Nonparametric Bayesian Inference with Positive Definite Kernels

Kenji Fukumizu

The Institute of Statistical Mathematics

Graduate University for Advanced Studies

Workshop on Mathematical Approaches to Large-Dimensional Data Analysis

2014 March 13 – 15. ISM, Tokyo

Nonlinearity and high-dimensionality

– Nonlinear / higher-order information in high-dimensional data.
  Biology, documents, social networks, ….

– Extracting nonlinear information of data
  Common practice:

  $(X, Y, Z) \rightarrow (X, Y, Z, X^2, Y^2, Z^2, XY, YZ, ZX, \ldots)$

– Computational problem for high dimensional data

  e.g. Up to the 2nd order for 10,000 dim data

  Dim of feature space:

  ${}_{10000}C_1 + {}_{10000}C_2 = 50{,}005{,}000$ (!)

Nonparametric inference

- Smoothing kernel: KDE, local polynomial fitting

$$h^{-d}K(x/h)$$

- Characteristic function: $E[e^{i\omega X}]$
- Spline, wavelet, order statistics, etc, etc,

→ Curse of dimensionality

- Smoothing kernel: usually not strong for high-dimensional data
- Characteristic function: Integral on high-dimensional spaces is difficult.

→**Kernel method:** a new approach to nonparametric inference. Computationally efficient, good performance for high-dimensional data in theory and practice.

Outline

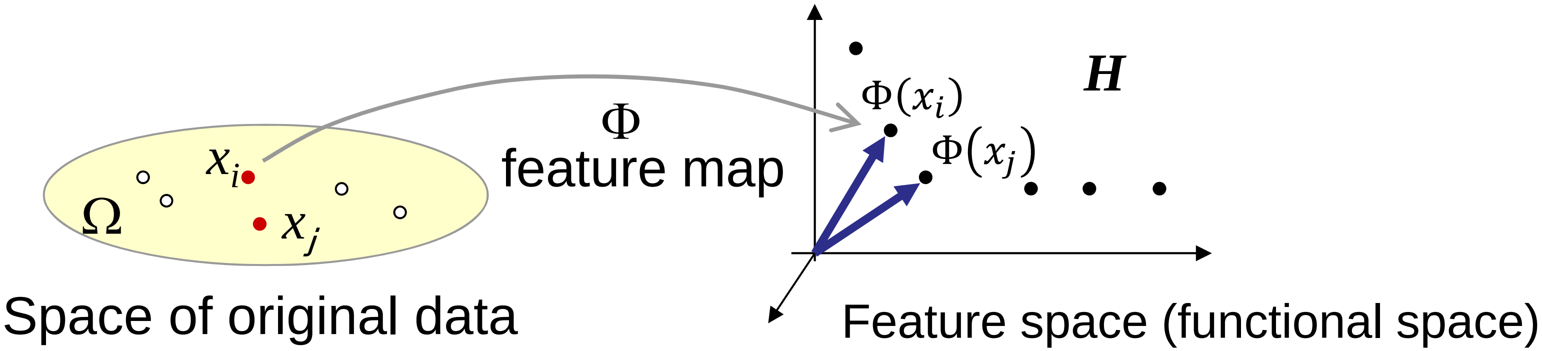
1. Introduction
2. Representing probabilities with kernels
3. Conditional probabilities
4. Kernel methods for Bayesian inference
5. Conclusions

References:

- K. Fukumizu, L. Song, A. Gretton (2014) Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels. *Journal of Machine Learning Research. 14:3753−3783.*
- Song, L., Gretton, A., and Fukumizu, K. (2013) Kernel Embeddings of Conditional Distributions. *IEEE Signal Processing Magazine 30(4), 98-111*

Introduction

Kernel methods: a big picture



Do linear analysis in the feature space.

$$\Phi: \quad \Omega \quad \rightarrow \quad H, \qquad x \mapsto \Phi(x)$$

- Support vector machine is known most .
- This talk: more recent advances for nonparametric inference.

Positive definite kernel

Def. Ω: set. $k: \Omega \times \Omega \to \mathbf{R}$

k is positive definite if k is symmetric, and for any $n \in \mathbf{N}$, $x_1, \dots, x_1 \in \Omega$, $c_1, \dots, c_n \in \mathbf{R}$, the matrix $\left(k(X_i, X_j)\right)_{ij}$ (Gram matrix) satisfies

$$\sum_{i,j=1}^{n} c_i c_j k(X_i, X_j) \geq 0.$$

– Examples on $\mathbf{R}^m$:

- Gaussian RBF kernel $$k_G(x,y) = \exp\left(-\frac{1}{2\sigma^2}||x-y||^2\right) \quad (\sigma > 0)$$
- Laplace kernel $$k_L(x,y) = \exp\left(-\alpha \sum_{i=1}^{m} |x_i - y_i|\right) \quad (\alpha > 0)$$
- Polynomial kernel $$k_P(x,y) = (x^T y + c)^d \qquad (c \geq 0, d \in \mathbf{N})$$

Reproducing kernel Hilbert space

Feature space = reproducing kernel Hilbert space (RKHS)

Positive definite kernel k on Ω uniquely defines a RKHS H_k (Aronzajn 1950).

- Function space: functions on Ω.
- Very special inner product: for any $f \in H_k$

$$\langle f, k(\cdot, x) \rangle = f(x) \qquad \text{(reproducing property)}$$

c.f. L^2 space

- Its dimensionality may be infinite (Gaussian, Laplace).

Note: from reproducing property

$$\langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y)$$

Mapping data into RKHS

– Feature Map

$$\Phi: \Omega \to H_k, \quad x \mapsto k(\cdot, x)$$

– Data transform

$$X_1, \dots, X_n \quad \mapsto \quad \Phi(X_1), \dots, \Phi(X_n):$$ functional data (artificially made)

– Inner product

For $f = \sum_i \alpha_i \Phi(X_i)$, $g = \sum_i \beta_i \Phi(X_i) \in H_k$,

$$\langle f, g \rangle = \sum_{i,j=1}^{n} \alpha_i \beta_j k(X_i, X_j) = \alpha^T G_X \beta$$

Computation with Gram matrices of size n.

Covariance matrix

$$
\begin{array}{cc}
 & n \\
p & \begin{bmatrix} X_1^1 & \cdots & X_n^1 \\ X_1^2 & \cdots & X_n^2 \\ \vdots & \ddots & \vdots \\ X_1^p & \cdots & X_n^p \end{bmatrix}
\end{array}
\begin{bmatrix} X_1^1 & X_1^2 & \cdots & X_1^p \\ \vdots & \vdots & \ddots & \vdots \\ X_n^1 & X_n^2 & \cdots & X_n^p \end{bmatrix} = n\hat{V}_{XX}
$$

Gram matrix

$$
\begin{array}{cc}
 & p \\
 & \dim H \\
n & \begin{bmatrix} \Phi(X_1\ )^1 & \Phi(X_1\ )^2 & \cdots & \\ \vdots & \vdots & & \\ \Phi(X_n\ )^1 & \Phi(X_n\ )^2 & \cdots & \end{bmatrix}
\end{array}
\begin{bmatrix} \Phi(X_1\ )^1 & \cdots & \Phi(X_n\ )^1 \\ \Phi(X_1\ )^2 & \cdots & \Phi(X_n\ )^2 \\ \vdots & & \vdots \\ & & \\ \vdots & & \vdots \end{bmatrix} = \left(k\left(X_i, \mathrm{X_j}\right)\right)
$$

RKHS → low-cost computation

Linear methods on H_k are computable by Gram matrices of size n (sample size).

- Suitable for high-dimensional data of moderate sample size.
  c.f. power expansion / L^2 basis expansion.

Remark: If sample size n is large, low rank approximation of Gram matrices works well.

» Incomplete Cholesky factorization (Fine & Scheinberg 2001)

» Nyström approximation (Williams & Seeger 2000).

Representing probabilities with kernels

Mean on RKHS

X: random variable taking value on a measurable space Ω, $\sim P$.
k: pos.def. kernel on Ω. H: RKHS defined by k.

Def. kernel mean on H :

$$m_P := E[\Phi(X)] = E[k(\cdot, X)] = \int k(\cdot, x) dP(x) \quad \in H_k$$

– Reproducing expectations

$$\langle f, m_P \rangle = E[f(X)] \qquad \text{for any } f \in H_k.$$

– Kernel mean can express higher-order moments of X.

Suppose $k(u, x) = c_0 + c_1 ux + c_2 (ux)^2 + \cdots \quad (c_i \geq 0)$, e.g., e^{ux}

$$m_P(u) = c_0 + c_1 E[X] u + c_2 E[X^2] u^2 + \cdots$$

c.f. moment generating function

Characteristic kernel

(Fukumizu et al. JMLR 2004, AoS 2009; Sriperumbudur et al. JMLR2010)

Def. A bounded pos. def. kernel k is called characteristic if

$$\mathcal{P} \to H_k, \quad P \mapsto m_P$$

is injective, i.e., $E_{X\sim P}[k(\cdot, X)] = E_{Y\sim Q}[k(\cdot, Y)] \iff P = Q$.

m_P with a characteristic kernel uniquely determines a probability.

Examples: Gaussian, Laplace kernel

(polynomial: not characteristic.)

c.f. characteristic functions $E[e^{iuX}]$.

Kernel mean → advantage in efficient computation.

Nonparametric inference with kernels

Principle: with characteristic kernels,

Inference on P $\Rightarrow$ Inference on m_P

- Two sample test → $m_P = m_Q$?

  (Gretton et al. JMLR 2012)

- Independence test → $m_{XY} = m_X \otimes m_Y$?
  - Close connection to *distance covariance,* which is a popular dependence measure (Székely, Rizzo, Bakirov 2007)

    (Sejdinovic, Sriperumbudur, Gretton, Fukumizu, AoS 2013)

- Bayesian Inference → this talk.

Covariance

(X, Y) : random vector taking values on $\Omega_X \times \Omega_Y$.

(H_X, k_X), (H_Y, k_Y): RKHS on Ω_X and Ω_Y, resp.

Def. (uncentered) covariance operators $C_{YX}: H_X \to H_Y$, $C_{XX}: H_X \to H_X$

$$C_{YX} = E[\Phi_Y(X)\Phi_X(Y)^T], \qquad C_{XX} = E[\Phi_X(X)\Phi_X(X)^T]$$

Reproducing property

$$\langle g, C_{YX} f \rangle = E[f(X)g(Y)] \qquad \text{for all } f \in H_X, g \in H_Y.$$

Simply, extension of covariance matrix (linear map) $V_{YX} = E[XY^T]$

Empirical estimators

Given $(X_1, Y_1,), \dots, (X_n, Y_n) \sim P$, i.i.d.,

Empirical Estimator:

$$\widehat{m}_X = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i), \qquad \hat{C}_{YX}f = \frac{1}{n}\sum_{i=1}^{n} k_Y(\cdot, Y_i)\langle k(\cdot, X_i), f\rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n} k_Y(\cdot, Y_i) f(X_i)$$

– Typically Gram matrix expression is obtained.

e.g. $$\left\|\hat{C}_{YX}\right\|_{HS}^2 = Tr[G_X G_Y]$$

– $\sqrt{n}$-consistency (in norm) and CLT are guaranteed. (Berlinet & Thomas-Agnan 2004, Gretton et al. 2005)

Conditional probabilities

Conditional kernel mean

- X, Y: Gaussian random vectors ($\in R^m, R^\ell$, resp.)

$$\operatorname*{argmin}_{A \in R^{\ell \times m}} \int \|Y - AX\|^2 dP(X,Y) \;=\; V_{YX} V_{XX}^{\;\;-1}$$

$$E[Y|X = x] = V_{YX} V_{XX}^{\;\;-1} x$$

- With characteristic kernels, for general X and Y,

$$\operatorname*{argmin}_{F \in H_X \otimes H_Y} \int \|\Phi_Y(Y) - \underbrace{F(X)}_{\langle F, \Phi_X(X) \rangle}\|_{H_Y}^2 dP(X,Y) = C_{YX} C_{XX}^{\;\;-1}$$

$$\boxed{E[\Phi(Y)|X = x] = C_{YX} C_{XX}^{\;\;-1} \Phi_X(x)}$$

Representing the conditional probability of Y given $X = x$.
In practice, regularized inverse must be used.

– How to use the conditional kernel mean?

- Nonparametric estimator of regression

$$\hat{E}[g(Y)|X=x] = \mathbf{k}_X^T(x)(G_X + \varepsilon_n I_n)^{-1}\mathbf{g}$$

$$\mathbf{k}_X(\cdot) = \left(k_X(\cdot, X_1), \ldots, k_X(\cdot, X_n)\right)^T \in H_X^n,$$
$$\mathbf{g} = \left(g(Y_1), \ldots, g(Y_n)\right)^T \in R^r$$

ε_n: regularization coefficient

c.f. Gaussian process / kernel ridge regression

- Conditional independence (Fukumizu et al. JMLR 2004, AoS 2009, NIPS 2010)

- Bayesian inference (discussed later)

– Note: for consistency, kernel is fixed, regularization coefficient $\varepsilon_n \to 0$. *c.f*. smoothing kernel.

Comparison: nonparametric regression

Assume Y is 1 dim., and kernel is used only for X

→ $$\hat{E}[Y|X=x] := \mathbf{k}_X^T(x)(G_X + \varepsilon_n I_n)^{-1}Y$$

Gaussian process / kernel ridge regression

– Consistency 1 (Eberts & Steinwart 2011)

If k_X is Gaussian, and $E[Y|X] \in W_2^\alpha(P_X)$, (under some technical assumptions) for any $\rho > 0$,

$$E\left|\hat{E}[Y|X] - E[Y|X]\right|^2 = O_p\left(n^{-\frac{2\alpha}{2\alpha+m}+\rho}\right) \qquad (n \to \infty)$$

Note: $O_p\left(n^{-\frac{2\alpha}{2\alpha+m}}\right)$ is the optimal rate for a linear estimator (Stone 1982).

* $W_2^\alpha(P_X)$: Sobolev space of order α.

– Consistency 2 (case: $E[Y|X] \in H_X$)

Suppose $E[Y|X] \in R(C_{XX}^{\beta})$ with $\beta \geq 0$, Then, with a characteristic kernel k_X,

$$\left\| \hat{E}[Y|X] - E[Y|X] \right\|_{H_X}^2 = O_p\left(n^{-\min\left\{\frac{1}{2}, \frac{\beta}{\beta+1}\right\}} \right)$$

- The rates do not depend on m (dim of X), since the analysis can be done within the RKHS.
- $|| \cdot ||_{H_X}$ is stronger than $|| \cdot ||_{sup}$. Thus,

$$\sup_x \left| \hat{E}[Y|X = x] - E[Y|X = x] \right| = O_p\left(n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2\beta+2}\right\}} \right)$$

Numerical studies

Comparisons

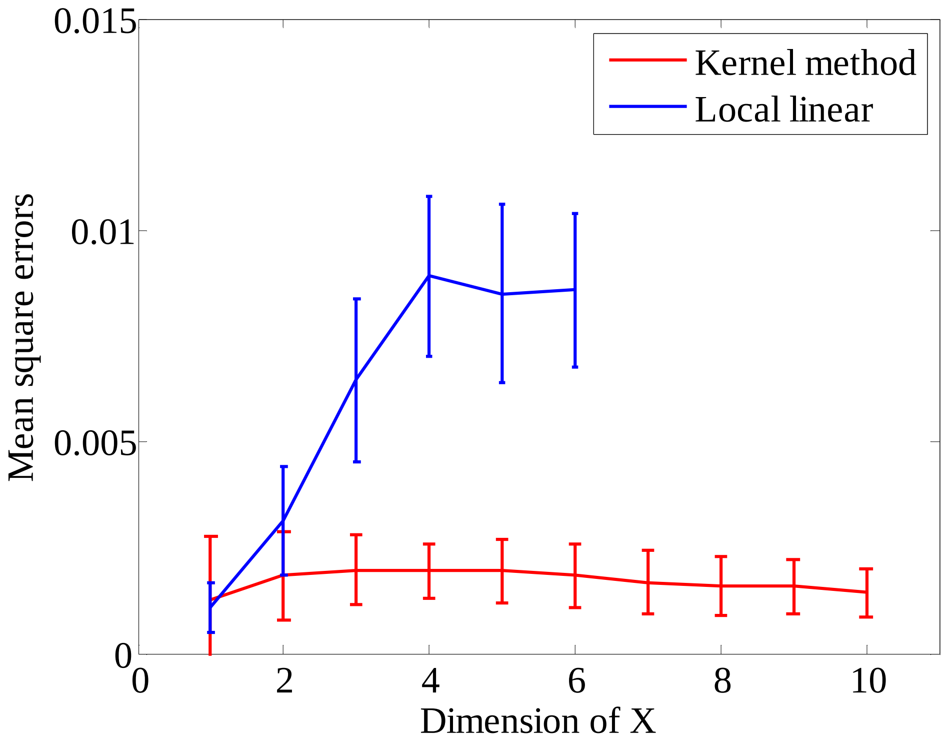
$$Y = 1/(1.5 + ||X||^2) + Z, \qquad X \sim N(0, I_d),\ Z \sim N(0, 0.1^2)$$

n = 100,   500 runs

Kernel ridge regression
with Gaussian kernel

Local linear regression
with Epanechnikov kernel
('locfit' in R is used)

Bandwidth parameters
are chosen by CV.



Kernel method for Bayesian inference

Kernel realization of Bayes' rule

- Bayes' rule

$$q(x|y) = \frac{p(y|x)\pi(x)}{q(y)}, \qquad q(y) = \int p(y|x)\pi(x)dx.$$

Π: prior with p.d.f π

$p(y|x)$: conditional probability (likelihood).

- Kernel realization:

Goal: estimate the kernel mean of the posterior

$$m_{post|y_*} := \int k_X(\cdot, x)q(x|y_*)dx$$

given

– m_Π: kernel mean of prior Π,

– C_{XX}, C_{YX}: covariance operators for $(X, Y) \sim P$,

where P is the joint probability to give $p(y|x)$ by conditioning.

Kernel Bayes' Rule

(Fukumizu, Song, Gretton JMLR2014)

Input: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ (to give cond. probability).

$\hat{m}_\Pi = \sum_{j=1}^{\ell} \gamma_j \Phi_X(U_j)$ (prior) a consistent estimator of m_Π.

1. [Expression of $q(x,y) = p(y|x)\pi(x)$: ← regression $((Y,X), U_\gamma) \,|X$]

   Compute $\Lambda = \mathrm{Diag}[(G_X/n + \varepsilon_n I_n)^{-1} G_{XU}\gamma\ ]$

2. [Conditioning: ← regression with $(W,Z) \sim q(x,y)$]

   Compute $R_{W|Z} = \Lambda G_Y \left((\Lambda G_Y)^2 + \delta_n I_n \right)^{-1} \Lambda.$

* ε_n, δ_n: regularization coefficients

Output: estimator for kernel mean of posterior given observation y_*

$$\boxed{\hat{m}_{post|y_*}(\cdot) = \mathbf{k}_X(\cdot)^T R_{W|Z} \mathbf{k}_Y(y_*) = \sum_{i=1}^{n} w_i(y_*) k_X(\cdot, X_i)}$$

Inference with KBR

- Weighted sample expression

$$\hat{m}_{post|y_*}(\cdot) = \sum_{i=1}^{n} w_i(y_*) k_X(\cdot, X_i)$$

Equivalent to the kernel mean of

$$\sum_{i=1}^{n} w_i(y_*) \delta_{X_i} \qquad (\delta_x\text{: Dirac's delta})$$

which is a signed measure (not necessarily a probability).
Some weights may be negative.

– $\sum_{i=1}^{n} w_i(y_*) \to 1 \ (n \to \infty)$ in probability under mild assumption.

How to use?

- Expectation: if $\frac{\pi}{p_X} \in \overline{\mathrm{Range}(C_{XX})}$ and $f \in L^2(P_X)$ satisfies $\int f(x)p(y|x)\pi(x)dx \in \mathrm{Range}(C_{YY})$,

$$\sum_{i=1}^{n} w_i(y_*)f(X_i) \to \int f(x)q(x|y_*)dx, \qquad (n \to \infty). \quad \text{(consistent)}$$

  e.g.

  - $f(x) = I_B(x)$: $\sum_{X_i \in B} w_i \to$ posterior prob. of set B.
  - $f(x) = x^r$: $\sum_i w_i X_i^r \to r$-th moment of posterior.

    (More general discussions in Kanagawa and Fukumizu, AISTATS 2014)

- Point estimation (quasi-MAP):

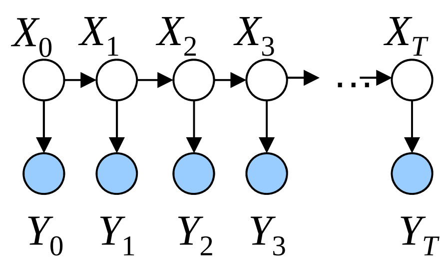
$$\hat{x} = \mathrm{argmin}_x \left\| \widehat{m}_{post|y_*} - \Phi_X(x) \right\|_{H_X}$$

  Solved numerically

– Completely nonparametric way of computing Bayes rule.
  - No parametric models are needed, but data or samples are used to express the probabilistic relations.
  - Bayesian inference is done with matrix computation.

Examples:

1. Nonparametric HMM



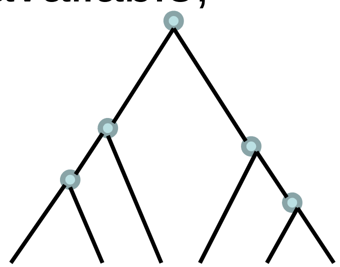
$$p(X,Y) = p(X_0,Y_0)\prod_{t=1}^{T} p(Y_t|X_t)q(X_t|X_{t-1})$$

$p(Y_t|X_t)$ and /or $q(X_t|X_{t-1})$ are unknown, but data are available.

2. Explicit form of likelihood $p(y|x)$ or prior π is unavailable, but sampling is possible.

   c.f. Approximate Bayesian Computation (ABC)

   (Kernel ABC: Nakagome, Mano, Fukumizu 2013)



Practical example

State $X_t \in \mathbf{R}^3$: 2-D coordinate and orientation of a robot

Observation Y_t : image sequence.

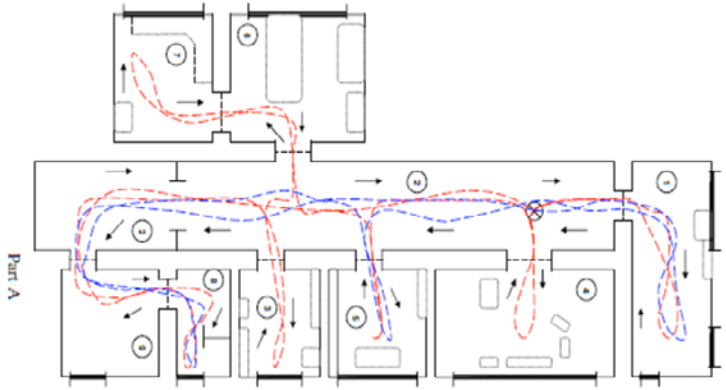
Training sample $(X_t, Y_t) : t = 1, \ldots, T$

Estimate the location of a robot from image sequences

- Observation: $p(Y_t | X_t)$

  Very difficult to model with a simple parametric model.

  → KBR !



Convergence rate

Theorem (Fukumizu, Song, Gretton 2012)

Let $f \in H_X$, $(Z,W) \sim Q$ with p.d.f. $p(y|x)\pi(x)$.

Assumptions:

- $\|\widehat{m}_\Pi - m_\Pi\|_{H_X} = O_p(n^{-\alpha})$ for some $0 < \alpha \le 1/2$.
- $\pi(x)/p_X(x) \in \mathrm{Range}\left(C_{XX}^{1/2}\right)$ for some $\beta \ge 0$.
- $E[f(Z)|W=\cdot] \in \mathrm{Range}(C_{XX}^2)$ for some $\nu \ge 0$.

Then, with $\varepsilon_n = n^{-2\alpha/3}$ and $\delta_n = n^{-8\alpha/27}$, for any y,

$$\mathbf{f}_X^T R_{X|Y} \mathbf{k}_Y (y) - E[f(Z)|W = y] = O_p(n^{-\frac{8\alpha}{27}}) \qquad (n \to \infty).$$

- Remark: the rate depending on the smoothness of the functions π/p_X and $E[f(Z)|W=\cdot]$ is also available.
- If $\alpha = 1/2$, the rate is $n^{-4/27}$ (very slow, unsatisfactory....).

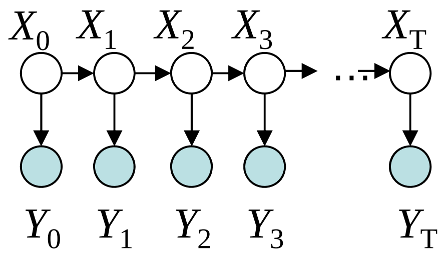
Choice of kernel and hyperparameter

– Parameters to be chosen for kernel methods: kernel (parameters in kernel) and regularization parameter for regression and KBR.

– In general, cross-validation is recommended, if possible.
  - Straightforward in supervised setting.
  - Make a relevant supervised problem and apply CV (e.g. HMM).

  Supports
  - CV has been used successfully for SVM.
  - The rate $O_p(n^{-\frac{2\alpha}{2\alpha+m}+\rho})$ for the regression is attained with parameter choice by validation (Eberts & Steinwart 2011).

Example: KBR for nonparametric HMM



- Assume:
  $p(y_t|x_t)$ and/or $q(x_t|x_{t-1})$ is not known.
  But, data $(X_t, Y_t)_{t=0}^{T}$ is available
  in training phase.

  Examples:

  - Measurement of hidden states is expensive,
  - Hidden states are measured with time delay.

- Testing phase (e.g., filtering, e.g.):

  given $\tilde{y}_0, \ldots, \tilde{y}_t$, estimate hidden state x_s.

  →KBR point estimator: $\mathrm{argmin}_{x_s} \left\| \widehat{m}_{x_s|\tilde{y}_0,\ldots,\tilde{y}_t} - \Phi(x) \right\|_{H_X}$

- General sequential inference uses Bayes' rule ➔ KBR applied.

Numerical examples

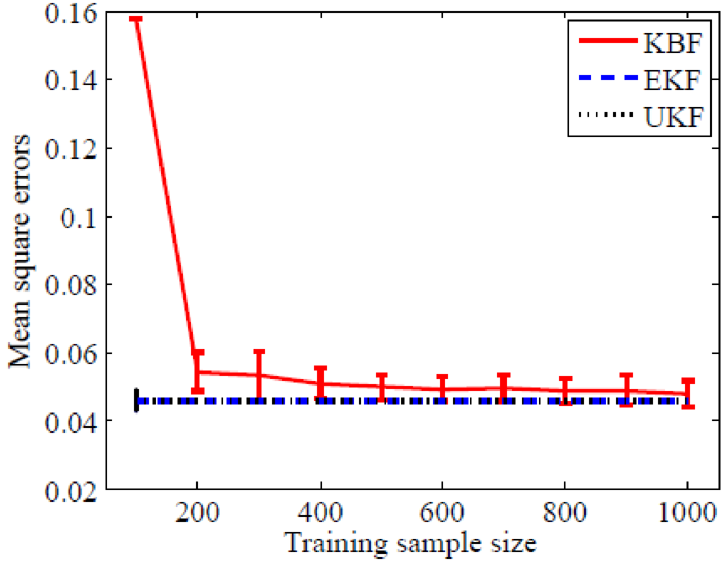
(a) Noisy rotation

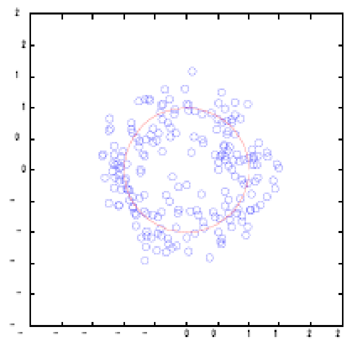
$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} + Z_t, \qquad \theta_{t+1} = \arctan\left(\frac{v_t}{u_t}\right) + 0.3,$$

$$Y_t = (u_t, v_t)^T + W_t,$$

$$Z_t, W_t \sim N(0, 0.04 I_2) \ (i.i.d.)$$

Filtering with the point estimator by KBR.





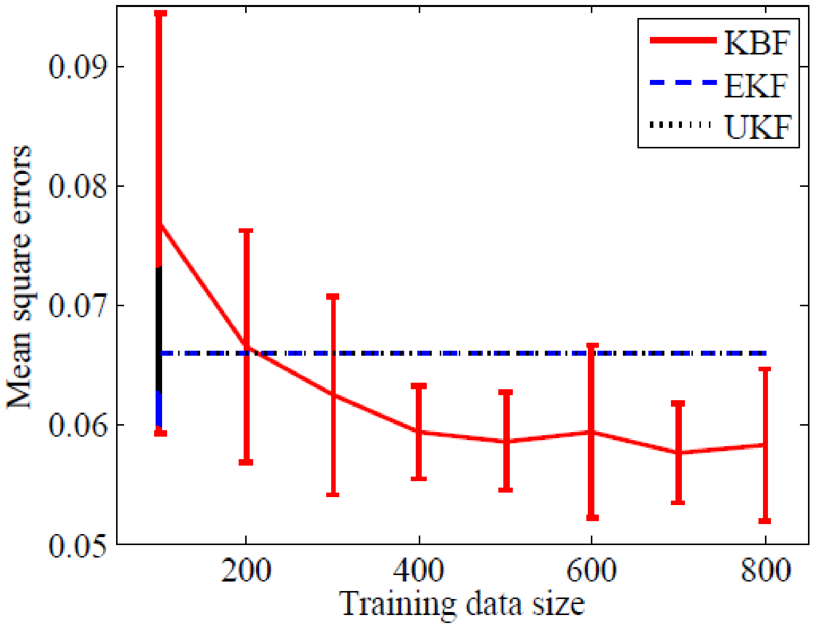
KBR does NOT know the dynamics, while the EKF and UKF use it.

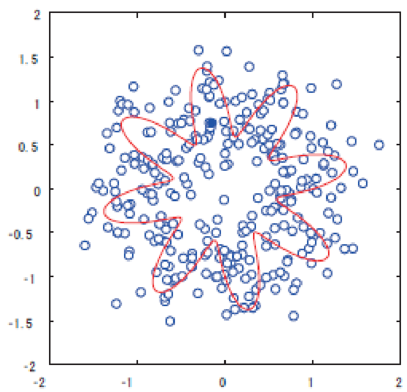
(b) Noisy oscillation

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = (1 + 0.4\sin(8\theta_t)) \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} + Z_t, \qquad \theta_{t+1} = \arctan\left(\frac{v_t}{u_t}\right) + 0.4,$$

$$Y_t = (u_t, v_t)^T + W_t,$$

$$Z_t, W_t \sim N(0, 0.04 I_2)\ (i.i.d.)$$





Camera angles

- Hidden X_t: angles of a video camera located at a corner of a room.
- Observed Y_t: movie frame of a room + additive Gaussian noise.
- X_t: 3600 downsampled frames of 20 x 20 RGB pixels (1200 dim.).
- The first 1800 frames for training, and the second half for testing.

noise	KBR (Trace)	Kalman filter(Q)
$\sigma^2 = 10^{-4}$	$0.15 \pm < 0.01$	0.56 ± 0.02
$\sigma^2 = 10^{-3}$	0.21 ± 0.01	0.54 ± 0.02

Average MSE for camera angles (10 runs)

To represent SO(3) model, Tr[AB^{-1}] for KBR, and quaternion expression for Kalman filter are used .

Robot localization (Re)

- COLD (COsy Localization Dataset, IJRR 2009)

  State $X_t \in \mathbf{R}^3$: 2-D coordinate and orientation of a robot

  Observation Y_t : image sequence（SIFT feature, 4200dim）

  Training sample $(X_t, Y_t) : t = 1, \ldots, T$

  Estimate the location of a robot from image sequences

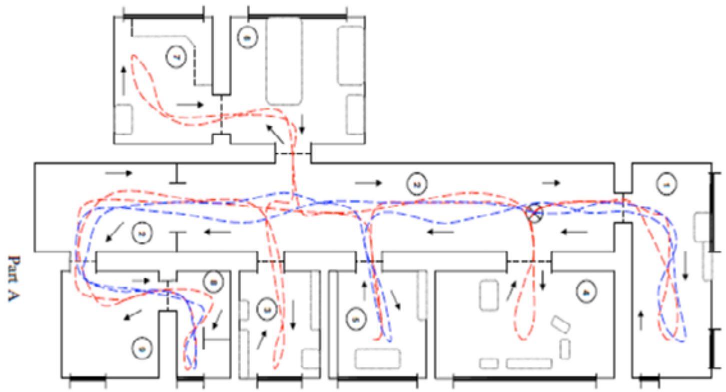
  – Observation: $p(Y_t|X_t)$ difficult to model.

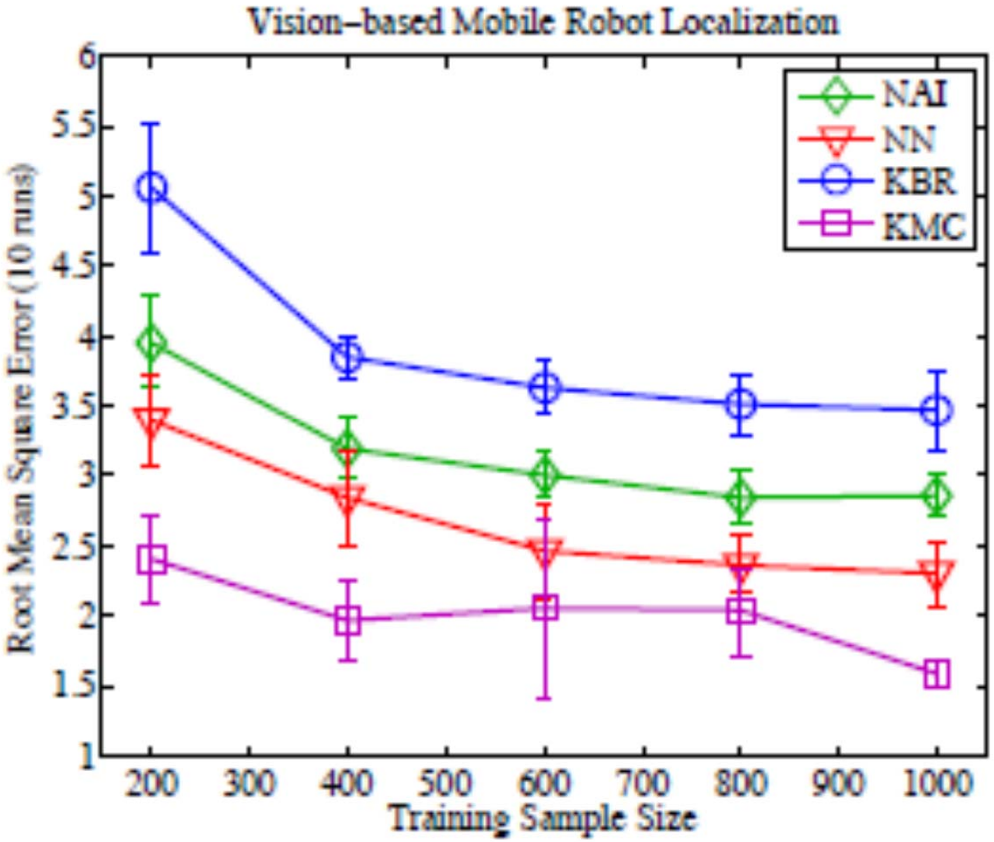
    → KBR

  – State transition: linear Gaussian

    Kernel Monte Carlo,

    (Kanagawa, Nishiyama, KF. 2013)



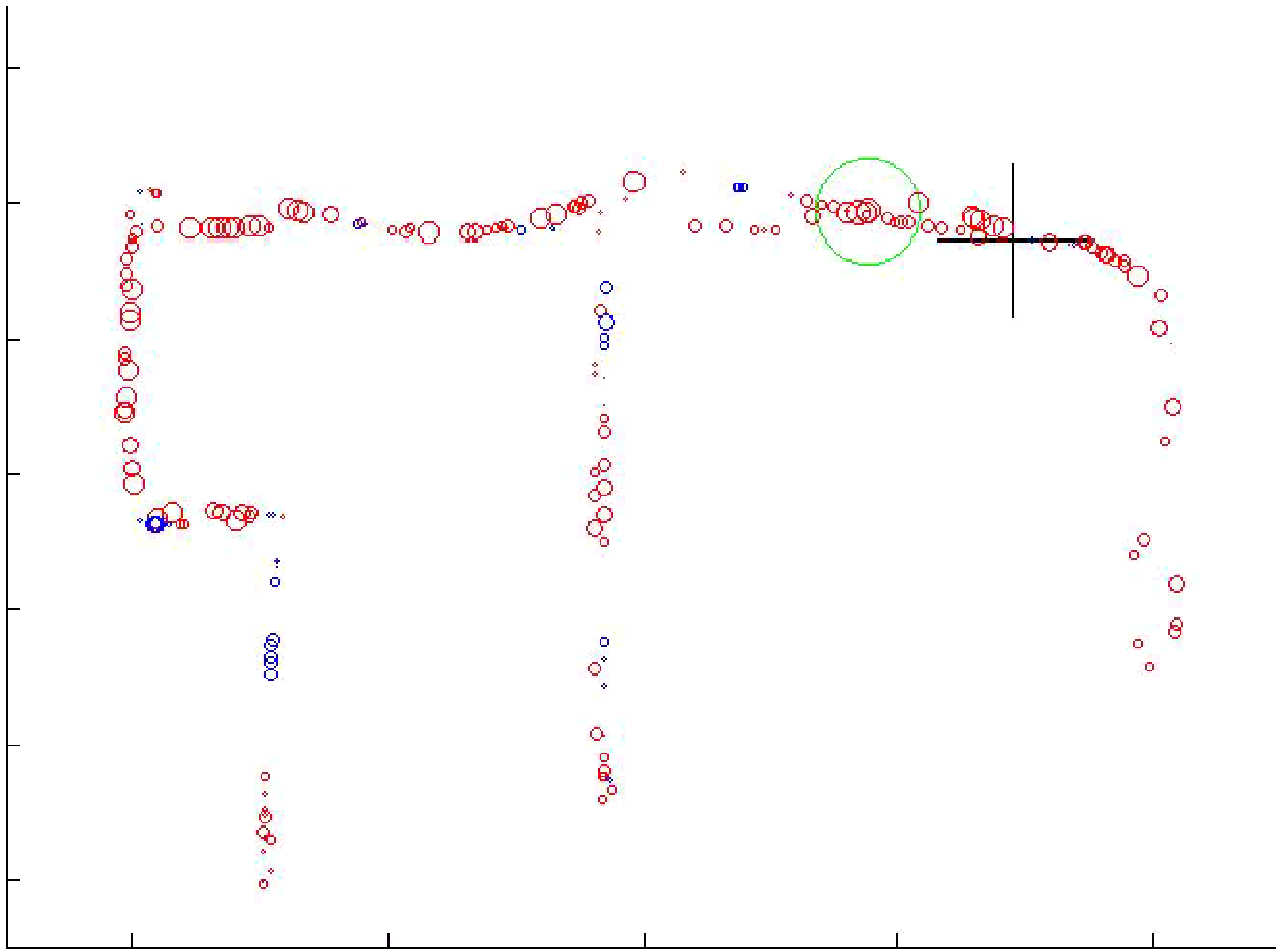


NAI: naïve method
(closest image
in training data)

NN: PF + K-nearest
neighbor
(Vlassis, Terwijn, Kröse 2002)

KBR: KBR + KBR

KMC: KBR + Monte Carlo



training sample = 200

✚: true location

◯ : estimate

red (+)/ blue (-) circles: weights on the training sample

Conclusions and discussions

- "Kernel methods": useful, general tool for nonparametric inference.
  - Suitable for high-dimensional data.
    - Efficient computation with Gram matrices.
    - Good performance for high-dimensional data.
  - Can be used for representing probabilities and conditional probabilities.
  - "Nonparametric" way for general Bayesian inference with matrix computation.

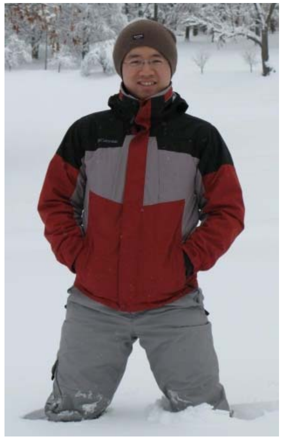
- Theoretical study is yet to be done
  - How can we justify the good performance of high-dimensionality theoretically?

    Large dimensional asymptotics?

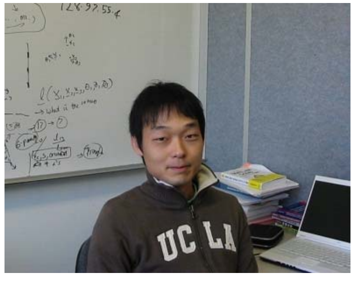
Collaborators

Arthur Gretton (UCL/MPI)

Bernhard Schölkopf (MPI)



Le Song (Georgia Tech)



Yu Nishiyama (ISM)

Motonobu Kanagawa (GUAS/ISM)