カーネル法の新展開
― その理論と応用 ―

福水　健次

統計数理研究所 ／ 総合研究大学院大学

第15回情報論的学習理論ワークショップ（IBIS2012）

2012.11.7-9. @筑波大学東京キャンパス

Outline

1. イントロダクション：　カーネル法の概要
2. 確率分布の表現としてのカーネル法
3. 条件付確率の表現と推定精度
4. カーネル推論則
5. おわりに

データの高次元性，非線形性

- 高次元データに潜む非線形性，複雑な構造

  生物，遺伝，文書，ソーシャルネットワーク，宇宙，気象，…

- データの非線形性の抽出

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

  - 高次元性に伴う計算量爆発

    例） 10,000 次元データに 2 次特徴を加えると

    特徴空間の次元 = ${}_{10000}C_1 + {}_{10000}C_2 = 50{,}005{,}000$ (!)

  - より効率的な方法は？ → カーネル法

正定値カーネルと再生核ヒルベルト空間

定義. Ω: 集合. $k : \Omega \times \Omega \to \mathbf{R}$ が正定値であるとは,

$k(x, y) = k(y, x)$ [対称] かつ任意の $n \in \mathbf{N}$, $x_1, \ldots, x_n \in \Omega$ に対し

Gram行列 $\left(k(x_i, x_j)\right)_{ij}$ が半正定値行列.

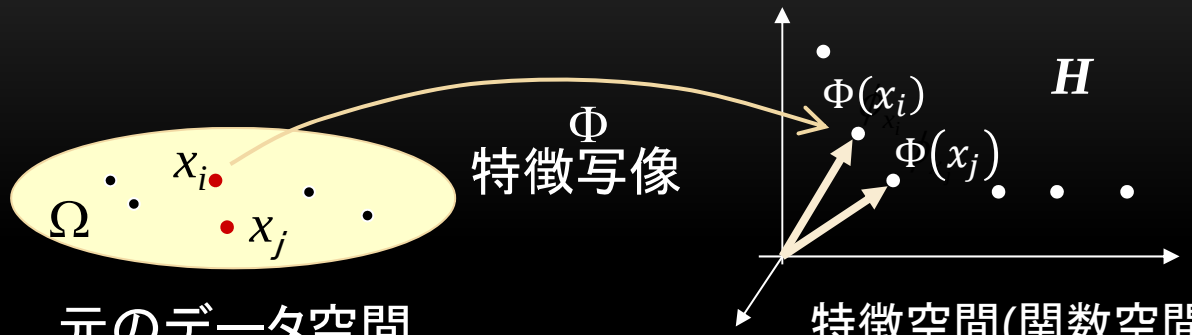
例） Gaussian RBFカーネル $\exp(-\|x - y\|^2 / 2\sigma^2)$

Laplace カーネル $\exp(-\alpha \sum_{i=1}^{m} |x_i - y_i|)$

- 再生核ヒルベルト空間（Reproducing kernel Hilbert space, RKHS）
  - 正定値カーネルにより定まる，Ω上の関数からなる関数空間.
  - 特殊な内積を持つ.

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

データ変換：RKHSへの特徴写像



特徴空間上で
線形データ解析を行う.
SVM, PCA, CCA, etc

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

- 特徴ベクトル: $X_1, \ldots, X_n \quad \mapsto \quad \Phi(X_1), \ldots, \Phi(X_n)$
- カーネルトリック

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

内積計算： $f = \sum_i \alpha_i \Phi(X_i),\ g = \sum_i \beta_i \Phi(X_i) \in H_k,$

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

- 線形アルゴリズムのカーネル化（既存）
  - 特徴ベクトルに対する線形アルゴリズム
    → さまざまなカーネル法
    e.g. カーネルPCA, 非線形SVM, カーネルCCA, etc.
  - 大きさ n（データ数）の Gram 行列計算に還元される.
    - 高次元データに適する.
      c.f. 高次項による展開
    - データ数 n が大きい時は低ランク近似 が有効.

- カーネル法の新しい展開
  - カーネル平均，条件付カーネル平均を用いた確率分布の表現
  - Gram行列計算によるノンパラメトリックな推論計算の実現

- 既存のノンパラメトリック推定
  - 平滑化カーネル（正定値とは限らない）：
    カーネル密度推定, 局所多項式　　$h^{-d}K(x/h)$
  - 特性関数の方法:　$E[e^{i\omega X}]$
  - スプライン，ウェーブレット, etc, etc,

  → 「次元の呪い」
    - 平滑化カーネル：　高次元（4, 5次元ぐらい）で困難

- カーネル法によるノンパラメトリック推定
  - 何ができるのか？
  - 「次元の呪い」を解決しているか？

確率分布の表現としてのカーネル法

カーネル平均：特徴ベクトルの平均

X: 可測空間 Ω 上に値を取る確率変数 $\sim P$.
k: Ω 上の正定値カーネル.　　H: k で定まるRKHS.

Def.　H における X のカーネル平均:

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

- 期待値の再生性

$$\langle f, m_P \rangle = E[f(X)] \qquad \forall f \in H_k.$$

- カーネル平均は X の高次モーメントの情報を持つ.

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

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

モーメント母関数

特性的なカーネル

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

Def. 有界で可測な正定値カーネル k が特性的(characteristic) とは,

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

が単射であることをいう. すなわち

$$E_{X\sim P}[k(\cdot, X)] = E_{Y\sim Q}[k(\cdot, Y)] \Leftrightarrow P = Q.$$

特性的なカーネルによるカーネル平均m_P は, 確率分布を一意に定める.

例: Gaussian, Laplace カーネル　(多項式カーネル: 非特性的.)

c.f. 特性関数 $E[e^{iuX}]$.

カーネル平均 → 効率的な計算が可能.

カーネル法によるノンパラメトリック推論の原理

特性的なカーネルにより，
　確率分布に関する推論　⇒　カーネル平均に関する推論・計算

- 分布の均一性検定（$P = Q$ か？）　→　$m_P = m_Q$?
- 独立性検定（XとYは独立か？）　→　$m_{XY} = m_X \otimes m_Y$?

RKHSにおける分散共分散

(X, Y): $\Omega_X \times \Omega_Y$ に値を取る確率変数
$(H_X, k_X), (H_Y, k_Y)$: Ω_X, Ω_Y 上のRKHSと正定値カーネル

Def. (非心化) 共分散作用素 $C_{YX}: H_X \to H_Y, C_{XX}: H_X \to H_X$

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

- 通常の共分散行列（線形写像）のRKHS値変数への一般化

$$V_{YX} = E[XY^T]$$

- テンソル（積空間）との同一視：

$$C_{YX} = E[\Phi_Y(Y) \otimes \Phi_X(X)] \quad \in H_Y \otimes H_X$$

一般に　線形写像 ≅ 2階のテンソル

$[f \mapsto g\langle h, f\rangle] \leftrightarrow \ g \otimes h$　(ランク1の場合)

サンプルによる推定

$(X_1, Y_1,), \dots, (X_n, Y_n) \sim P$, i.i.d.,
推定量:　標本平均，標本共分散でOK

$$\hat{m}_X = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i), \quad \hat{C}_{YX} = \frac{1}{n}\sum_{i=1}^{n} k_Y(\cdot, Y_i) \otimes k_X(\cdot, X_i)$$

- さまざまな量がGram行列で表現可能
- $1/\sqrt{n}$-オーダーでの一致性 (in norm) や中心極限定理が保証される (see e.g., Berlinet & Thomas-Agnan 2004)

条件付確率の表現と推定精度

条件付カーネル平均

- X, Y: ガウス確率変数($\in R^m, R^\ell$, resp.)

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

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

- 特性的なカーネルを用いると, 一般のX, Yに対し

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

$$E[\Phi(Y)|X = x] = C_{YX} {C_{XX}}^{-1} \Phi_X(x)$$
$$\| $$
$$\int \Phi(y) p(y|x) dy$$

$X = x$を与えたときの Y の条件付確率のカーネル平均

条件付カーネル平均の応用

- 確率モデルに基づく推論 e.g. グラフィカルモデル　（後述）
- 条件付独立性／依存性 (Fukumizu et al. JMLR 2004, AoS 2009, NIPS 2010)
- ノンパラメトリック回帰（c.f. ガウス過程/ カーネルリッジ回帰)

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

$$\mathbf{k}_X(\cdot) = (k_X(\cdot, X_1), \ldots, k_X(\cdot, X_n))^T \in H_X^n,$$

$$\mathbf{g} = \big(g(Y_1), \ldots, g(Y_n)\big)^T \in R^r$$

ε_n: 正則化係数

収束の速さ

Y: 1次元と仮定．X のみに カーネルを用いる

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

ガウス過程/ カーネルリッジ回帰

- 一致性 1 (Eberts & Steinwart , NIPS2011)

  $X \in \mathbf{R}^m$，k_X：ガウスカーネル，$E[Y|X] \in W_2^\alpha(P_X)$, ある種の緩い仮定のもと，任意の$\rho > 0$ に対し，

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

Note: $O_p\left(n^{-\frac{2\alpha}{2\alpha+m}}\right)$ は，線形推定量の最適オーダー (Stone 1982).

* $W_2^\alpha(P_X)$: **オーダーαのSobolev空間**

- 一致性 2 (case: $E[Y|X] \in H_X$)

  k_Xを特性的，$E[Y|X] \in \mathrm{Range}(C_{XX}^{\beta})$ （$\beta \geq 0$）とするとき，

$$E\left|\hat{E}[Y|X] - E[Y|X]\right|^2 = O_p\left(n^{-\min\left\{\frac{1}{2}, \frac{2\beta+1}{2\beta+3}\right\}}\right)$$

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

  - 収束のレートは m （Xの次元）に依存しない（「次元の呪い」の解消？）
    しかし「一致性1」の$\alpha \to \infty$の場合より遅い.
  - Gaussian RKHS $\subset W_2^{\alpha}(P_X)$, 稠密
  - 「一致性1」では， $W_2^{\alpha}(P_X)$の任意の関数をGaussian RKHSの$\|f\|_H < R$ なる関数により近似する際の，モデル誤差の議論が必要．モデル誤差は次元に依存.

比較実験

- 実験条件

  カーネルリッジ回帰（Gaussカーネル）

  局所線形回帰（Epanechnikov kernel） (R: 'locfit'パッケージ)

  $Y = f(X) + Z, \qquad X \sim N(0, I_d), \;\; Z \sim N(0, 0.1^2)$

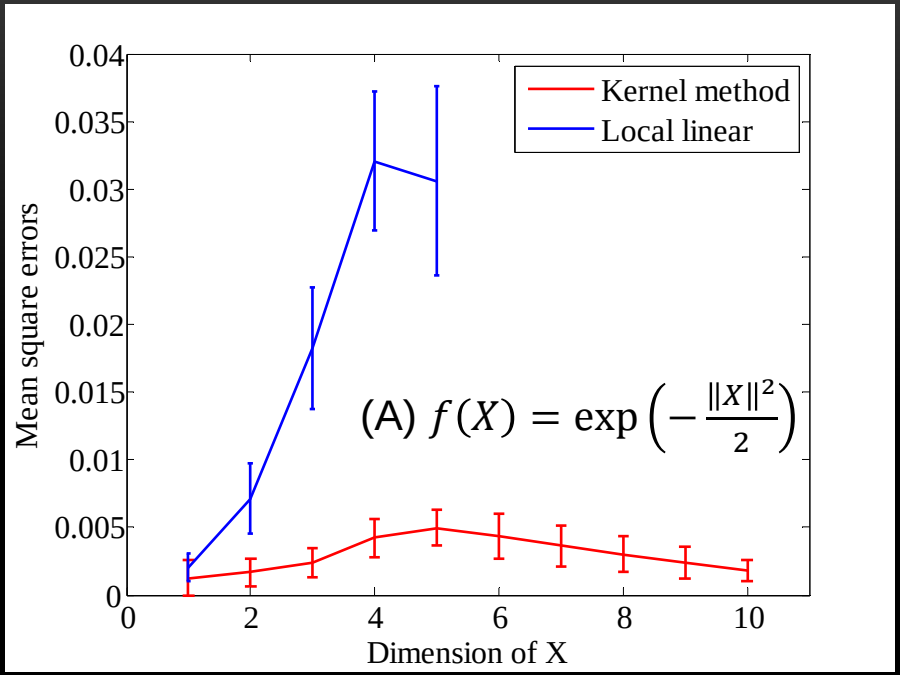
  (A) $f(X) = \exp\left(-\frac{\|X\|^2}{2}\right)$

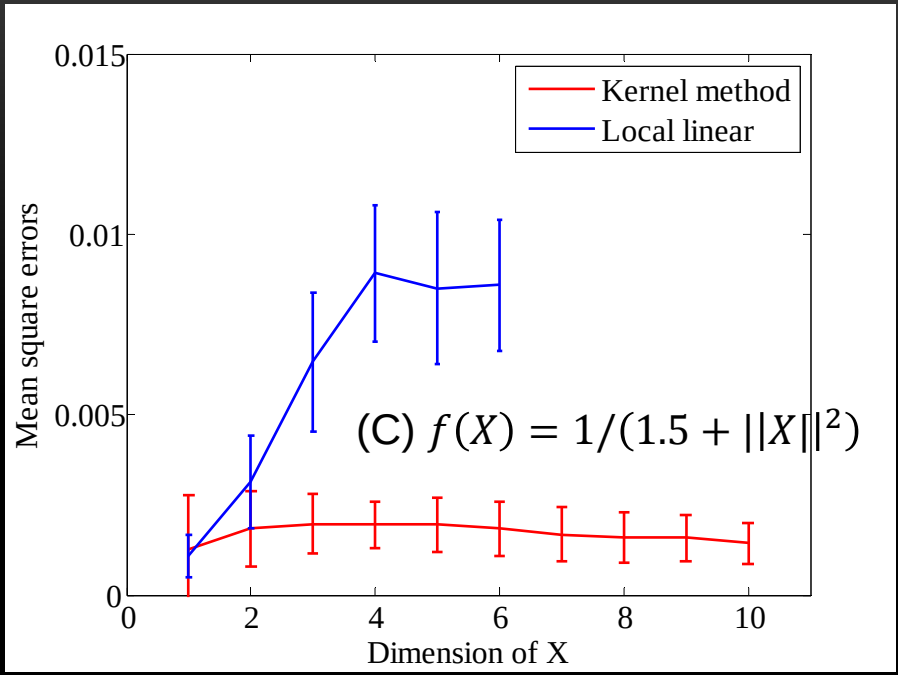
  (B) $f(X) = \frac{1}{m}(X_1 + \cdots, X_m)$ not included in Gaussian RKHS

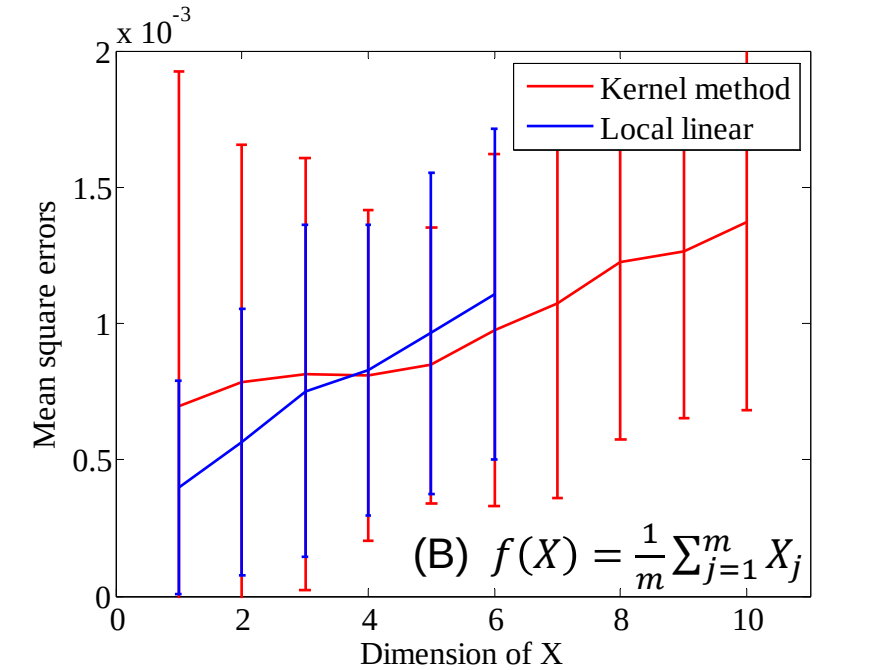
  (C) $f(X) = 1/(1.5 + ||X||^2)$ not included in Gaussian RKHS

  データ数 $n = 100$, 次元 $m = 1,\ldots,10$ (at most) , 500 runs

  バンド幅, 正則化係数はCVで選択.


Kernel method
Local linear
Mean square errors
Dimension of X
(A) $f(X) = \exp\left(-\frac{\|X\|^2}{2}\right)$


Kernel method
Local linear
Mean square errors
Dimension of X
(C) $f(X) = 1/(1.5 + ||X||^2)$


x 10^-3
Kernel method
Local linear
Mean square errors
Dimension of X
(B) $f(X) = \frac{1}{m}\sum_{j=1}^{m} X_j$

カーネル推論則

条件付確率を用いた推論則

- Sum rule： $q(y) = \int p(y|x)\pi(x)dx$
- Chain rule： $q(x,y) = p(y|x)\pi(x)$
- Bayes' rule： $q(x|y) = \frac{p(y|x)\pi(x)}{\int p(y|x)\pi(x)dx}$
- カーネル化：
  - 変数の関係をすべてデータで表す（ノンパラメトリック！）
  - 分布をすべて（重み付）カーネル平均，共分散作用素で表現
  - Gram行列計算によって上記の計算ルールを実現する．

Kernel Sum Rule

- Sum rule: $q(y) = \int p(y|x)\pi(x)dx$
- カーネル化： $m_Y = C_{YX}C_{XX}^{-1}m_\pi$

$$\hat{m}_Y := \hat{C}_{YX}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1}\hat{m}_\pi$$

- Gram行列表現：

Input: $\hat{m}_\pi = \sum_{i=1}^{\ell}\alpha_i\Phi(\tilde{X}_i)$, $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{XY}$,

➡ $\hat{m}_Y = \sum_{i=1}^{n}\beta_i\Phi(Y_i)$, $\beta = (G_X + n\varepsilon_n I_n)^{-1}G_{X\tilde{X}}\alpha$.

$$G_{X\tilde{X}} = \left(k(X_i, \tilde{X}_j)\right)_{ij}$$

- Intuition: $m_\pi = \int \Phi(x)\pi(x)dx$, $C_{YX}C_{XX}^{-1}\Phi(x) = \int \Phi(y)p(y|x)dy$

$$\begin{aligned} C_{YX}C_{XX}^{-1}m_\pi &= \int C_{YX}C_{XX}^{-1}\Phi(x)\pi(x)dx \\ &= \int\int \Phi(y)p(y|x)\pi(x)dydx \\ &= \int \Phi(y)q(y)dy = m_Y \end{aligned}$$

Kernel Chain Rule

- Chain rule: $q(x,y) = p(y|x)\pi(x)$
- カーネル化： $C_Q = C_{(YX)X} C_{XX}^{-1} m_\pi$

$$\hat{C}_Q := \hat{C}_{(YX)X}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1} \hat{m}_\pi$$

- Gram行列表現：

Input: $\hat{m}_\pi = \sum_{i=1}^{\ell} \alpha_i \Phi\left(\tilde{X}_i\right), (X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{XY}$

$$\Rightarrow \hat{C}_Q = \sum_{i=1}^{n} \beta_i \Phi(Y_i) \otimes \Phi(X_i),\ \beta = (G_X + n\varepsilon_n I_n)^{-1} G_{X\tilde{X}} \alpha.$$

- Intuition: Note $C_{(YX)X}: H_X \to H_Y \otimes H_X$, $E\left[\left(\Phi(Y) \otimes \Phi(X)\right) \otimes \Phi(X)\right]$

From Sum Rule,

$$C_{(YX)X} C_{XX}^{-1} m_\pi = \int\int\int \Phi(y) \otimes \Phi(x) \underbrace{p(y|x)\delta(x - x')}_{p(y,x|x')} \pi(x') dy dx dx'$$

$$= \int\int \Phi(y) \otimes \Phi(x) p(y|x)\pi(x) dy dx = C_Q$$

Kernel Bayes' Rule

- ベイズルール

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

- カーネルベイズルール (KBR, Fukumizu et al NIPS2011)

  ベイズ則は，$q(x,y) = p(y|x)\pi(x)$ に対する，$y \to x$ の回帰.

$$m_{Q_x|y} = C_{XY}^{\pi} {C_{YY}^{\pi}}^{-1} \Phi(y)$$

where $\quad C_{YX}^{\pi} = C_{(YX)X} C_{XX}^{-1} m_{\pi}, \quad C_{YY}^{\pi} = C_{(YY)X} C_{XX}^{-1} m_{\pi}$

- Gram行列表現： Input: $\hat{m}_{\pi} = \sum_{i=1}^{\ell} \alpha_i \Phi(\tilde{X}_i), (X_1, Y_1), \dots, (X_n, Y_n) \sim P_{XY},$

  ➡ $\hat{m}_{Q_x|y} = \sum_{i=1}^{n} w_i(y) \Phi(X_i),$

  $w(y) = R_{X|Y}\mathbf{k}(y), \quad \mathbf{k}(y) = \big(k(Y_1, y), \dots, k(Y_n, y)\big)^T$

  $R_{X|Y} = \Lambda G_{YY}\big((\Lambda G_{YY})^2 + \delta_n I_n\big)^{-1} \Lambda \mathbf{k}(y),$

  $\Lambda = \mathrm{Diag}[(G_{XX} + n\varepsilon_n I_n)^{-1} G_{X\tilde{X}} \alpha]$

KBRによる推論

KBR は事後確率$q(x|y)$自身ではなく，そのカーネル平均を推定.

Bayes推論にどのように用いるか？

- $f \in H_X$ に対する$f(X)$のノンパラメトリック回帰

$$\underbrace{\mathbf{f}_X^T R_{X|Y} k_Y(y)}_{\text{KBR推定量}} \to \int f(x) q(x|y) dx. \qquad \text{(一致性)}$$

$$\text{where } \mathbf{f}_X = \left(f(X_1), \ldots, f(X_n)\right)^T.$$

- 点推定:

$$\hat{x} = \text{argmin}_x \left\| \hat{m}_{X|Y=y} - \Phi_X(x) \right\|_{H_X}$$

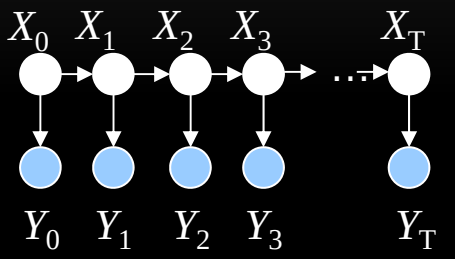
数値的解法が必要

カーネル推論則の応用例

完全な「ノンパラメトリック」ベイズ推論

分布の情報もサンプルで表現

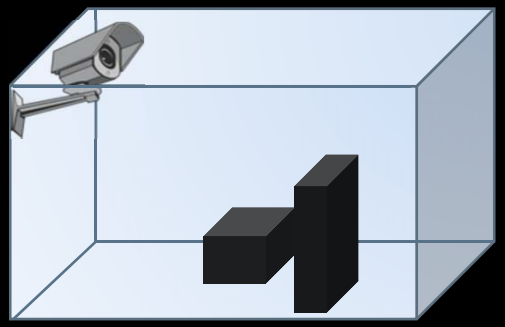
c.f. “いわゆる”ノンパラメトリックベイズ



応用例

- ノンパラメトリックHMM (Fukumizu et al. NIPS 2011)
  $p(Y_t|X_t)$ や $q(X_t|X_{t-1})$ をサンプルで表現.
- Kernel Approximate Bayesian Computation (Kernel ABC)
  (Nakagome , Fukumizu, Mano. 2012)
  尤度関数が陽に書けない場合
- カーネルBelief Propagation (Song et al, AISTATS2011)
- POMDP: Bellman方程式のカーネル化　(Nishiyama et al UAI2012)

- ノンパラメトリックHMMへの応用：カメラ角度の推定
  - 隠れ変数 X_t: 室内に位置を固定されたビデオカメラの角度.
  - 観測変数 Y_t: 部屋の動画像 + 人工ガウスノイズ.
  - 20 x 20 RGB 画素 (1200次元).



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 .

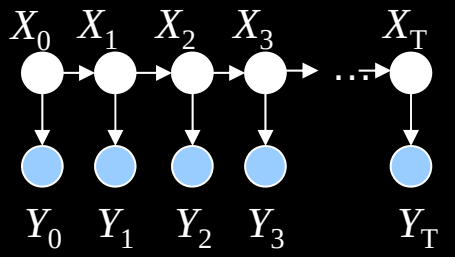
おわりに

- カーネル法は，ノンパラメトリック推論の有力な方法論
  - 高次元データに対する適性：　計算量，推定精度
- カーネル平均，共分散作用素に基づく推論則.
  - 完全な「ノンパラメトリック」な推論
  - ベイズ推論が実現可能.
- セミパラメトリックなカーネルベイズへの展開
  - パラメトリック部分 ＋ ノンパラメトリック部分（カーネル法）

    例）セミパラメトリックHMM

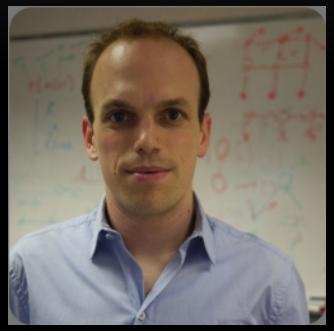
    遷移過程：条件付確率known

    観測過程：unknownだがデータが存在
    - 粒子フィルタ＋カーネル法（D40, 金川ら）
    - 厳密計算＋カーネル法（D46, 西山ら）



Collaborators

Bernhard Schölkopf (MPI)



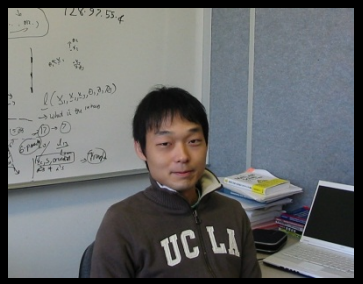
Arthur Gretton (UCL/MPI)

Bharath Sriperumbudur (Cambridge)

Le Song (Georgia Tech)

中込滋樹(ISM)

間野修平(ISM)



西山悠(ISM)

金川元信(奈良先端)