Kernel Method for Bayesian Inference

Kenji Fukumizu

The Institute of Statistical Mathematics, Tokyo.

Joint work with Le Song (CMU) and Arthur Gretton (Univ. College London, Max Planck Institute)

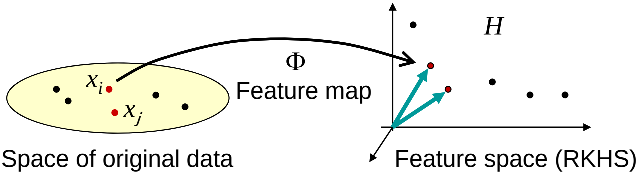
Nov. 10, 2010. ACML2010, Tokyo

(Slides revised Nov. 13)

Outline

Kernel Method in a Big Picture

- Kernel method = a systematic way of mapping data into a high-dimensional reproducing kernel Hilbert space (RKHS) to extract higher order moments or nonlinearity.



$$X \quad \Longrightarrow \quad \Phi(X) \qquad \text{(random vector on } \mathcal{H}\text{)},$$

- Linear statistical methods are applied on RKHS:
  SVM, kernel PCA, etc.

Overview: Inference with Kernel Mean

Basic statistics on RKHS are already useful.

- Kernel mean: $E[\Phi(X)]$ can characterize the probability of X.
- Applied to nonparametric statistical inference.
  - homogeneity test (Gretton et al. 2007),
  - independence test (Gretton et al 2008)
  - conditional independence test (Fukumizu et al 2008),
  - dimension reduction (F., Bach, Jordan, 2004, 2010), etc.

Overview: Kernel Bayesian Inference

- Bayes' rule:

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

- Of course, there are many ways of computing / approximating Bayes' rule. e.g. MCMC, importance sampling, sequential MC, variational method, EP, etc. Yet, its computation is challenging.

- This talk: kernel way of computing Bayes' rule.

  Express the kernel mean of posterior by that of posterior and likelihood.

Positive Definite Kernel

Def. Let $\mathcal{X}$ be a set. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if $k(x,y) = k(y,x)$ and for any $x_1, \ldots, x_n \in \mathcal{X}$ the symmetric matrix

$$(k(x_i, x_j))_{i,j=1}^n = \begin{pmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{pmatrix} \qquad \text{(Gram matrix)}$$

is positive semidefinite.

Examples. (on $\mathbb{R}^m$)

- Gaussian kernel: $\exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$.
- Polyn. kernel: $(x^T y + c)^d \quad (c \geq 0, d \in \mathbb{N})$.
  $\mathcal{H}_k = \{$poly. deg $\leq d\}$.

Reproducing Kernel Hilbert Space

Theorem (Moore-Aronszajn (1950))

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ (or $\mathbb{R}$) be a positive definite kernel on a set $\mathcal{X}$. Then, there uniquely exists a Hilbert space $\mathcal{H}_k$ consisting of functions on $\mathcal{X}$ such that

1. $k(\cdot, x) \in \mathcal{H}_k$ for every $x \in \mathcal{X}$,
2. $\mathrm{Span}\{k(\cdot, x) \mid x \in \mathcal{X}\}$ is dense in $\mathcal{H}_k$,
3. k is the reproducing kernel on $\mathcal{H}_k$, i.e.

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = f(x) \qquad (\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k).$$

(reproducing property)

RKHS is used as a feature space, which may be infinite dimensional.

Data Analysis with Positive Definite Kernels

- Feature map: mapping random variable or data:

$$X \quad \mapsto \quad \Phi(X) = k(\cdot, X) \qquad \text{random variable on } \mathcal{H}_k,$$

$$\mathcal{X} \ni X_1, \ldots, X_n \quad \mapsto \quad \Phi(X_1), \ldots, \Phi(X_n) \in \mathcal{H}_k$$

- Kernel trick: inner product is easily computable.

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j) \qquad \text{(Gram matrix)}$$

- Linear methods are extendable to RKHS with efficient computation.
- Typically, problems can be reduced to Gram matrices of sample size.

Mean and Covariance on RKHS I

$X \sim P$: random variable on $\mathcal{X}$. $k_{\mathcal{X}}$: pos. def. kernel on $\mathcal{X}$.

- Def. m_P: kernel mean of X on $\mathcal{H}_k$

$$m_P := E[\Phi(X)] = E[k(\cdot, X)] = \int k(\cdot, x)dP(x) \quad \in \mathcal{H}_k.$$

- Fact: $\langle f, m_P \rangle = E[f(X)]$. (reproducing property)

- m_P expresses higher-order moments of X.
  e.g. suppose $k(u, x) = c_0 + c_1(ux) + c_2(ux)^2 + \cdots$ ($c_i > 0$).

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

Mean and Covariance on RKHS II

(X, Y): random vector on $\mathcal{X} \times \mathcal{Y}$, $\sim P$. $k_{\mathcal{X}}, k_{\mathcal{Y}}$: pos. def. kernels on $\mathcal{X}, \mathcal{Y}$ (resp).

- Def. (uncentered) cross-covariance operator

$$C^P_{\mathcal{Y}\mathcal{X}} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}, \qquad \langle g, C^P_{\mathcal{Y}\mathcal{X}} f \rangle = E[g(Y)f(X)].$$

- Covariance operator

$$C^P_{\mathcal{X}\mathcal{X}} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{X}}, \quad \langle h, C^P_{\mathcal{X}\mathcal{X}} f \rangle = E[h(X)f(X)].$$

- (Cross-)covariance operator = mean in the product space.

$$C^P_{\mathcal{Y}\mathcal{X}} \iff m_P = E[k_{\mathcal{Y}}(\cdot, Y) \otimes k_{\mathcal{X}}(\cdot, X)] \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}.$$

$$\because \ \langle g, C^P_{\mathcal{Y},\mathcal{X}} f \rangle = \langle g \otimes f, m_P \rangle.$$

Mean and Covariance on RKHS III

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

- Empirical Estimation:

$$\widehat{m}_X = \frac{1}{n} \sum_{i=1}^{n} k_{\mathcal{X}}(\cdot, X_i),$$

$$\widehat{C}_{YX} = \frac{1}{n} \sum_{i=1}^{n} k_{\mathcal{Y}}(\cdot, Y_i) \otimes k_{\mathcal{X}}(\cdot, X_i).$$

- Typically, Gram matrix expression is obtained.

- $O_p(n^{-1/2})$-consistency in RKHS-norm is guaranteed (Gretton et al. 2005, etc).

Characteristic Kernel: Representing Class

$\mathcal{P}$: the set of all probabilities on a measurable space $(\mathcal{X}, \mathcal{B})$.

Def. (F., Bach, Jordan 2004, 2009) k is called characteristic if

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

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

- Example. Gaussian kernel, Laplacian kernel. (Sriperumbudur et al. 2010)
- With characteristic kernels,

$$\text{Inference on } P \quad \Longrightarrow \quad \text{Inference on } m_P$$

  - two sample test $\Rightarrow$ $m_P = m_Q$?
  - independence test $\Rightarrow$ $m_{XY} = m_X \otimes m_Y$?
- Hereafter, all kernels are assumed to be characteristic.

Bayes' Rule

Bayes' Rule:

$$q(x|y) = \frac{q(x,y)}{q_{\mathcal{Y}}(y)} = \frac{p(y|x)\pi(x)}{q_{\mathcal{Y}}(y)}, \quad q_{\mathcal{Y}}(y) = \int q(x,y)dx.$$

Π: prior (p.d.f. $\pi(x)$).
P: joint distribution to give likelihood $p(y|x)$.

Kernel realization:

- Given m_Π (kernel mean of Π) and $C^P_{\mathcal{Y}\mathcal{X}}, C^P_{\mathcal{X}\mathcal{X}}$ (covariance operators of $p(x,y)$), express the kernel mean of the posterior

$$m_{Q_{\mathcal{X}|y}} := \int k_{\mathcal{X}}(\cdot, x) q(x|y) dx$$

Conditional Probabilities with Kernels I

- Basic Proposition (F., Bach, Jordan 2004)
  If $E[g(Y)|X = \cdot] \in \mathcal{H}_\mathcal{X}$ for $g \in \mathcal{H}_\mathcal{Y}$, then

$$C^P_{\mathcal{X}\mathcal{X}} E[g(Y)|X = \cdot] = C^P_{\mathcal{X}\mathcal{Y}} g.$$

$\because$) $\langle f, C^P_{\mathcal{X}\mathcal{X}} E[g(Y)|X = \cdot] \rangle = E[f(X)E[g(Y)|X]] = E[f(X)g(Y)] = \langle f, C^P_{\mathcal{X}\mathcal{Y}} g \rangle.$

- Expression of kernel mean of conditional probability $p(y|x)$:

$$E[k_\mathcal{Y}(\cdot, Y)|X = x] = C^P_{\mathcal{Y}\mathcal{X}} {C^P_{\mathcal{X}\mathcal{X}}}^{-1} k_\mathcal{X}(\cdot, x).$$

(A bit naive, but can be justified.)

$\because$) $E[g(Y)|X = \cdot] = C^{-1}_{\mathcal{X}\mathcal{X}} C_{\mathcal{X}\mathcal{Y}} g \quad \Longrightarrow$

$\langle g, E[k_\mathcal{Y}(\cdot, Y)|X = x] \rangle = \langle C^{-1}_{\mathcal{X}\mathcal{X}} C_{\mathcal{X}\mathcal{Y}} g, k_\mathcal{X}(\cdot, x) \rangle = \langle g, C_{\mathcal{Y}\mathcal{X}} C^{-1}_{\mathcal{X}\mathcal{X}} k_\mathcal{X}(\cdot, x) \rangle.$

Kernel Mean of Posterior

Q: joint probability with p.d.f. $q(x,y) = p(y|x)\pi(x)$.

Kernel mean of posterior

$$m_{Q_{\mathcal{X}|y}} := E_Q[k_{\mathcal{X}}(\cdot,X)|Y=y] = C^Q_{\mathcal{X}\mathcal{Y}} {C^Q_{\mathcal{Y}\mathcal{Y}}}^{-1} k_{\mathcal{Y}}(\cdot,y).$$

- Ingredients:

$$m_{Q_{\mathcal{Y}}} = C^P_{YX} {C^P_{XX}}^{-1} m_\Pi \quad \longleftrightarrow \quad q_{\mathcal{Y}}(y) = \int p(y|x)\pi(x)dx.$$

Recall: covariance = mean of product.

$$C^Q_{\mathcal{X}\mathcal{Y}} \iff \qquad m_Q = C^P_{(\mathcal{X}\mathcal{Y})\mathcal{X}} {C^P_{\mathcal{X}\mathcal{X}}}^{-1} m_\Pi \quad \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}},$$

$$C^Q_{\mathcal{Y}\mathcal{Y}} \iff \quad m^Q_{\mathcal{Y}\times\mathcal{Y}} = C^P_{(\mathcal{Y}\mathcal{Y})\mathcal{X}} {C^P_{XX}}^{-1} m_\Pi \quad \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}},$$

$$C^P_{(\mathcal{X}\mathcal{Y})\mathcal{X}} = E_P\big[\big(k_{\mathcal{X}}(\cdot,X)\otimes k_{\mathcal{Y}}(\cdot,Y)\big)\otimes k_{\mathcal{X}}(\cdot,X)\big] : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{Y}},$$

$$C^P_{(\mathcal{Y}\mathcal{Y})\mathcal{X}} = E_P\big[\big(k_{\mathcal{Y}}(\cdot,Y)\otimes k_{\mathcal{Y}}(\cdot,Y)\big)\otimes k_{\mathcal{X}}(\cdot,X)\big] : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}\otimes\mathcal{H}_{\mathcal{Y}}.$$

Kernel Bayes' Rule

Kernel Bayes' Rule

$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$, i.i.d. $\widehat{m}_\Pi = \sum_{j=1}^{\ell} \gamma_j k_\mathcal{X}(\cdot, U_j)$.
Gram matrix expression of the kernel mean of posterior is

$$\widehat{m}_{Q_\mathcal{X}|y} = \sum_{i=1}^{n} w_i(y) k_\mathcal{X}(\cdot, X_i), \quad w(y) = L_Y (L_Y^2 + \delta_n I_n)^{-1} \Lambda \mathbf{k}_Y(y),$$

for any $y \in \mathcal{Y}$, where

$$L_Y = \Lambda G_Y, \quad \Lambda = \mathrm{Diag}\big((G_X + n\varepsilon_n I_n)^{-1} G_{XU}\gamma\big),$$
$$\mathbf{k}_Y = (k_\mathcal{Y}(\cdot, Y_1), \ldots, k_\mathcal{Y}(\cdot, Y_n))^T,$$
$$G_X = (k_\mathcal{X}(X_i, X_j)), G_Y = (k_\mathcal{Y}(Y_i, Y_j)), G_{XU} = (k_\mathcal{X}(X_i, U_j)),$$

and ε_n, δ_n are regularization constants.

- The posterior is given by a weighted sample (X_i, w_i), while the weights may not be positive.

Consistency

Theorem

Assumptions:

- $\pi/p_X \in \mathcal{R}(A_X {C^P_{\mathcal{X}\mathcal{X}}}^{1/2})$.
- $\|\widehat{m}_\Pi - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ $(n \to \infty)$ for some $0 < \alpha \leq 1/2$.
- $A_Y : \mathcal{H}_\mathcal{Y} \to L^2(P_Y), f \mapsto f$ is injective.
- $E[f(X)|Y = \cdot] \in \mathcal{H}_\mathcal{Y}$ for any $f \in \mathcal{H}_\mathcal{X}$, and $S : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$, $f \mapsto E[f(X)|Y = \cdot]$ makes $(C^Q_{\mathcal{Y}\mathcal{Y}})^{-\nu} S$ bounded for $\nu > 0$.

With $\varepsilon_n = n^{-\frac{2}{3}\alpha}$ and $\delta_n = n^{-\max\{\frac{4}{15}\alpha, \frac{4}{3(\nu+3)}\alpha\}}$, for any $y \in \mathcal{Y}$

$$\left\|\widehat{m}_{Q_\mathcal{X}|y} - m_{Q_\mathcal{X}|y}\right\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\min\{\frac{4}{15}\alpha, \frac{2\nu}{3(\nu+3)}\alpha\}}), \quad (n \to \infty).$$

Note: the rate does not depend on the dimensionality.

c.f. Kernel density estimation.

Kernel Bayesian Inference I

Kernel Bayesian Inference (KBI) = 'Nonparametric' Bayesian inference using kernel Bayes' rule.

- Likelihood $p(y|x)$ and the prior $\pi(x)$ are given by samples.

Case I Explicit form of likelihood $p(y|x)$ is unavailable, but sampling from $p(y|x)$ is easy.
c.f. Approximate Bayesian Computation (ABC).

Case II Likelihood $p(y|x)$ is unknown, but sample from $p(x, y)$ is given in training phase (discussed later).

- If both of $p(y|x)$ and $\pi(x)$ are known, there are many good numerical / approximation methods, such as MCMC, SMC, variational Bayes, etc.

Kernel Bayesian Inference II

- Kernel Bayesian Inference estimates the kernel mean $m_{Q_{\mathcal{X}|y}} = \int k_{\mathcal{X}}(\cdot, x) q(x|y) dx$, not the posterior $q(x|y)$ itself.
- How to use for inference?
  - Expectation: for $f = \sum_{i=1}^{n} f_i k_{\mathcal{X}}(\cdot, X_i)$,

$$\int f(x) q(x|y) dx \quad \longleftarrow \quad \sum_{i=1}^{n} f_i w_i(y).$$

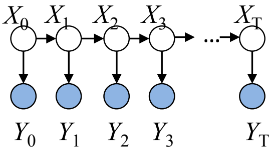
  - Approximate MAP solution:

$$\max_x \widehat{m}_{Q_{\mathcal{X}|y}}(x)$$

    may be solved iteratively.
- Time complexity: matrix inversion costs $O(n^3)$, but with low-rank (r) approximation, KBI costs $O(nr^2)$.

KBI for Nonparametric Hidden Markov Model

Model: $p(X,Y) = \pi(X_1)\prod_{t=1}^{T} p(Y_t|X_t)\prod_{t=1}^{T-1} q(X_{t+1}|X_t)$,



- Assume
  - $p(y|x)$ and/or $q(x|x')$ is not known.
  - But, sample $(X_t, Y_t)_{t=1}^{T}$ is available in training phase.
- Testing phase:
  - given $\tilde{y}_1, \ldots, \tilde{y}_t$, compute $\max_{x_s} p(x_s|\tilde{y}_1, \ldots, \tilde{y}_t)$.

    $\Longrightarrow$ Kernel Bayesian inference: $\max_{X_s} \widehat{m}_{x_s|\tilde{y}_1,\ldots,\tilde{y}_t}$.
- *E.g.* when measurement of hidden states is expensive, or when hidden states are measured with time delay in predicting future state.

- Sequential filtering:

$$\widehat{m}_{x_t|\tilde{y}_1,\ldots,\tilde{y}_t} = \sum_{i=1}^{T} \alpha_i^{(t)} k_{\mathcal{X}}(\cdot, X_i), \quad \alpha^{(t)} = \alpha^{(t)}(\tilde{y}_1, \ldots, \tilde{y}_t).$$

- Update rule:

$$\widehat{\mu}^{(t+1)} = (G_X + T\varepsilon_T I_T)^{-1} G_{X,X_{+1}} (G_X + T\varepsilon_T I_T)^{-1} G_X \alpha^{(t)}.$$

$$\alpha^{(t+1)} = L_Y^{(t+1)} \left((L_Y^{(t+1)})^2 + \delta_T I_T \right)^{-1} \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1}).$$

$G_{X,X_{+1}}$: "transfer" matrix $\left(G_{X,X_{+1}}\right)_{ij} = k_{\mathcal{X}}(X_i, X_{j+1})$.
$\Lambda^{(t+1)} = \mathrm{diag}(\widehat{\mu}_1^{(t+1)}, \ldots, \widehat{\mu}_T^{(t+1)})$ and $L_Y^{(t+1)} = \Lambda^{(t+1)} G_Y$,

- Prediction and smoothing are similar.
- The computational cost for each update is $O(Tr^2)$, once low-rank (r) approximation is used for training sample.

Comparison with Approx. Bayesian Computation

Assume: $p(y|x)$ is not explicitly known, but sampling is possible.

Approximate Bayesian Computation (ABC): existing method of sampling from posterior.

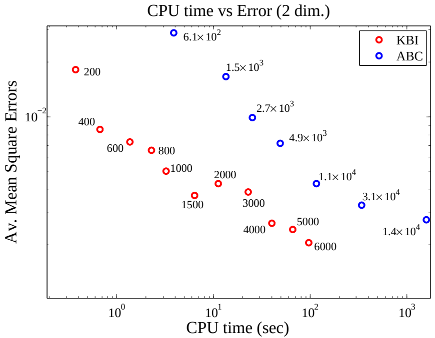
- Procedure (sampling and rejection):
  1. y given.
  2. Sample $X_i \sim \Pi$. $Y_i \sim p(Y|X_i)$.
  3. If $d(y, Y_i) < \epsilon$, accept X_i.
  4. Repeat 2 and 3.

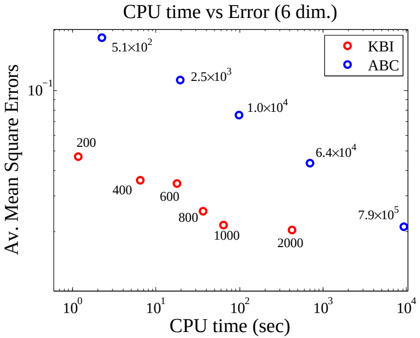
  Accepted sample $X_1, \ldots, X_N \sim q(X|y)$ approximately.

- Exact if $\varepsilon \to 0$, but acceptance rate is small particularly when the dimension of X is large.
- Proposed and used mainly in population genetics.

Experimental results

- task: $E[X|Y=y]$, evaluated at 10 different points of y. 10 random runs.
- Gaussian prior and likelihood so that the truth can be calculated.
- Gaussian kernels are used for KBI.
- Incomplete Cholesky is used for the low-rank approximation in KBI ($\varepsilon_n =$ tolerance $\propto 1/n, \delta_n = 2\varepsilon_n$).





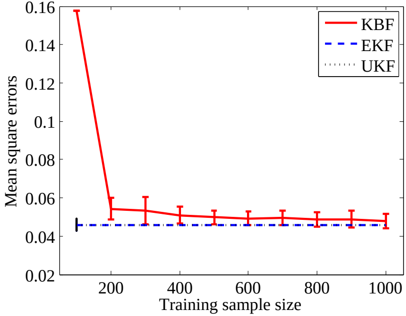
Experiments on Nonparametric Filtering

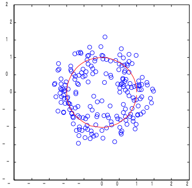
(a) Noisy rotation

$$\begin{cases} \binom{u_{t+1}}{v_{t+1}} = \binom{\cos\theta_{t+1}}{\sin\theta_{t+1}} + Z_t, \qquad \theta_{t+1} = \arctan(v_t/u_t) + 0.3, \\ Y_t = (u_t, v_t)^T + W_t, \end{cases}$$

$$Z_t \sim N(0, 0.2^2 I_2), W_t \sim N(0, 0.2^2 I).$$

Approximate MAP solution are computed by KBI.



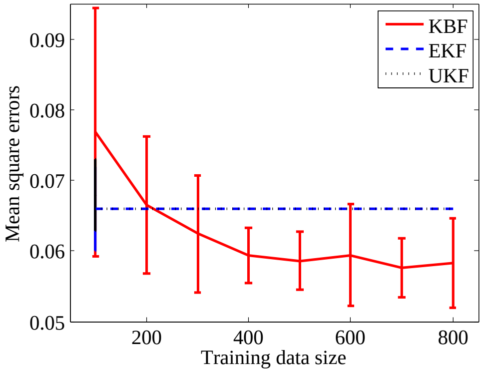


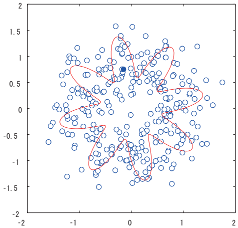
Note: KBI does not know the dynamics, while EKF and UKF use the exact knowledge.

(b) Noisy oscillation

$$\begin{cases} \begin{pmatrix} u_{t+1} \\ v_{t+1} \end{pmatrix} = (1 + 0.4\sin(8\theta_{t+1}))\begin{pmatrix} \cos\theta_{t+1} \\ \sin\theta_{t+1} \end{pmatrix} + Z_t, \quad \theta_{t+1} = \arctan(v_t/u_t) + 0.4, \\ Y_t = (u_t, v_t)^T + W_t, \end{cases}$$

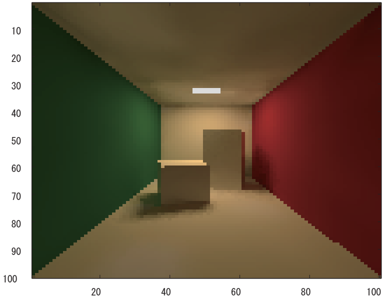
$$Z_t \sim N(0, 0.2^2 I_2), W_t \sim N(0, 0.2^2 I).$$





Estimation of Camera Angle

- Hidden X_t: angle of a camera.
- Observed Y_t: movie frame of a room + additive Gaussian noise.
- Data: Synthesized by POV-Ray (http://www.povray.org).
  X_t: 3600 downsampled frames of 20×20 RGB pixels (1200 dim.). The first 1800 frames are used for training, and the second half is used for test.



Results

	$\mathbb{R}^9$		$SO(3)$	
	KBI (Gauss)	Kalman ($\mathbb{R}^9$)	KBI (Tr)	Kalman (Q^*)
$\sigma^2 = 10^{-4}$	0.21 ± 0.02	1.98 ± 0.08	$0.15 \pm < 0.01$	0.56 ± 0.02
$\sigma^2 = 10^{-3}$	0.22 ± 0.01	1.94 ± 0.06	0.21 ± 0.01	0.54 ± 0.02

Average MSE of estimating camera angles (10 runs)

- For Kalman filter, dynamics is estimated with linear Gaussian model.
- In $\mathbb{R}^9$ model, Gaussian kernel for KBI.
- In $SO(3)$ model, $\mathrm{Tr}[AB]$ for KBI, and quaternion expression for Kalman filter.

Conclusions

- Kernel Bayes' rule: kernel way of realizing Bayes' rule nonparametrically.

  Kernel mean of posterior can be computed given samples from the prior and likelihood.

  - No explicit form of the likelihood is needed.
  - Consistency is guaranteed, and the rate does not depend on the dimensionality.
  - Computational cost is linear w.r.t. sample size if low-rank approximation is used.

- Future / on-going works:
  - Kernel (parameter) choice?
  - Applications to various Bayesian inference.
  - Combination of parametric and non-parametric HMM (parametric for transition, nonparametric for observation).

Thank you!

References

[1] Fukumizu, K., L. Song and A. Gretton. (2010) Kernel Bayes' rule. arXiv:1009.5736v2 [stat.ML]

[2] Song, L., J. Huang, A. Smola and K. Fukumizu. (2009) Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. *Proc. ICML2009*, 961–968.

[3] Fukumizu, K., F.R. Bach and M.I. Jordan. (2009) Kernel dimension reduction in regression. *Ann. Stat.* 37(4), pp.1871–1905.

[4] Fukumizu, K., F.R. Bach and M.I. Jordan. (2004) Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *JMLR.* 5, pp.73–99.