Hilbert Space Embeddings of Hidden Markov Models

ICML 2010 Best Paper

Le Song, Byron Boots, Sajid Siddiqi,
Geoff Gordon and Alex Smola

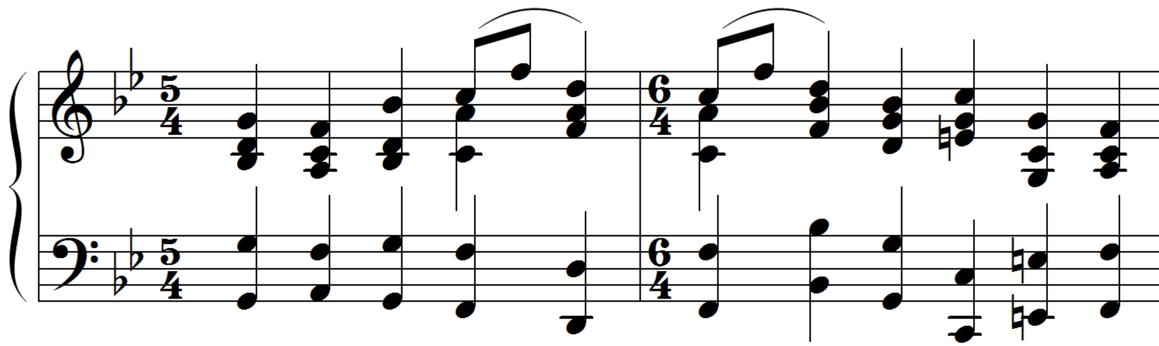
presented by Shouyuan Chen

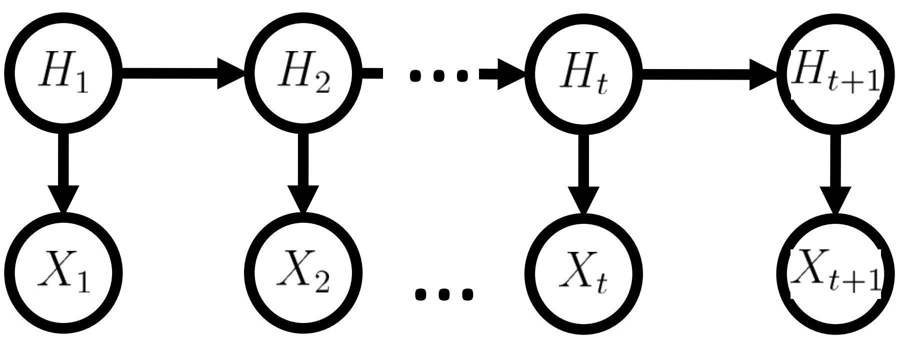
modified from: http://www.cs.cmu.edu/~beb/files/HSE_HMMs.pdf

Hidden Markov Models (HMMs)

Video sequence

Music





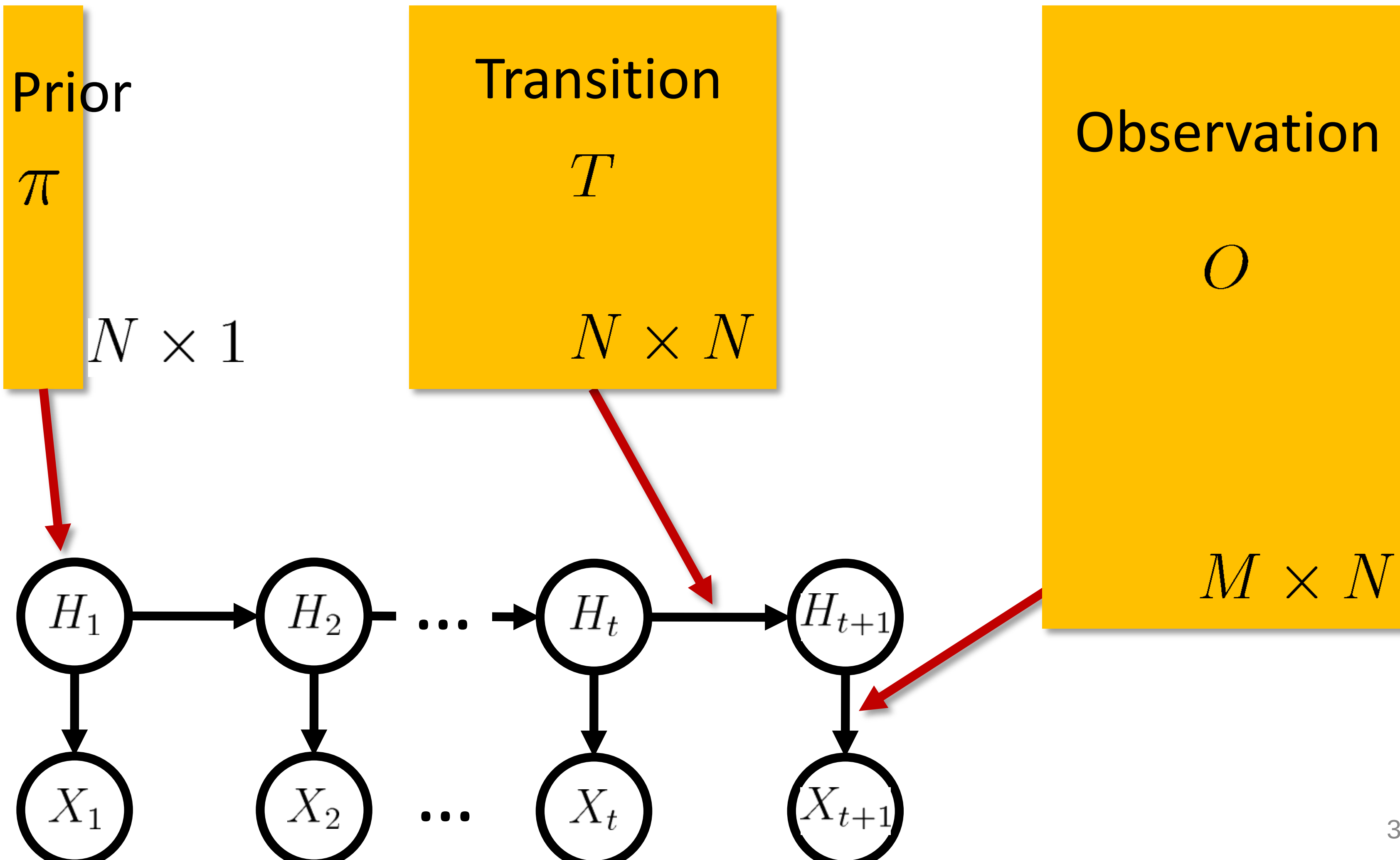
High-dimensional features

Hidden variables

Unsupervised learning

Notation

$\pi_i = \mathbb{P}(H_1 = i)$ $\quad T_{i,j} = \mathbb{P}(H_{t+1} = i | H_t = j)$ $\quad O_{i,j} = \mathbb{P}(X_t = i | H_t = j)$



Preliminaries: Learning Discrete HMMs

- EM algorithm [Dempster et al. 77]:
  - Maximum likelihood solution (Baum-Welch)
  - ✗ Local maxima
  - ✗ Curse of dimensionality
- **Spectral algorithm for HMM** [Hsu et al. 09]:
  - ✓ No local optima
  - ✓ Consistent

A Spectral Algorithm for Learning HMM

- Background
  - Learning linear system [van Overschee and de Moor 1996]

$$\mathbf{x}^{(n+1)} = A\mathbf{x}^{(n)} + B\mathbf{u}^{(n)} + \boldsymbol{\varepsilon}_x^{(n)}$$

$$\mathbf{y}^{(n)} = C\mathbf{x}^{(n)} + D\mathbf{u}^{(n)} + \boldsymbol{\varepsilon}_y^{(n)}$$

    - Observation: Y, U
    - **Subspace identification**
      - estimate A,B,C,D using SVD (up to a linear transformation)
    - Deterministic model (except noise)
- HMM is also a linear system (probabilistic)

$$\mathbb{E}[\vec{h}_{t+1} \mid \vec{h}_t] = T\vec{h}_t \quad \text{and} \quad \mathbb{E}[\vec{x}_t \mid \vec{h}_t] = Oh_t$$

[Hsu, 09] Use subspace identification to learn HMM

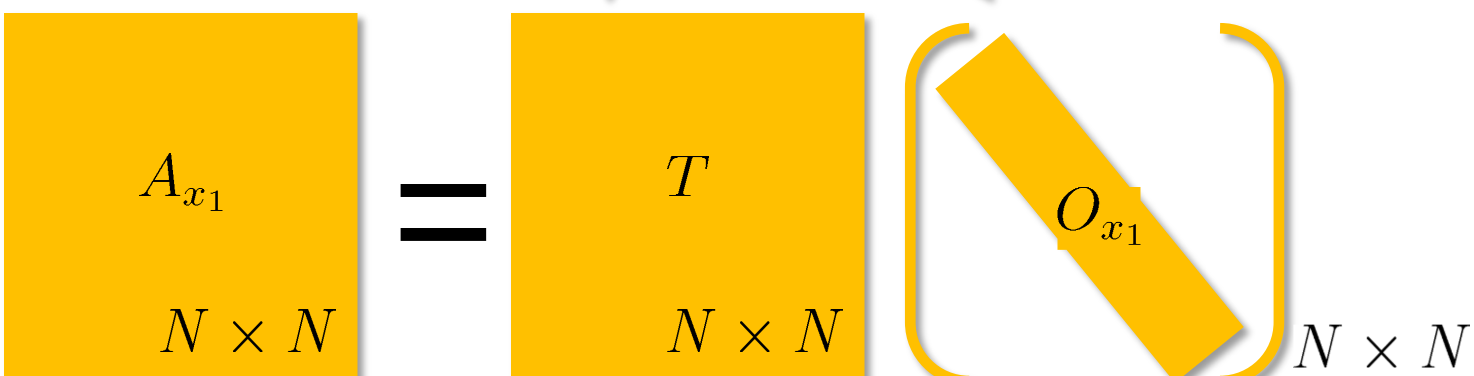
Predictive Distributions of HMMs

- Input $x_{t:1} := (x_t, \ldots, x_1)$ output $\mathbb{P}(X_{t+1} | x_{t:1})$
- Variable elimination:

$$\mathbb{P}(X_{t+1} = i | x_{t:1}) \propto \sum_{H_t} \mathbb{P}(X_{t+1} = i | H_t) \sum_{H_{t-1}} \mathbb{P}(H_t | H_{t-1}) \ldots$$

$$\ldots \sum_{H_1} \mathbb{P}(H_2 | H_1) \mathbb{P}(X_1 = x_1 | H_1) \mathbb{P}(H_1)$$

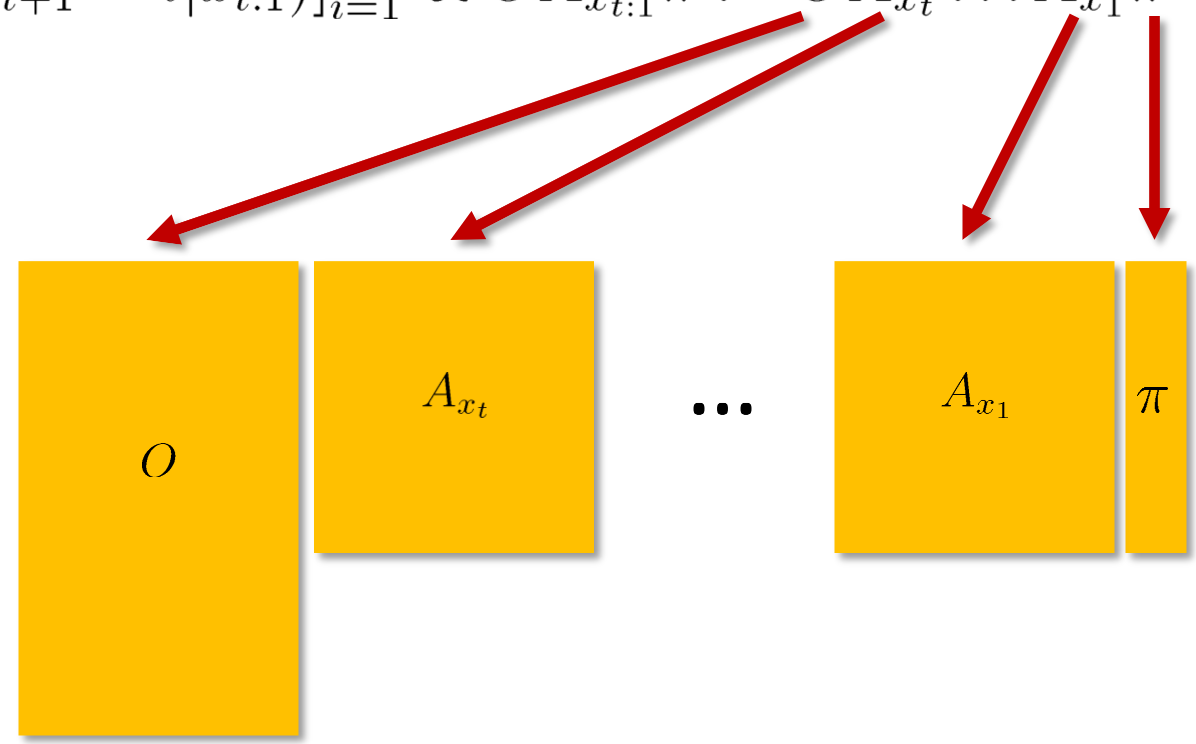
Observable Operator
[Jaeger 00]



Predictive Distributions of HMMs

- Input $x_{t:1} := (x_t, \ldots, x_1)$ output $\mathbb{P}(X_{t+1} | x_{t:1})$
- Variable elimination (matrix representation):

$$[\mathbb{P}(X_{t+1} = i | x_{t:1})]_{i=1}^{M} \propto O A_{x_{t:1}} \pi := O A_{x_t} \ldots A_{x_1} \pi$$



Observable representation of HMM

- Key observation: need *not* recover A_{x_t} :

$$
\begin{aligned}
[\mathbb{P}(X_{t+1} = i | x_{t:1})]_{i=1}^{M} &\propto O A_{x_t} \ldots A_{x_1} \pi \\
&= \underbrace{(O S^{-1})}_{b_\infty} \underbrace{(S A_{x_t} S^{-1})}_{B_{x_t}} \ldots (S A_{x_1} S^{-1}) \underbrace{(S \pi)}_{b_1}
\end{aligned}
$$

Only need to estimate *O*, *Ax and* π up to invertible transformation *S*

- $S = U^\top O$ where U are singular vectors of joint probability of sequence pairs [Hsu et al. 09]

Observable representation for HMM

$$C_{2,1} = \mathbb{P}(X_{t+1} = i, X_t = j)$$

$M \times M$

pairs

$$C_{3,x,1} = \mathbb{P}(X_{t+2} = i, x, X_t = j)$$

$M \times M$

triplets

$$u = \mathbb{P}(X_t = i)$$

$M \times 1$

singletons

sequence

Thin SVD of $C_{2,1}$, get principal left singular vectors U

Observable representation for HMMs

$$[\mathbb{P}(X_{t+1} = i | X_{t:1})]_{i=1}^{M} \propto (OS^{-1})(SA_{x_t}S^{-1}) \ldots (SA_{x_1}S^{-1})(S\pi)$$
$$= b_\infty B_{x_t} \ldots B_{x_1} b_1$$

$C_{2,1}(U^\top C_{2,1})^\dagger$

$(U^\top C_{3,x,1})(U^\top C_{2,1})^\dagger$

$U^\top u$

$C_{2,1} =$
$\mathbb{P}(X_{t+1} = i, X_t = j)$

$M \times M$

$C_{3,x,1} =$
$\mathbb{P}(X_{t+2} = i, x, X_t = j)$

$M \times M$

$u =$
$\mathbb{P}(X_t = i)$

$M \times 1$

pairs

triplets

singletons

Works only for discrete case

Key Objects in Graphical Models

- Marginal distributions $\mathbb{P}(Y)$
- Joint distributions $\mathbb{P}(Y, X)$
- Conditional distributions $\mathbb{P}(Y|X)$
- Sum rule $\mathbb{P}(Y) = \int_X \mathbb{P}(Y|X)\mathbb{P}(X)$
- Product rule $\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$

Use kernel representation for distributions, do probabilistic inference in feature space

Embedding distributions

- Summary statistics for distributions $\mathbb{P}(Y)$:

$\mathbb{E}_{Y\sim\mathbb{P}}[Y]$ Mean

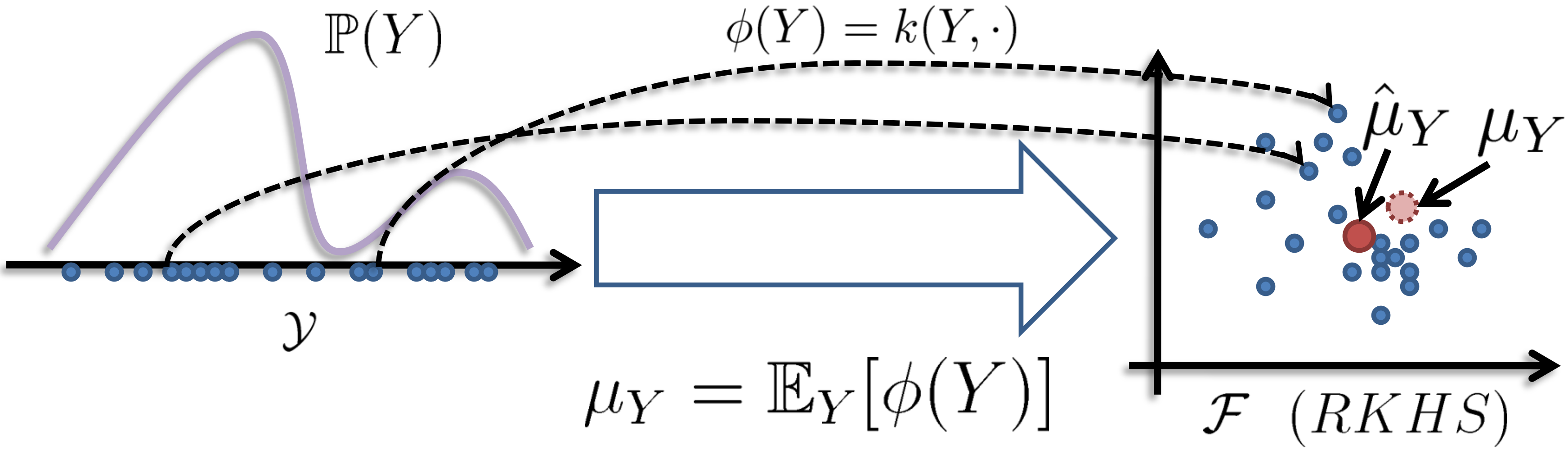
$\mathbb{E}_{Y\sim\mathbb{P}}[YY^\top]$ Covariance

$\mathbb{E}_{Y\sim\mathbb{P}}[\delta_{y_0}(Y)]$ Probability $P(y_0)$

$\mathbb{E}_{Y\sim\mathbb{P}}[\phi(Y)]$ expected features

- Pick a kernel $k(y,y') = \langle \phi(y), \phi(y') \rangle$, and generate a different summary statistic

Embedding distributions



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

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

- One-to-one mapping from $\mathbb{P}(Y)$ to μ_Y for certain kernels (RBF kernel)
- Sample average converges to true mean at $O_p(m^{-\frac{1}{2}})$

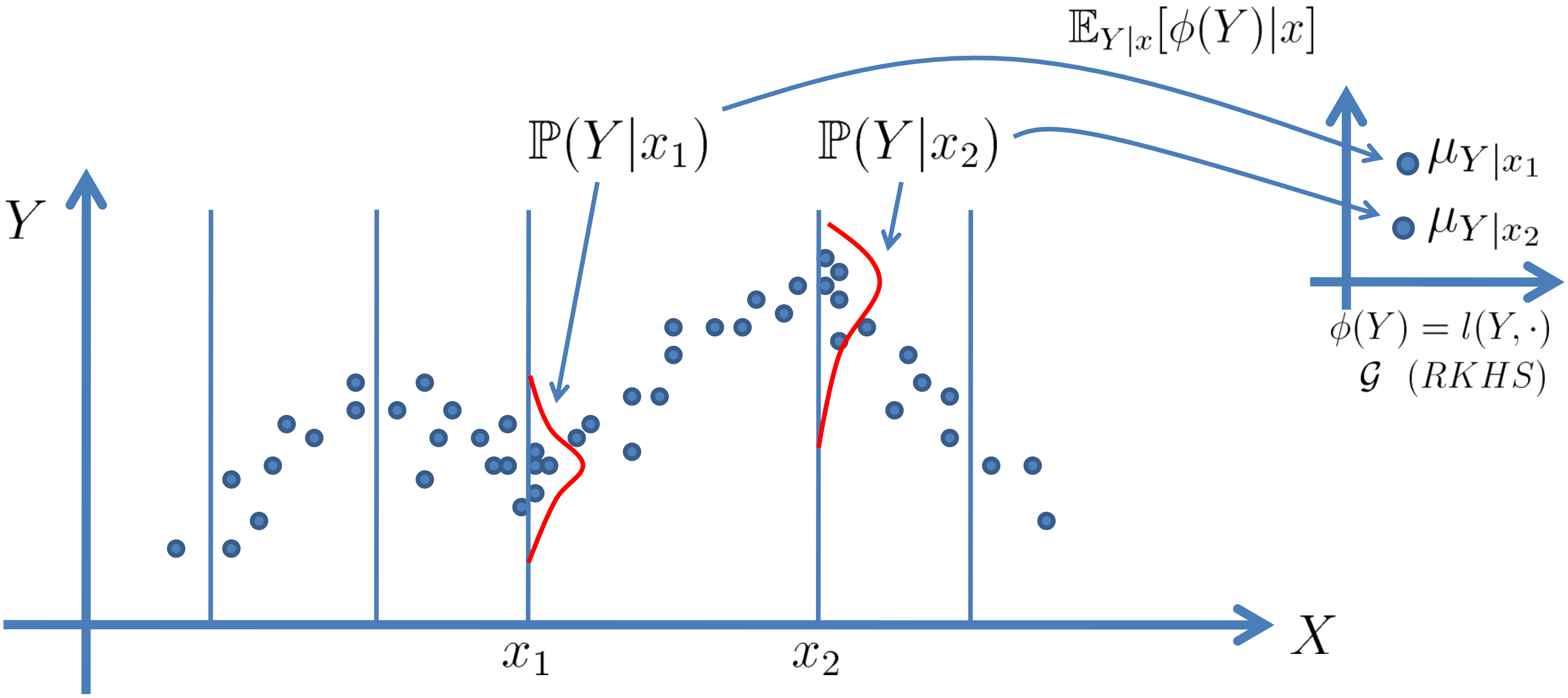
Embedding joint distributions

- Embedding joint distributions $\mathbb{P}(Y, X)$ using outer-product feature map $\phi(Y)\varphi(X)^\top$

$$\mu_{YX} = \mathbb{E}_{YX}[\phi(Y)\varphi(X)^\top]$$
$$\hat{\mu}_{YX} = \frac{1}{m}\sum_{i=1}^{m} \phi(y_i)\varphi(x_i)^\top$$

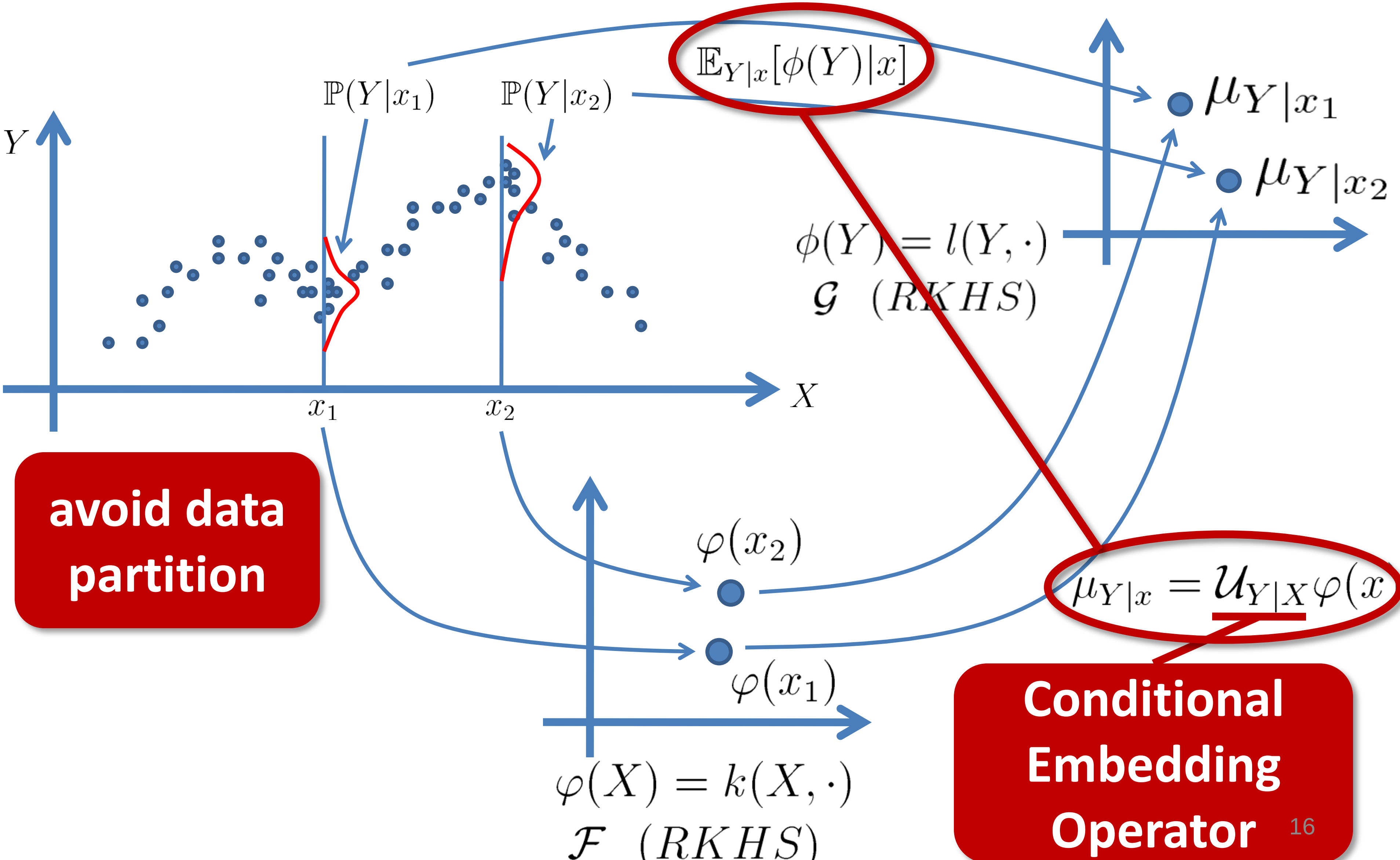
- μ_{YX} is also the covariance operator $\mathcal{C}_{YX}$
- Recover discrete probability with delta kernel
- Empirical estimate converges at $O_p(m^{-\frac{1}{2}})$

Embedding Conditionals



- For each value *X=x* conditioned on, return the summary statistic for $\mathbb{P}(Y|X = x)$
- Some *X=x* are *never* observed

Embedding conditionals



Conditional Embedding Operator

- Estimation via covariance operators [Song et al. 09]

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$
$$\hat{\mathcal{U}}_{Y|X} = \Phi(K + \lambda I)^{-1}\Upsilon$$
$$\Phi := (\phi(y_1), \ldots, \phi(y_m)), \quad L = \Phi^\top \Phi$$
$$\Upsilon := (\varphi(x_1), \ldots, \varphi(x_2)), \quad K = \Upsilon^\top \Upsilon$$

- Gaussian case: covariance matrix instead
- Discrete case: joint over marginal
- Empirical estimate converges at $O_p((\lambda m)^{-\frac{1}{2}} + \lambda^{\frac{1}{2}})$

Sum and Product Rules

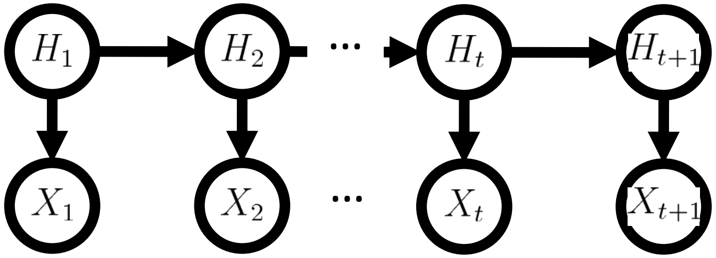
	Probabilistic Relation	Hilbert Space Relation
Sum Rule	$\mathbb{P}(Y) = \int_X \mathbb{P}(Y\|X)\mathbb{P}(X)$	$\mu_Y = \mathcal{U}_{Y\|X}\mu_X$
Product Rule	$\mathbb{P}(X,Y) = \mathbb{P}(Y\|X)\mathbb{P}(X)$	$\mu_{XY} = \mathcal{U}_{Y\|X}\mathcal{C}_{XX}$

$$\mu_Y = \mathbb{E}_Y[\phi(Y)] = \mathbb{E}_X\mathbb{E}_{Y|X}[\phi(Y)] = \mathbb{E}_X[\mathcal{U}_{Y|X}\varphi(X)] = \mathcal{U}_{Y|X}\mu_X$$

Total Expectation — Conditional Embedding — Linearity

Hilbert Space HMMs

$$[\mathbb{P}(X_{t+1} = i | x_{t:1})]_{i=1}^{M} \propto OA_{x_{t:1}}\pi := OA_{x_t} \ldots A_{x_1}\pi$$



$$\mu_{X_{t+1}|x_{t:1}} \propto \mathbb{E}_{X_{t+1}|x_{t:1}}[\varphi(X_{t+1})]$$

$$\propto \mathbb{E}_{H_{t+1}|x_{t:1}} \mathbb{E}_{X_{t+1}|H_{t+1}}[\varphi(X_{t+1})]$$

(Total expectation)

$$\propto \mathbb{E}_{H_{t+1}|x_{t:1}}[\mathcal{U}_{X_{t+1}|H_{t+1}} \phi(H_{t+1})]$$

(Conditional embedding operator)

$$\propto \mathcal{U}_{X_{t+1}|H_{t+1}} \mathbb{E}_{H_{t+1}|x_{t:1}}[\phi(H_{t+1})]$$

(Linearity)

$$\propto \mathcal{U}_{X_{t+1}|H_{t+1}} \mathcal{A}_{x_t} \mathbb{E}_{H_t|x_{t-1:1}}[\phi(H_t)]$$

$(\mathcal{A}_{x_t} := \mathbb{P}(X_t = x_t | h_t) \mathbb{E}_{H_{t+1}|h_t}[\phi(H_{t+1})])$

$$\propto \mathcal{U}_{X_{t+1}|H_{t+1}} \mathcal{A}_{x_t} \ldots \mathcal{A}_{x_1} \mu_1$$

(recursion and $\mu_1 := \mathbb{E}_{X_1}[\varphi(X_1)]$)

Hilbert space HMMs

$$\mu_{X_{t+1}|x_{t:1}} \propto (\mathcal{U}_{X_{t+1}|H_{t+1}}\mathcal{S}^{-1})(\mathcal{S}\mathcal{A}_{x_t}\mathcal{S}^{-1})\ldots(\mathcal{S}\mathcal{A}_{x_1}\mathcal{S}^{-1})(\mathcal{S}\mu_1)$$

$$\mathcal{C}_{2,1}(U^\top \mathcal{C}_{2,1})^\dagger \qquad (U^\top \mathcal{U}_{3,1|2}\varphi(x))(U^\top \mathcal{C}_{2,1})^\dagger \qquad U^\top \mu_1$$

$\mathcal{C}_{2,1} =$
$\mathbb{E}[\varphi(X_{t+1})\varphi(X_t)^\top]$

$\infty \times \infty$

pairs

$\mathcal{U}_{3,1|x}\varphi(x) =$
$\mathbb{E}[\varphi(X_{t+2})\varphi(X_t)^\top | x]$
for each middle x

$\infty \times \infty$

triplets

$\mu_1 =$
$\mathbb{E}[\varphi(X_t)]$

$\infty \times 1$

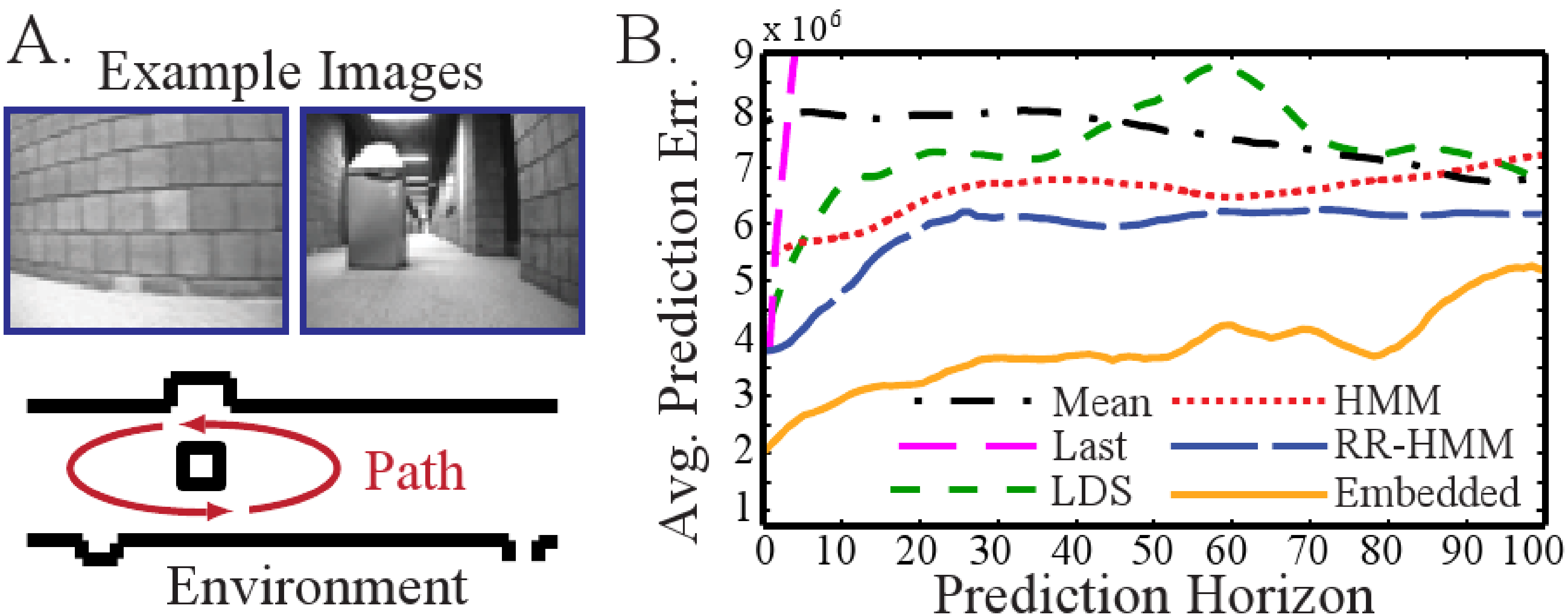
singletons

Experiment

- Video sequence prediction
- Slot car sensor measurement prediction
- Speech classification
- Compare with
  - Discrete HMMs learned by EM [Dempster et al. 77],
  - Reduced rank HMM [Sajid et al. 10]
  - Spectral HMM [Hsu et al. 09]
  - Linear dynamical system approach (LDS) [Sajid et al. 08]

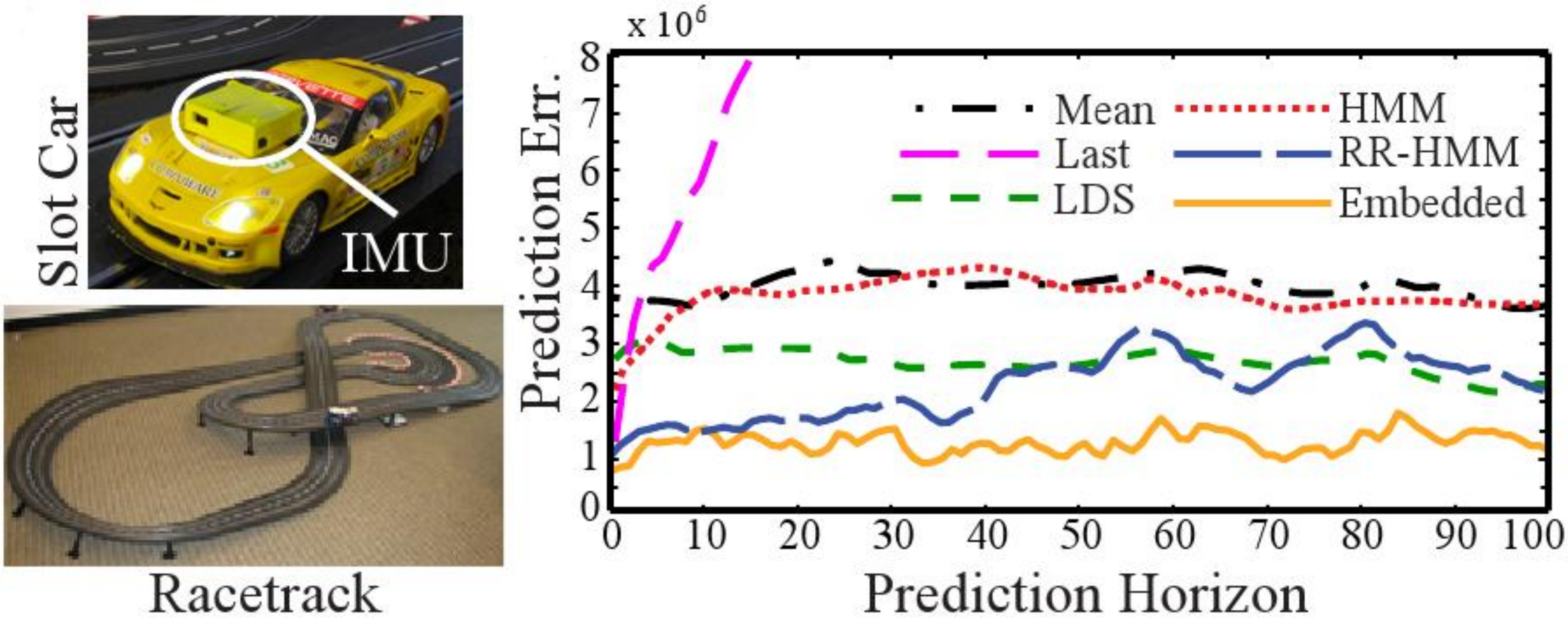
Predicting Video Sequences

- Sequence of grey scale images as inputs
- Latent space dimension 50



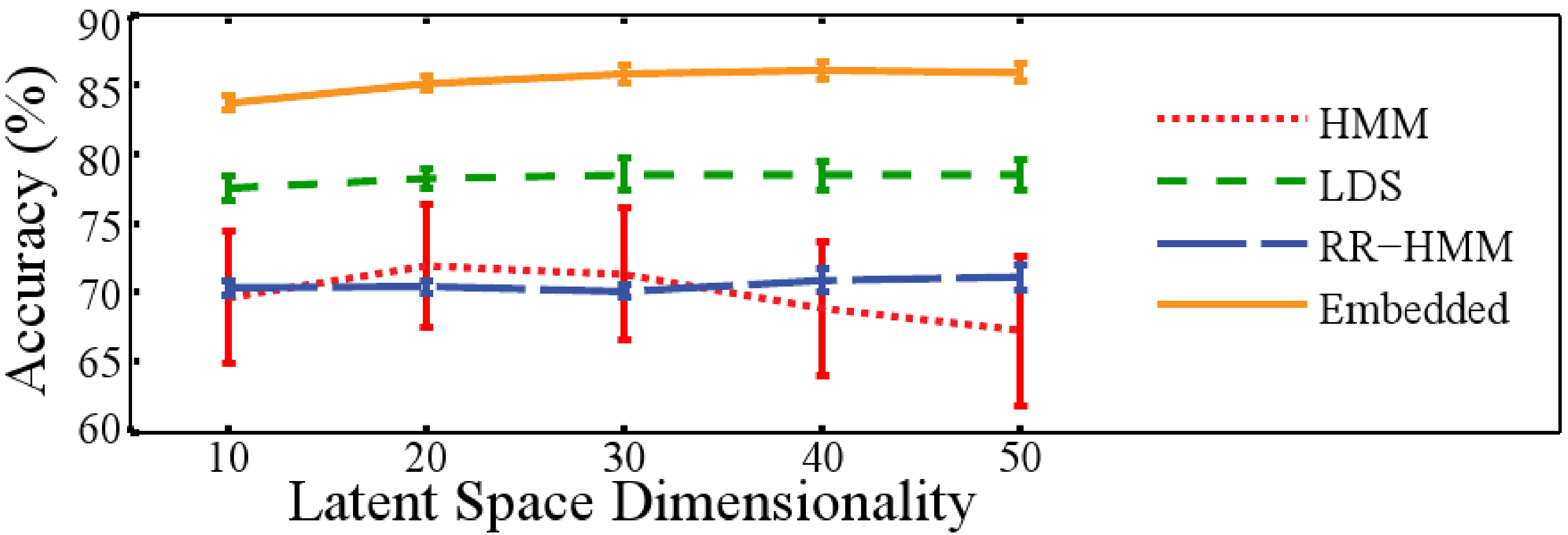
Predicting Sensor Time-series

- Inertial unit: 3D acceleration and orientation
- Latent space dimension 20



Audio Event Classification

- Mel-Frequency Cepstral Coefficients features
- Varying latent space dimension



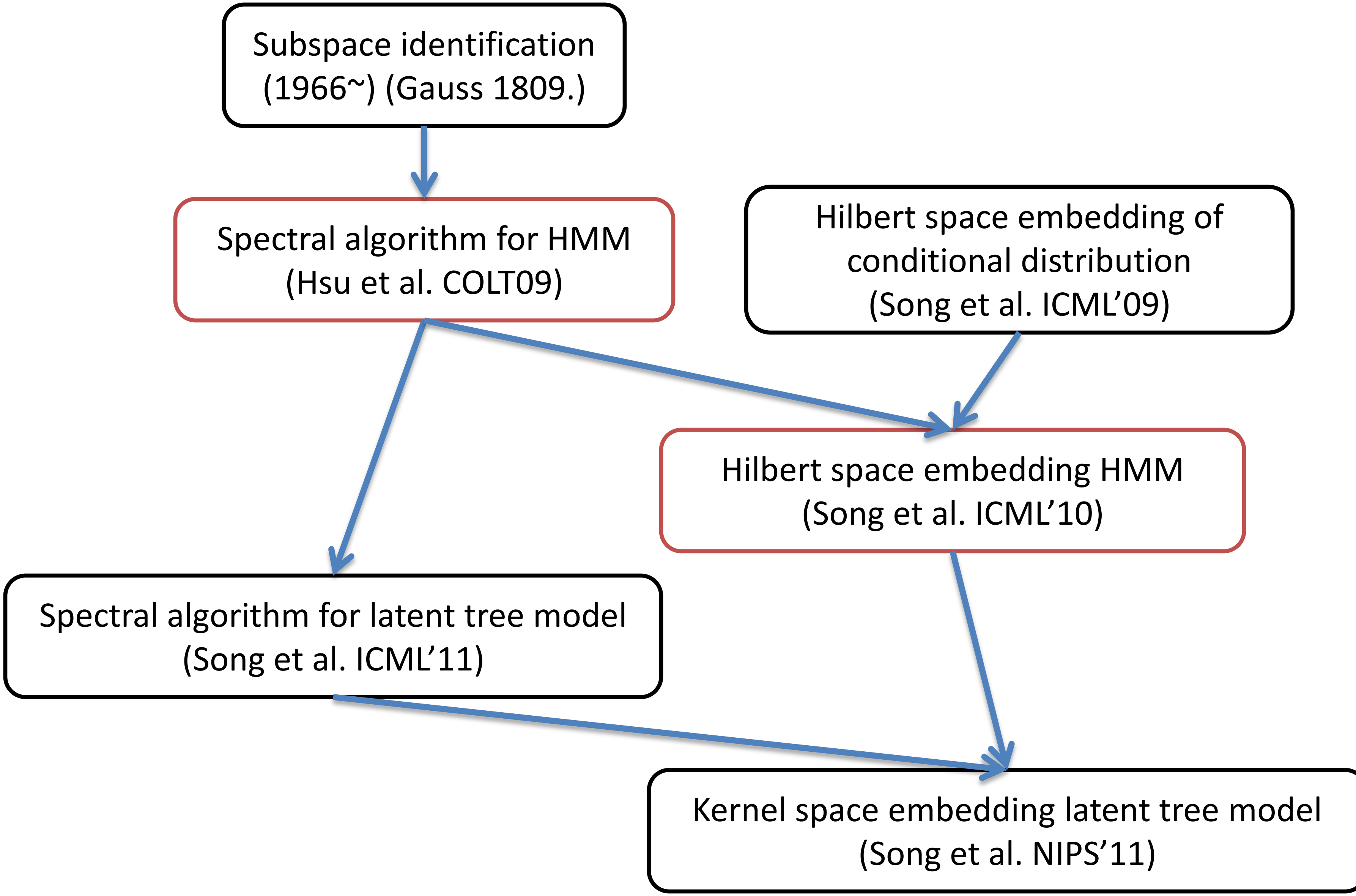
Summary

- Represent distributions in feature spaces, reason using Hilbert space sum and product rules

- Extends HMMs nonparametrically to domains with kernels

- Kernelize belief propagation, CRF and general graphical models with hidden variables?

Genealogy


Subspace identification
(1966~) (Gauss 1809.)
Spectral algorithm for HMM
(Hsu et al. COLT09)
Hilbert space embedding of
conditional distribution
(Song et al. ICML'09)
Hilbert space embedding HMM
(Song et al. ICML'10)
Spectral algorithm for latent tree model
(Song et al. ICML'11)
Kernel space embedding latent tree model
(Song et al. NIPS'11)

Thanks!

Big Picture Question

Graphical Models	Kernel Methods
✓ Dependent variables ✓ Hidden variables	✓ High dimensional ✓ Nonlinear ✓ Multimodal
✗ High dimensional ✗ Nonlinear ✗ Multimodal	✗ Dependent variables ✗ Hidden variables

Combine the best of graphical models and kernel methods.