Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems

Le Song Jonathan Huang Alex Smola Kenji Fukumizu

Carnegie Mellon University Yahoo! Research Institute of Statistical Mathematics

Hilbert Space Embeddings

Summary Statistics

- Summary statistics for distributions:

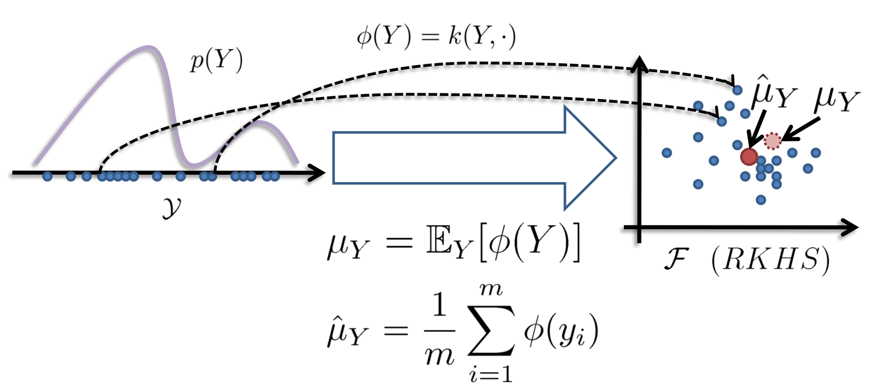
$\mathbb{E}_{Y\sim p}[Y]$ (mean)

$\mathbb{E}_{Y\sim p}[Y^2]$ (variance)

$\mathbb{E}_{Y\sim p}[\phi(Y)]$ (expected feature)

- For certain Á, can reconstruct *p* from the summary!

Hilbert Space Embeddings

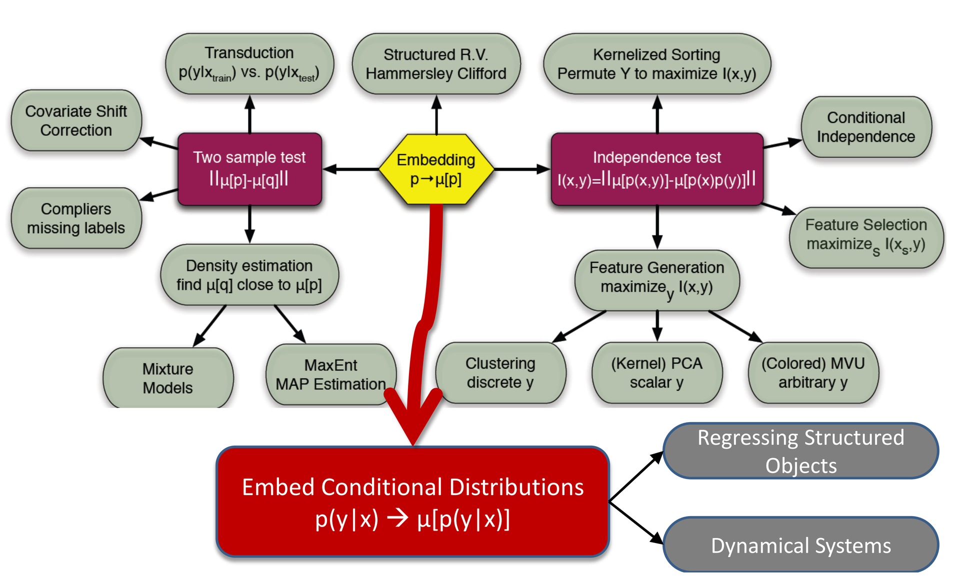


$$\mu_Y = \mathbb{E}_Y[\phi(Y)]$$

$$\hat{\mu}_Y = \frac{1}{m}\sum_{i=1}^{m}\phi(y_i)$$

- Can reconstruct $p(Y)$ from μ_Y for certain kernels!
- Sample average converges to true mean

Applications



Estimating Embedding for $p(Y|X)$

- Simple case: X binary, Y vector valued
- Can embed two distributions

$$p(Y|X=0) \rightarrow \mu_{Y|X=0} = \mathbb{E}_{Y|X=0}[\phi(Y)]$$

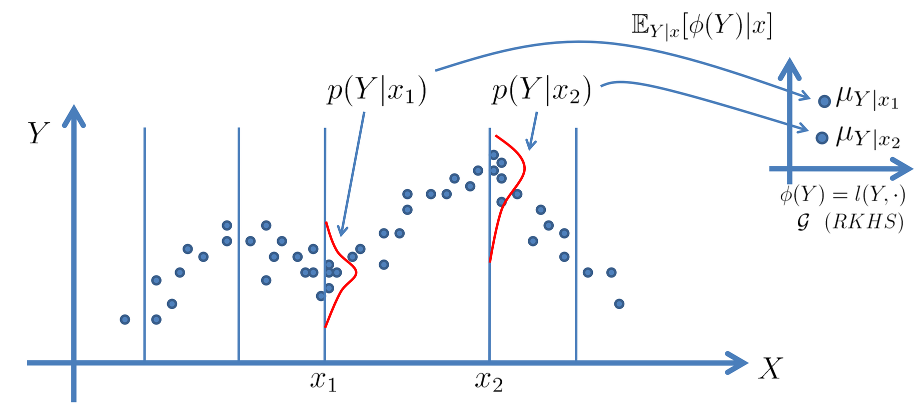
$$p(Y|X=1) \rightarrow \mu_{Y|X=1} = \mathbb{E}_{Y|X=1}[\phi(Y)]$$

$X \rightarrow Y$

What if X is *continuous or structured*?

Conditional Embeddings

Conditional Distributions



- Goal: Estimate embedding from finite sample
- Problem: How can we embed p(**Y**|**X**=*x*) for **unobserved** *x*?

Conditional Embedding

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$

- Kernel estimate (*m-by-m* matrix)

$$\hat{\mathcal{U}}_{Y|X} = \Phi(K + \lambda I)^{-1}\Upsilon^{T}$$

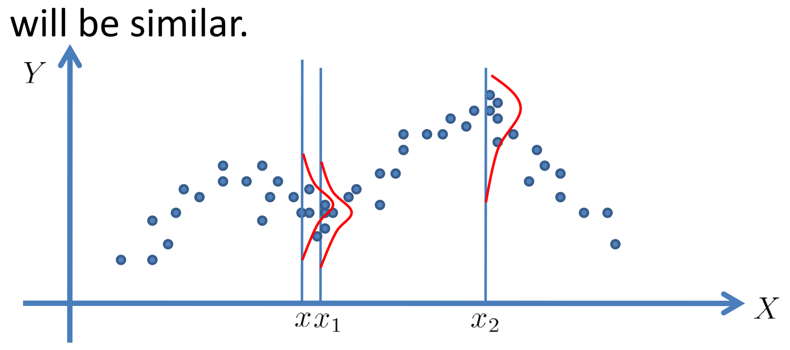
regularization parameter

$\Phi := (\phi(y_1), \ldots, \phi(y_m))$, $L = \Phi^\top\Phi$ Feature/Gram matrices for training y_i

$\Upsilon := (\varphi(x_1), \ldots, \varphi(x_2))$, $K = \Upsilon^\top\Upsilon$ Feature/Gram matrices for training x_i

Key Idea

- If x and x' are close, $p(Y|x)$ and $p(Y|x')$ will be similar.



Generalize to unobserved *x* via a kernel on *X*

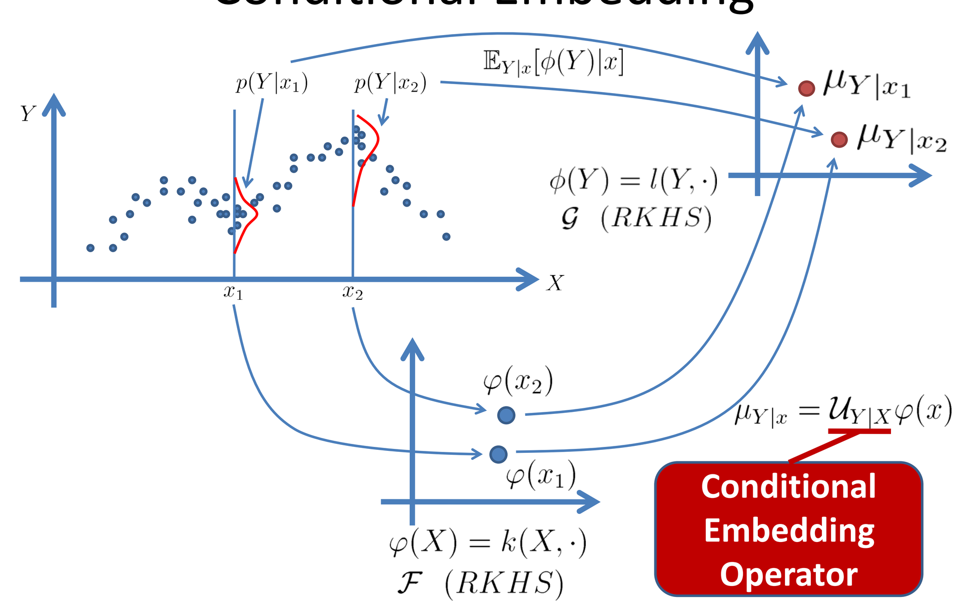
Sum and Product Rules

	Probabilistic Relation	Hilbert Space Relation
Sum Rule	$p(X) = \int_Y p(X\|Y)p(Y)$	$\mu_Y = \mathcal{U}_{Y\|X}\mu_X$
Product Rule	$p(X,Y) = p(Y\|X)p(X)$	$\mu_{XY} = \mathcal{U}_{Y\|X}\mu_X^{\otimes}$

$\mu_X^{\otimes} := \mathbb{E}_X[\varphi(X) \otimes \varphi(X)]$

Can do *probabilistic inference* with embeddings!

Conditional Embedding



"Ordinary" RKHS embedding

$$\hat{\mu}_Y = \sum_{i=1}^{m}\left(\frac{1}{m}\right)\phi(y_i)$$

Conditional embedding

$$\hat{\mu}_{Y|X=x} = \hat{\mathcal{U}}_{Y|X}\cdot\varphi(x) = \sum_{i=1}^{m}\beta_i\phi(y_i)$$

Uniform weights

Training feature vectors

Non-uniform weights

Defining the Conditional Embedding Operator

- Desired properties:
  1. $\mu_{Y|x} = \mathcal{U}_{Y|X}\varphi(x)$
  2. $\mathbb{E}_{Y|x}[g(Y)|x] := \langle g, \mu_{Y|x}\rangle = \langle g, \mathcal{U}_{Y|X}\varphi(x)\rangle \quad \forall g \in \mathcal{G}$
- Covariance operator:

$$\mathcal{C}_{XX}\mathbb{E}_{Y|X}[g(Y)|X] = \mathcal{C}_{XY}g$$

$$\underbrace{\mathcal{C}_{XY} := \mathbb{E}_{XY}[\varphi(X)\otimes\phi(Y)] - \mathbb{E}_X[\varphi(X)]\otimes\mathbb{E}_Y[\phi(Y)]}_{\text{generalization of covariance matrix}}$$

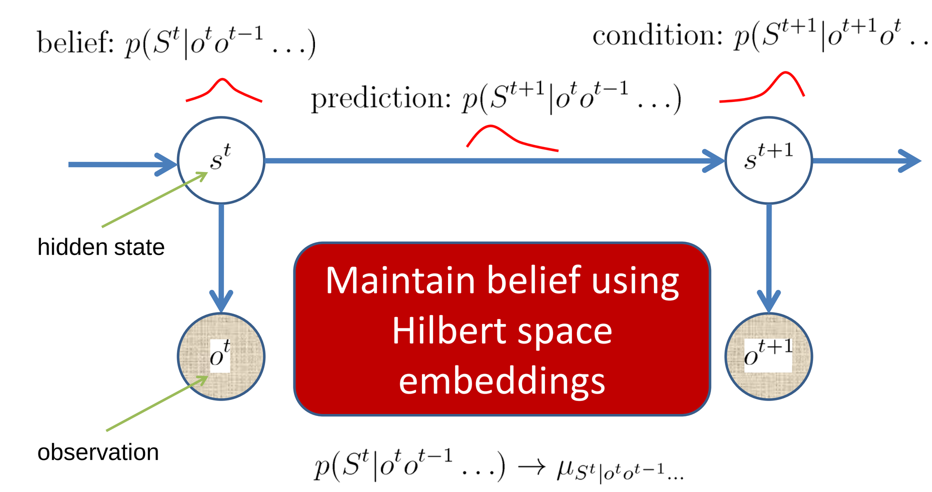
Consistency Guarantee

Theorem: $\hat{\mu}_{Y|X=x}$ converges to $\mu_{Y|X=x}$ at the rate: $O((m\lambda)^{-1/2} + \lambda^{1/2})$

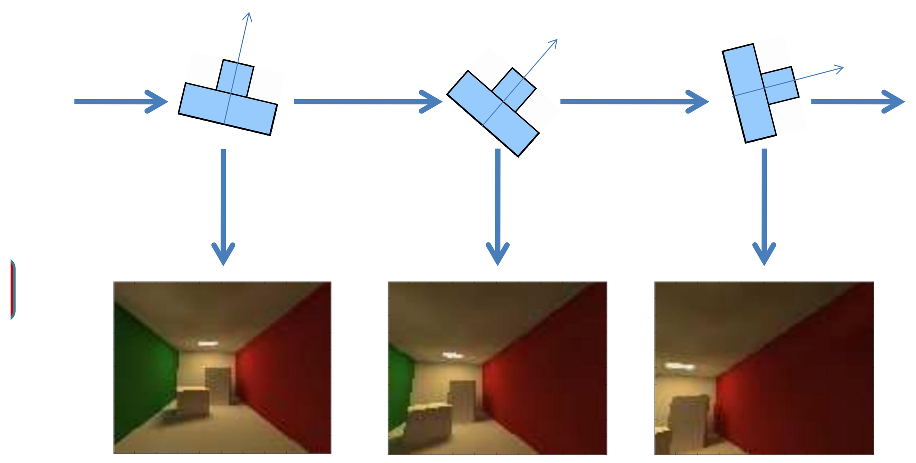
- Conditional embeddings *harder to estimate than ordinary embeddings*
- To obtain bias ϵ^2, fix $\lambda \sim O(\epsilon^2)$ to obtain $O(m^{-1/2})$ convergence

Dynamical Systems

Dynamical Systems



Camera Rotation

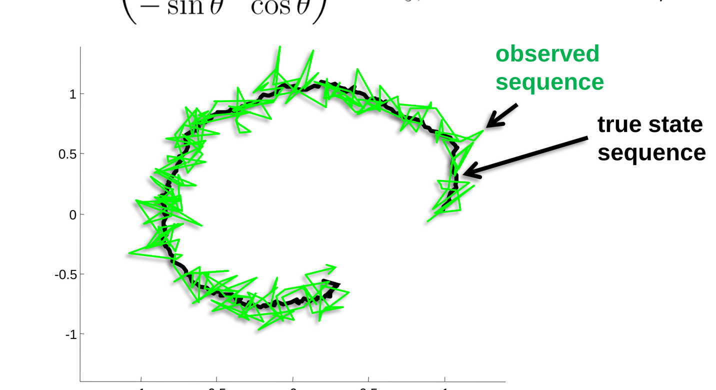


Rendered using POV-ray renderer

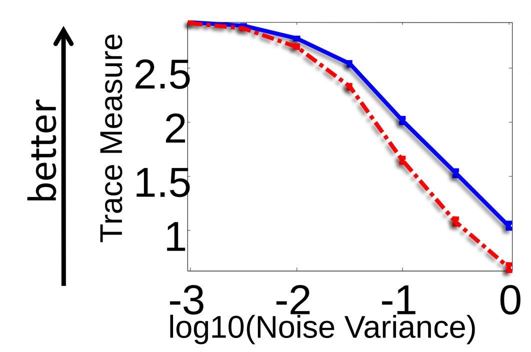
Evaluation with Linear Dynamics

- Linear dynamics model (Kalman filter optimal):

$$s^{t+1} = \begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}s^t + \xi, \quad o^{t+1} = s^{t+1} + \eta$$



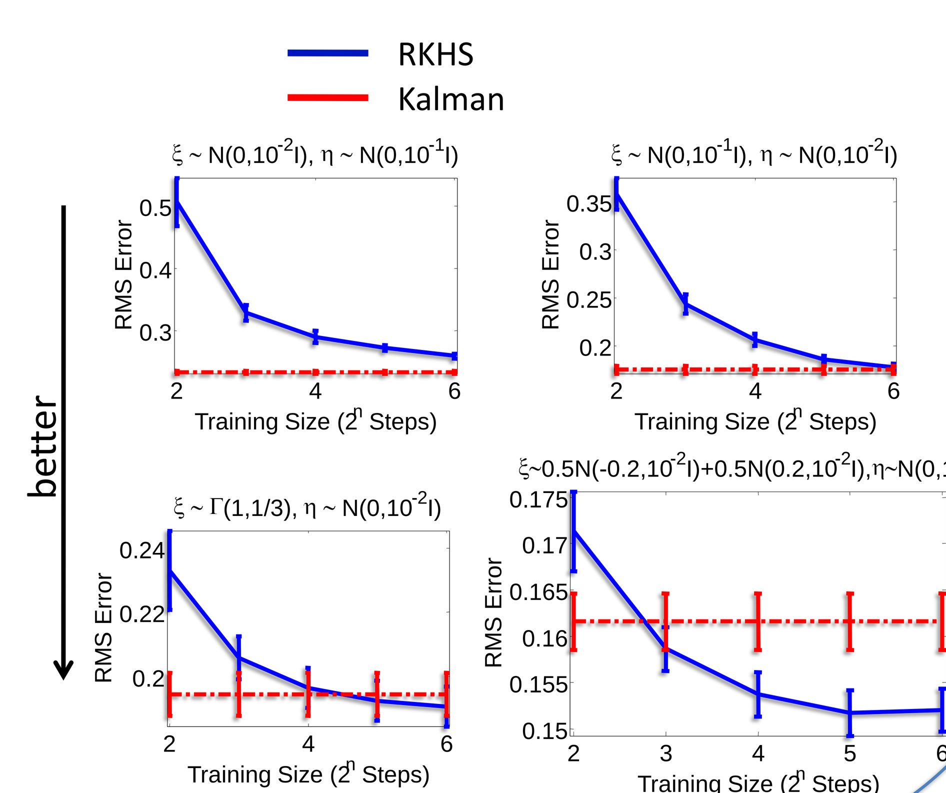
Camera Rotation



Dynamical systems with Hilbert space embeddings works better

trace measure: $\mathrm{tr}R^\top R'$
the larger the better

	Ground truth	RKHS	Kalman filter	Random
Trace measure	3	2.91	2.18	0.01



Advantages

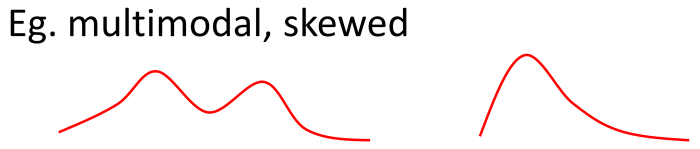
- Applies to *any* domain with a kernel

Eg. reals, rotations, permutations, strings

$$R = \begin{pmatrix}1 & 0 & 0\\ 0 & \cos\theta & \sin\theta\\ 0 & \sin\theta & \cos\theta\end{pmatrix} \quad \pi = \begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0\end{pmatrix} \quad AGCT\ldots$$

- Applies to more general distributions

Eg. multimodal, skewed



Discussion

Future Directions

- Learn dynamical systems from sequences of observations (without the hidden states)
- Generalize inference procedures to handle more complicated models (e.g., Bayesian networks, Markov random fields)
- Formulate conditional independence tests
- Establish sample complexity bounds for the dynamical systems setting

Conclusion

- Embedding conditional distributions
  - Generalizes to unobserved X using kernels
- Theoretical analysis of empirical estimator convergence
- Application to dynamical systems learning and inference