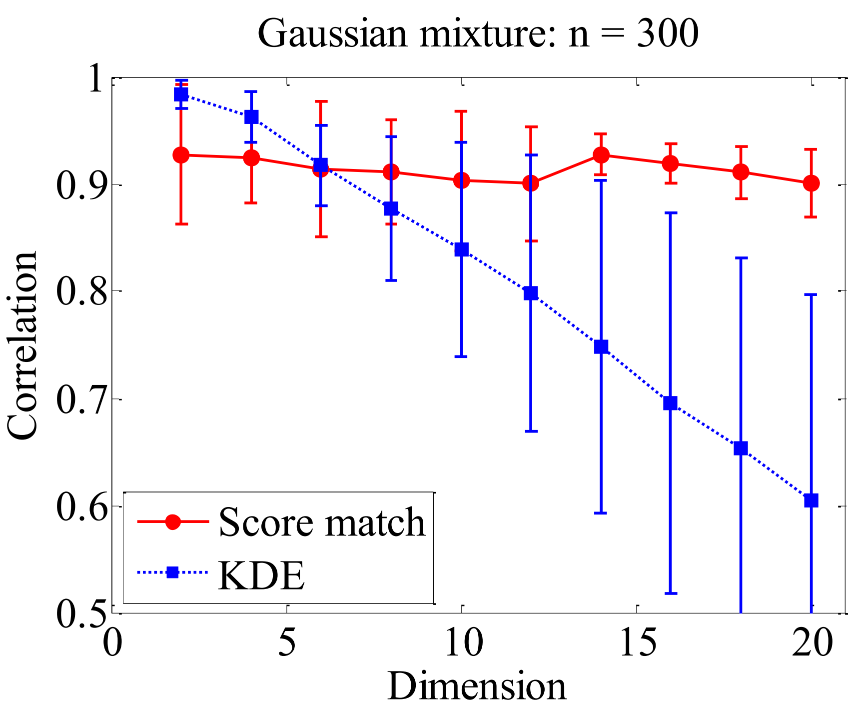
– Gaussian

$$p_0 = \phi_d(x; 0, I_d)$$

– Gaussian Mixture

$$p_0 = 0.5\phi_d(x; 41_d, I_d) + 0.5\phi_d(x; -41_d, I_d)$$





Conclusions

- Infinite dimensional exponential family with positive definite kernel
  - A natural extension of finite dimensional exponential family
  - Sufficient statistics and parameter are given by feature vector $k(\cdot, x)$ and function f, respectively.

- Score matching method gives a tractable estimator for kernel exponential family.
  - No need of computing normalization constants.
  - The estimator is given as a solution to a linear equation.
  - Non-normalized density function is estimated nonparametrically.

Thank you.

Reference

B. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvarinen. Density Estimation in Infinite Dimensional Exponential Families. *arXiv:1312.3516*.