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: \( \hat{\rho}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K((x - X_i)/h_n) \)
  - Characteristic function: \( \text{Ch. } f_X(u) = E[e^{iuX}], \text{ Ch. } \hat{f}_X(u) = \frac{1}{n} \sum_{i=1}^{n} e^{iuX_i} \)

- New alternative: kernel mean
  \( X \): random variable taking values on \( \Omega \), with probability \( P \).
  \( k \): positive definite kernel on \( \Omega \), \( H_k \): RKHS associated with \( k \).

**Def.** Kernel mean of \( X \) on \( H_k \):

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

\[\Phi(x) = k(\cdot, x): \text{feature vector}\]

Empirical estimation: \( \hat{m}_P = \frac{1}{n} \sum_{i=1}^{n} \Phi(X_i) \quad \text{for } X_1, \ldots, X_n \sim P \text{ i.i.d.} \)
• Reproducing expectation: \[ \langle f, m_P \rangle = E[f(X)] \quad \forall f \in H_k. \]

• Kernel mean has information of higher order moments of \( X \)
  e.g. \[ k(u, x) = c_0 + c_1ux + c_2(ux)^2 + \cdots \quad (c_i \geq 0), \quad \text{e.g., } e^{ux} \]
  \[ m_P(u) = c_0 + c_1E[X]u + c_2E[X^2]u^2 + \cdots \]
  Moment generating function
Characteristic kernel
(Fukumizu et al. JMLR 2004, AoS 2009; Sriperumbudur et al. JMLR 2010)

**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 $\text{Ch. } f_X(u) = E[e^{iuX}]$.
  - 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,

\[ \text{Inference on } P \quad \Rightarrow \quad \text{Inference on } m_P \]

• Two sample test \( \Rightarrow m_P = m_Q \) ? (Gretton et al. NIPS 2006, JMLR 2012, NIPS 2009, 2012)

• Independence test \( \Rightarrow m_{XY} = m_X \otimes m_Y \) ? (Gretton NIPS 2007)

• Bayesian inference
  \( \Rightarrow \) 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

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

$(H_X, k_X), (H_Y, k_Y) : \text{RKHS on } \Omega_X \text{ and } \Omega_Y, \text{ resp.}$

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

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

- Simply, extension of covariance matrix (linear map) $V_{YX} = E[XY^T]$
- Reproducing property:
  $$\langle g, C_{YX}f \rangle = E[f(X)g(Y)] \quad \text{for all } f \in H_X, g \in H_Y.$$
- $C_{YX}$ 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 ($\in \mathbb{R}^m, \mathbb{R}^\ell$, resp.)

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

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

• For general $X$ and $Y$

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

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

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

\[ \hat{E}[\Phi_Y(Y)|X = x] = k_Y^T(\cdot)(G_X + n\varepsilon_n I_n)^{-1}k_X(x) \]

\[ k_X(x) = (k_X(x, X_1), ..., k_X(x, X_n))^T \in \mathbb{R}^n, \]

\[ k_Y(\cdot) = (k_Y(\cdot, Y_1), ..., k_Y(\cdot, Y_n))^T \in H^n_Y, \]

\[ \varepsilon_n: \text{regularization coefficient} \]

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

c.f. kernel ridge regression

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

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

  Joint sample

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

  \[ G_X = \left(k(X_i, X_j)\right)_{i,j}, \quad G_{X\tilde{X}} = \left(k(X_i, \tilde{X}_j)\right)_{i,j} \]

• 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: \( m_Q = C_{(YX)X}C^{-1}_{XX}m_\pi \)

• Gram matrix expression:

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

  \[ \hat{m}_Q = \sum_{i=1}^{n} \beta_i \Phi(Y_i) \otimes \Phi(X_i), \quad \beta = (G_X + n\varepsilon_n I_n)^{-1}G_{XX}\alpha. \]

• Intuition: Note \( C_{(YX)X}: H_X \rightarrow H_Y \otimes H_X, \quad E[(\Phi(Y) \otimes \Phi(X)) \otimes \Phi(X)] \)

  From Sum Rule,

  \[ C_{(YX)X}C^{-1}_{XX}m_\pi = \int \int \Phi(y) \otimes \Phi(x)p(y|x)\delta(x - x')\pi(x')dydxdx' \]

  \[ = \int \int \Phi(y) \otimes \Phi(x)p(y|x)\pi(x)dydx = m_Q \]
Kernel Bayes’ Rule (KBR)

- Bayes’ rule is regression $y \rightarrow x$ with probability $q(x, y) = p(y|x)\pi(x)$
- Kernel Bayes’ Rule (KBR, Fukumizu et al NIPS2011)
  \[ m_{Qx|y} = C_{XY}^\pi C_{YY}^\pi \Lambda \Phi(y) \]
  where \( C_{YY}^\pi = C_{YY} C_{XX}^{-1} m_\pi \), \( C_{XX}^\pi = C_{XX} C_{XX}^{-1} m_\pi \)
  Recall: Mean on the product space = Covariance
- Gram matrix expression:
  \[
  \hat{m}_\pi = \sum_{i=1}^\ell \alpha_i \Phi(\tilde{X}_i), \quad (X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{XY},
  \]
  \[
  \hat{m}_{Qx|y} = \sum_{i=1}^n w_i(y) \Phi(X_i),
  \]
  \[
  w(y) = R_{X|Y} k_Y(y),
  \]
  \[
  R_{X|Y} = \Lambda G_{YY} ((\Lambda G_{YY})^2 + \delta_n I_n)^{-1} \Lambda k(y),
  \]
  \[
  \Lambda = \text{Diag}[(G_{XX} + n \varepsilon_n I_n)^{-1} G_{XX} \alpha]
  \]
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 $f \in H_X$,
    \[
    \mathbf{f}_X^T R_{X|Y} k_Y(y) \to \int f(x)q(x|y)dx. \quad \text{(consistent)}
    \]
    where $\mathbf{f}_X = (f(X_1), \ldots, f(X_n))^T$.
  - Point estimation:
    \[
    \hat{x} = \arg\min_x \| \mathbf{m}_{X|Y=y} - \Phi_X(x) \|_{H_X}
    \]
    (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:
  \[ p(X, Y) = p(X_0, Y_0) \prod_{t=1}^{T} p(Y_t|X_t)q(X_t|X_{t-1}) \]

  \[ p(y_t|x_t) \text{ and/or } q(x_t|x_{t-1}) \text{ is not known.} \]

  But, data \((X_t, Y_t)^T_{t=0}\) 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 \(\tilde{y}_0, ..., \tilde{y}_t\), estimate hidden state \(x_s\).

  \[ \text{KBR point estimator: } \arg\min_{x_s} \| m_{x_s|\tilde{y}_0, ..., \tilde{y}_t} - \Phi(x) \|_{H_X} \]

• General sequential inference uses Bayes’ rule \(\Rightarrow\) KBR applied.
• **Smoothing: noisy oscillation**

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

\[Y_t = (u_t, v_t)^T + W_t, \quad Z_t, W_t \sim N(0, 0.04I_2) \text{ (i.i.d.)}\]

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.

<table>
<thead>
<tr>
<th>noise</th>
<th>KBR (Trace)</th>
<th>Kalman filter(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 10^{-4}$</td>
<td>0.15 ±&lt; 0.01</td>
<td>0.56 ± 0.02</td>
</tr>
<tr>
<td>$\sigma^2 = 10^{-3}$</td>
<td>0.21±0.01</td>
<td>0.54 ± 0.02</td>
</tr>
</tbody>
</table>

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 probability.
  • “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

Arthur Gretton (UCL/MPI)

Bernhard Schölkopf (MPI)

Bharath Sriperumbudur (Cambridge)

Shigeki Nakagome (ISM)  Shuhei Mano (ISM)  Yu Nishiyama (ISM)

Motonobu Kanagawa (NAIST)  Le Song (Georgia Tech)