Kernel Bayes' Rule: Nonparametric Bayesian inference with kernels

Kenji Fukumizu

The Institute of Statistical Mathematics

NIPS 2012 Workshop
Confluence between Kernel Methods and Graphical Models

December 8, 2012 @Lake Tahoe

Introduction

- A new kernel methodology for nonparametric inference.
  - Kernel means are used in representing and manipulating the probabilities of variables.
  - "Nonparametric" Bayesian inference is also possible!
    - Completely nonparametric
    - Computation is done by linear algebra with Gram matrices.
    - Different from "Bayesian nonparametrics"

  → Today's main topic.

Outline

1. Kernel mean: a method for nonparametric inference
2. Representing conditional probability
3. Kernel Bayes' Rule and its applications
4. Conclusions

Kernel mean: representing probabilities

- Classical nonparametric methods for representing probabilities
  - Kernel density estimation: $\hat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K((x - X_i)/h_n)$
  - Characteristic function: $\mathrm{Ch.f}_X(u) = E[e^{iuX}]$, $\widehat{\mathrm{Ch.f}}_X(u) = \frac{1}{n}\sum_{i=1}^{n} e^{iuX_i}$

- New alternative: kernel mean

  X: random variable taking values on Ω, with probability P.

  k: positive definite kernel on Ω, H_k: RKHS associated with k.

  Def. Kernel mean of X on H_k:

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

$\Phi(x) = k(\cdot, x)$: feature vector

Empirical estimation: $\hat{m}_P = \frac{1}{n}\sum_{i=1}^{n} \Phi(X_i)$ for $X_1, \ldots, X_n \sim P$. i.i.d.

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

- Kernel mean has information of higher order moments of X

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

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

Moment generating function

Characteristic kernel

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

Def. A bounded measurable kernel k is characteristic, if

$$m_P = m_Q \iff P = Q.$$

- Kernel mean m_P with characteristic kernel k uniquely determines the probability.
- Examples: Gaussian, Laplace kernel (polynomial kernel is not)
- Analogous to the characteristic function $\mathrm{Ch.f}_X(u) = E[e^{iuX}]$.
  - Ch.f. uniquely determines the probability of X.
  - Positive definite kernel gives a better alternative:
    - efficient computation by kernel trick.
    - applicable to non-vectorial data.

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. NIPS 2006, JMLR 2012, NIPS 2009, 2012)

- Independence test → $m_{XY} = m_X \otimes m_Y$? (Gretton NIPS 2007)

- Bayesian inference

  → Estimate kernel mean of the posterior given kernel representation of prior and conditional probability.

- Conventional approaches to nonparametric inference
  - Smoothing kernel (not necessarily positive definite)

    Kernel density estimation, local polynomial fitting $h^{-d}K(x/h)$
  - Characteristic function: $E[e^{i\omega X}]$

    etc, etc, ...

→ "Curse of dimensionality"

e.g. smoothing kernel: difficulty for high (or several) dimension.

- Kernel methods for nonparametric inference
  - What can we do?
  - How robust to high-dimensionality?

Conditional probabilities

Conditional kernel mean

- Conditional probabilities are important to inference
  - Graphical modeling: conditional independence / dependence
  - Bayesian inference

- Kernel mean of conditional probability

$$E[\Phi(Y)|X = x] = \int \Phi(y)p(y|x)dy$$

- Question:
  - How can we estimate it in the kernel framework?
  - Accurate estimation of $p(y|x)$ is not easy.

→ Regression approach.

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 $C_{YX}: H_X \to H_Y$, $C_{XX}: H_X \to H_X$

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

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

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

- C_{YX} can be identified with the kernel mean $E[k_Y(\cdot, Y) \otimes k_X(\cdot, X)]$ on the product space $H_Y \otimes H_X$:

Conditional kernel mean

- Review: X, Y, Gaussian random variables ($\in \mathbf{R}^m, \mathbf{R}^\ell$, resp.)

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

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

- For general X and Y

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

$$\langle F, \Phi_X(X)\rangle_{H_X}$$

With characteristic kernel k_X,

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

↑

Conditional kernel mean given $X = x$

- Empirical estimation

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

$$\mathbf{k}_X(x) = (k_X(x, X_1), \dots, k_X(x, X_n))^T \in \mathbf{R}^n,$$

$$\mathbf{k}_Y(\cdot) = (k_Y(\cdot, Y_1), \dots, k_Y(\cdot, Y_n))^T \in H_Y^n,$$

ε_n: regularization coefficient

Note: joint sample $(X_1, Y_1), \dots, (X_n, Y_n) \sim P_{XY}$ is used to give the conditional kernel mean with $P_{Y|X}$.

c.f. kernel ridge regression

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

Kernel Bayes’ Rule

Inference with conditional kernel mean

- Sum rule: $q(y) = \int p(y|x)\pi(x)dx$
- Chain rule : $q(x, y) = p(y|x)\pi(x)$
- Bayes' rule: $q(x|y) = \frac{p(y|x)\pi(x)}{\int p(y|x)\pi(x)dx}$

- Kernelization
  - Express the probabilities by kernel means.
  - Express the statistical relations among variables with covariance operators.
  - Realize the above inference rules with Gram matrix computation.

Kernel Sum Rule

- Sum rule: $q(y) = \int p(y|x)\pi(x)dx$
- Kernelization: $m_Y = C_{YX}C_{XX}^{-1}m_\pi$
- Gram matrix expression

Joint sample

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

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

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

- Proof: $\int \Phi(y)p(y|x)dy = C_{YX}C_{XX}^{-1}\Phi(x)$

⬇ $\int \cdot\ \pi(x)dx$

$\int\int \Phi(y)p(y|x)\pi(x)dxdy = C_{YX}C_{XX}^{-1}m_\Pi$

Kernel Chain Rule

- Chain rule: $q(x,y) = p(y|x)\pi(x)$
- Kernelization: $m_Q = C_{(YX)X} C_{XX}^{-1} m_\pi$
- Gram matrix expression:

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

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

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

  From Sum Rule,

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

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

Kernel Bayes' Rule (KBR)

- Bayes' rule is regression $y \to x$ with probability $q(x, y) = p(y|x)\pi(x)$
- Kernel Bayes' Rule (KBR, Fukumizu et al NIPS2011)

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

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

Recall: Mean on the product space = Covariance

- Gram matrix expression:

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

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

$$w(y) = R_{X|Y} \mathbf{k}_Y(y),$$

$$R_{X|Y} = \Lambda G_{YY} \left((\Lambda G_{YY})^2 + \delta_n I_n \right)^{-1} \Lambda \mathbf{k}(y),$$
$$\Lambda = \mathrm{Diag}[(G_{XX} + n\varepsilon_n I_n)^{-1} G_{X\tilde{X}} \alpha]$$

Inference with KBR

- KBR estimates the kernel mean of the posterior $q(x|y)$, not itself.
- How can we use it for Bayesian inference?
  - Expectation: for any $f \in H_X$,

$$\mathbf{f}_X^T R_{X|Y} k_Y(y) \to \int f(x) q(x|y) dx. \qquad \text{(consistent)}$$

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

  - Point estimation:

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

(pre-image problem) solved numerically

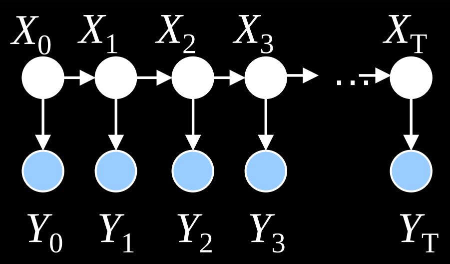
- Completely nonparametric way of computing Bayes rule.

  No parametric models are needed, but data or samples are used to express the probabilistic relations nonparametrically.

  Examples:

  1. Nonparametric HMM

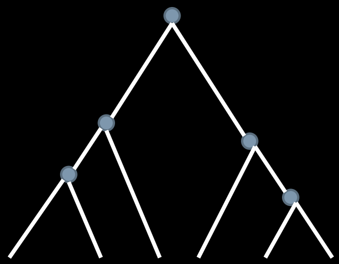
     See next.



  2. Kernel Approximate Bayesian Computation (Nakagome, F., Mano 2012)

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

     c.f. Approximate Bayesian Computation (ABC)

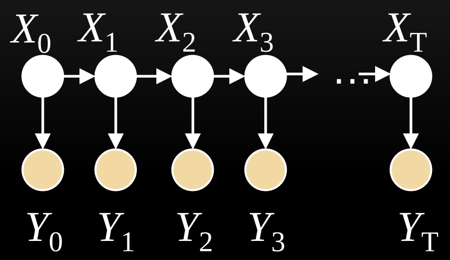


  3. Kernelization of Bellman equation in POMDP (Nishiyama et al UAI2012)

Example : KBR for nonparametric HMM

- Assume:

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

  $p(y_t|x_t)$ and/or $q(x_t|x_{t-1})$ 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: $\mathrm{argmin}_{x_s} \left\| \hat{m}_{x_s|\tilde{y}_0,\ldots,\tilde{y}_t} - \Phi(x) \right\|_{H_X}$

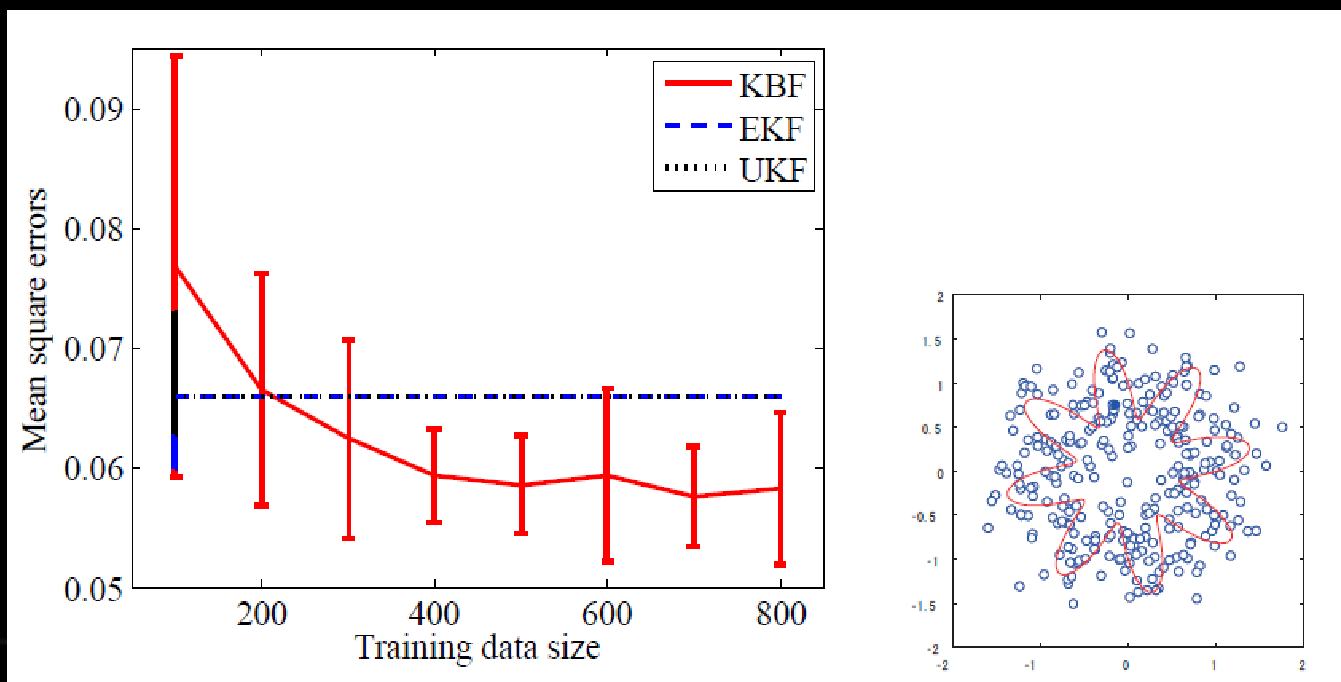
- General sequential inference uses Bayes' rule → KBR applied.

- Smoothing: noisy oscillation

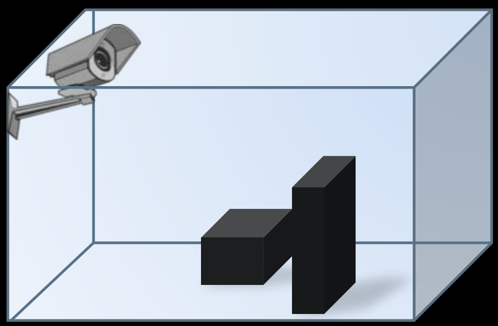
$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = (1 + 0.4\sin(8\theta_t)) \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} + Z_t, \qquad \theta_{t+1} = \arctan\left(\frac{v_t}{u_t}\right) + 0.4,$$

$$Y_t = (u_t, v_t)^T + W_t, \quad Z_t, W_t \sim N(0, 0.04 I_2)\ (i.i.d.)$$

Note: KBR does not know the dynamics, while the EKF and UKF use it.



- Rotation angle of camera
  - 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 $\pm$ 0.02
$\sigma^2 = 10^{-3}$	0.21$\pm$0.01	0.54 $\pm$ 0.02

Average MSE for camera angles (10 runs)

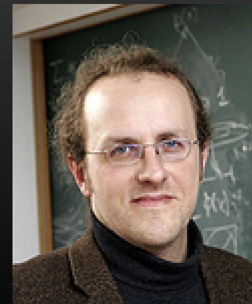
* For the rotation matrices, Tr[AB^{-1}] kernel for KBR, and quaternion expression for Kalman filter are used .

Concluding remarks

- "Kernel methods": useful, general tool for nonparametric inference.
    - Efficient linear algebraic computation with Gram matrices.
- Kernel Baeys' rule.
    - Inference with kernel mean of conditional probablity.
    - "Completely nonparametric" way for general Bayesian inference.

- Ongoing / future works
  - Combination of parametric model and kernel nonparametric method:
    - Exact integration + kernel nonparametrics (Nishiyama et al IBIS2012)
    - Particle filter + kernel nonparametrics (Kanagawa et al IBIS 2012)
  - Theoretical analysis in high-dimensional situation.
  - Relation to other recent nonparametric approaches?
    - Gaussian process
    - Bayesian nonparametrics

Collaborators



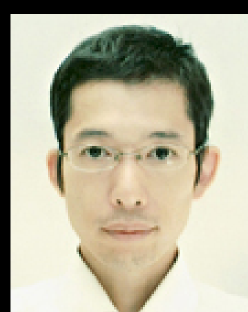
Bernhard Schölkopf (MPI)

Arthur Gretton (UCL/MPI)

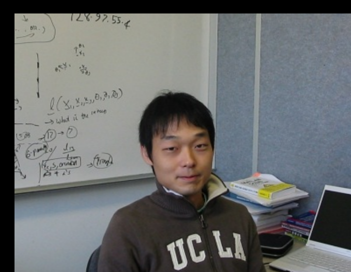
Le Song (Georgia Tech)

Bharath Sriperumbudur (Cambridge)

Shigeki Nakagome (ISM)



Shuhei Mano (ISM)



Yu Nishiyama (ISM)

Motonobu Kanagawa (NAIST)