

Hilbert Space Embeddings of Hidden Markov Models

Le Song

Carnegie Mellon University

Joint work with Byron Boots, Sajid Siddiqi, Geoff Gordon and Alex Smola



Big Picture Question

Graphical Models	Kernel Methods
<ul style="list-style-type: none">✓ Dependent variables✓ Hidden variables	<ul style="list-style-type: none">✓ High dimensional✓ Nonlinear✓ Multimodal
<ul style="list-style-type: none">✗ High dimensional✗ Nonlinear✗ Multimodal	<ul style="list-style-type: none">✗ Dependent variables✗ Hidden variables

Graphical models and kernel methods complement each other

Big Picture Question

Kernel Graphical Models

- ✓ High dimensional
- ✓ Nonlinear
- ✓ Multimodal
- ✓ Dependent variables
- ✓ Hidden variables

Combine the best of graphical models and kernel methods?

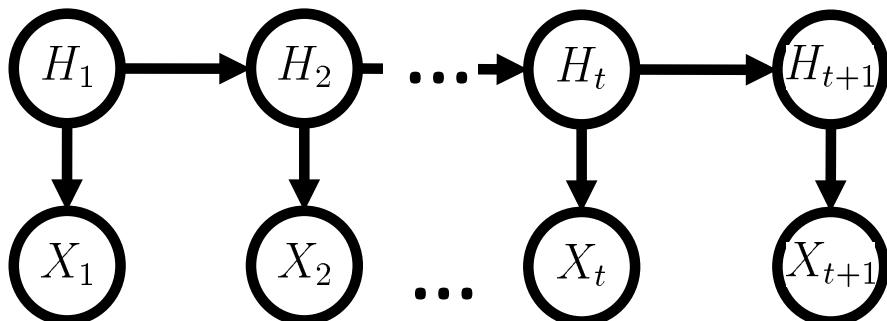
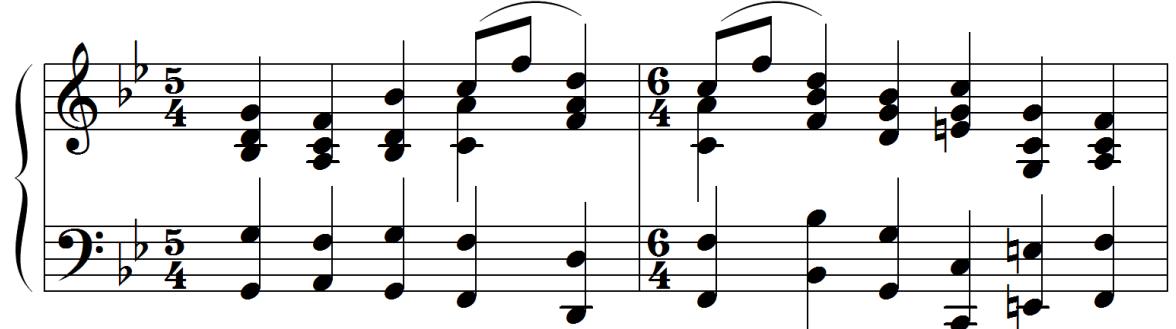
Kenji Fukumizu, Carlos Guestrin, Arthur Gretton,
Jonathan Huang, Eric Xing

Hidden Markov Models (HMMs)



Video sequence

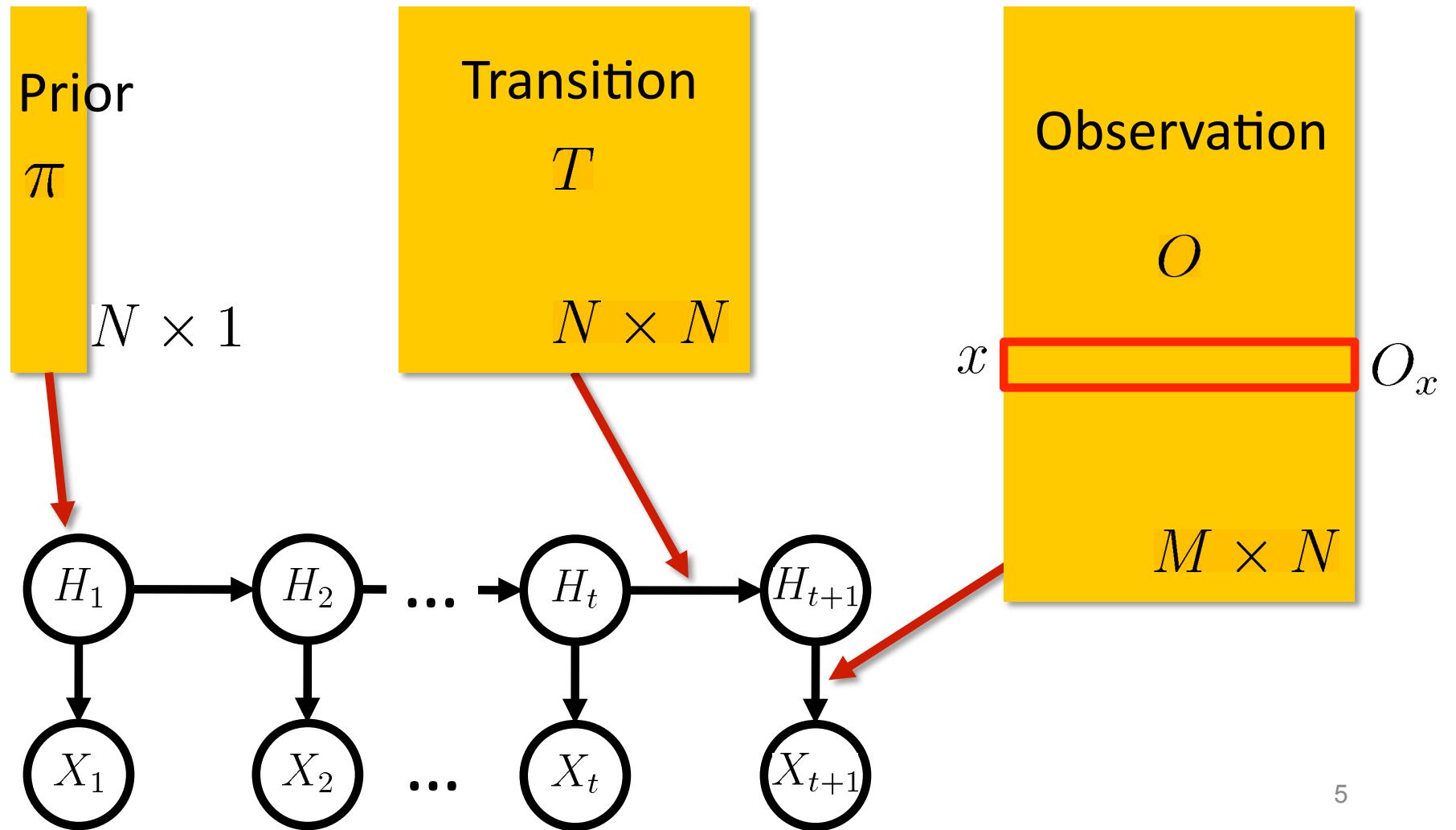
Music



High-dimensional features
Hidden variables
Unsupervised learning

Notation

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



Previous Work on HMMs

- Expectation maximization [Dempster et al. 77]:

Interested in prediction, not hidden states

- Singular value decomposition (SVD) for surrogate hidden states

✓ No local optima

✓ Consistent

Spectral HMMs [Hsu et al. 09, Siddiqi et al. 10], Subspace Identification [Van Overschee and De Moor 96]

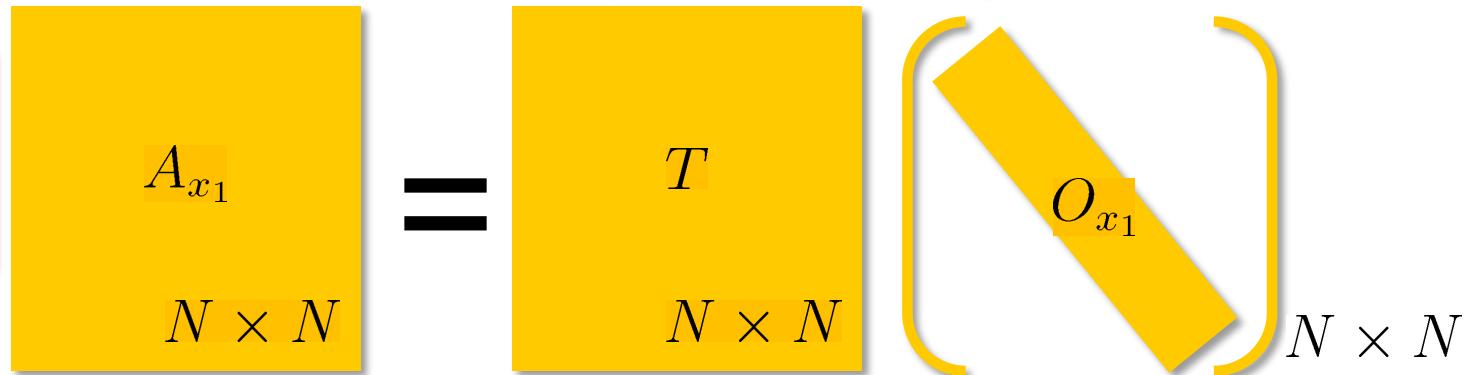
Predictive Distributions of HMMs

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

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

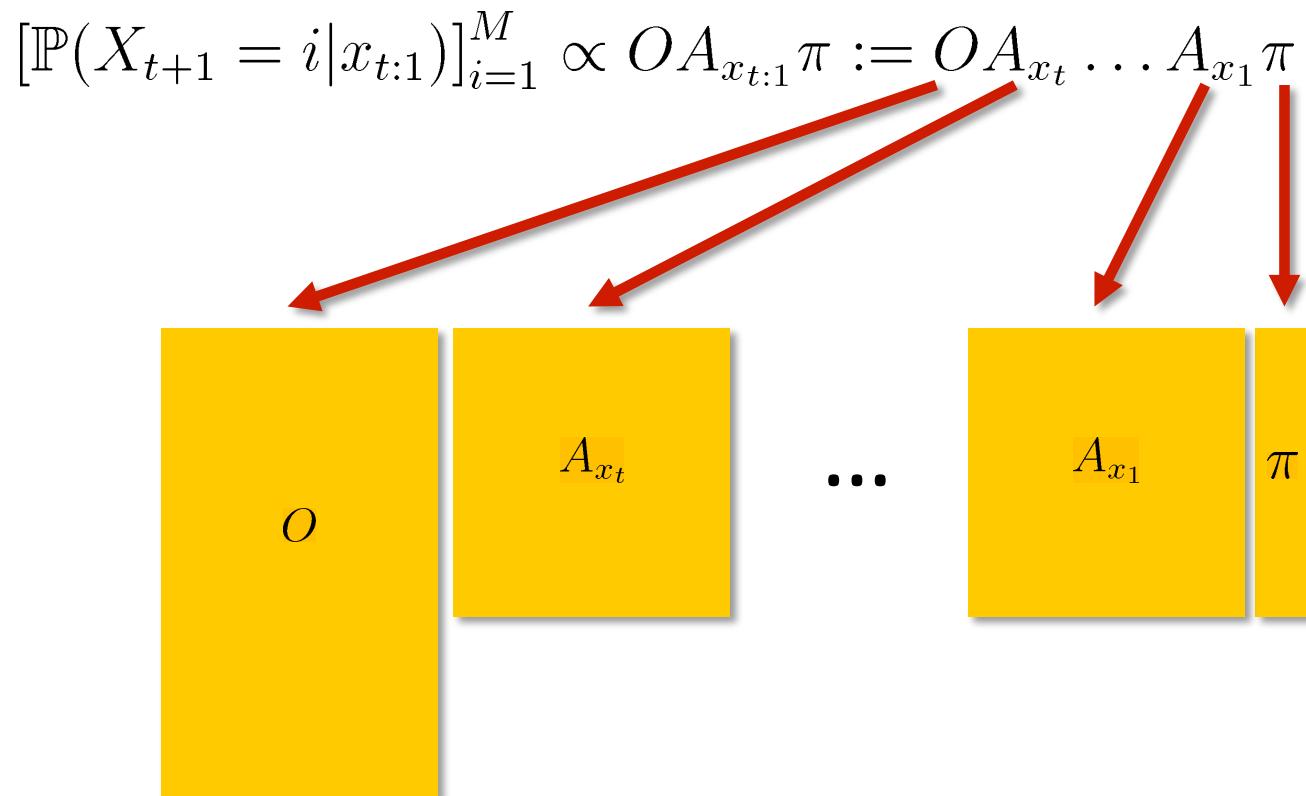
$$\dots \sum_{H_1} \mathbb{P}(H_2|H_1) \mathbb{P}(X_1 = x_1|H_1) \mathbb{P}(H_1)$$

Observable
Operator
[Jaeger 00]



Predictive Distributions of HMMs

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



Observable representation of HMM

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

$$\begin{aligned} [\mathbb{P}(X_{t+1} = i | x_{t:1})]_{i=1}^M &\propto O A_{x_t} \dots A_{x_1} \pi \\ &= (OS^{-1})(SA_{x_t}S^{-1}) \dots (SA_{x_1}S^{-1})(S\pi) \end{aligned}$$


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

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

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

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

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

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

Estimable from observation sequence!

A thin SVD of $C_{2,1}$ to get U

Works only for discrete case

1

ns

Key Objects in Graphical Models

- Marginal distributions $\mathbb{P}(Y)$
- Joint distributions $\mathbb{P}(Y, X)$
 - Condition distributions, do probabilistic inference in feature space
- Sum rule $\mathbb{P}(Y) = \int_X \mathbb{P}(Y|X)\mathbb{P}(X)$
- Product rule $\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$

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

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

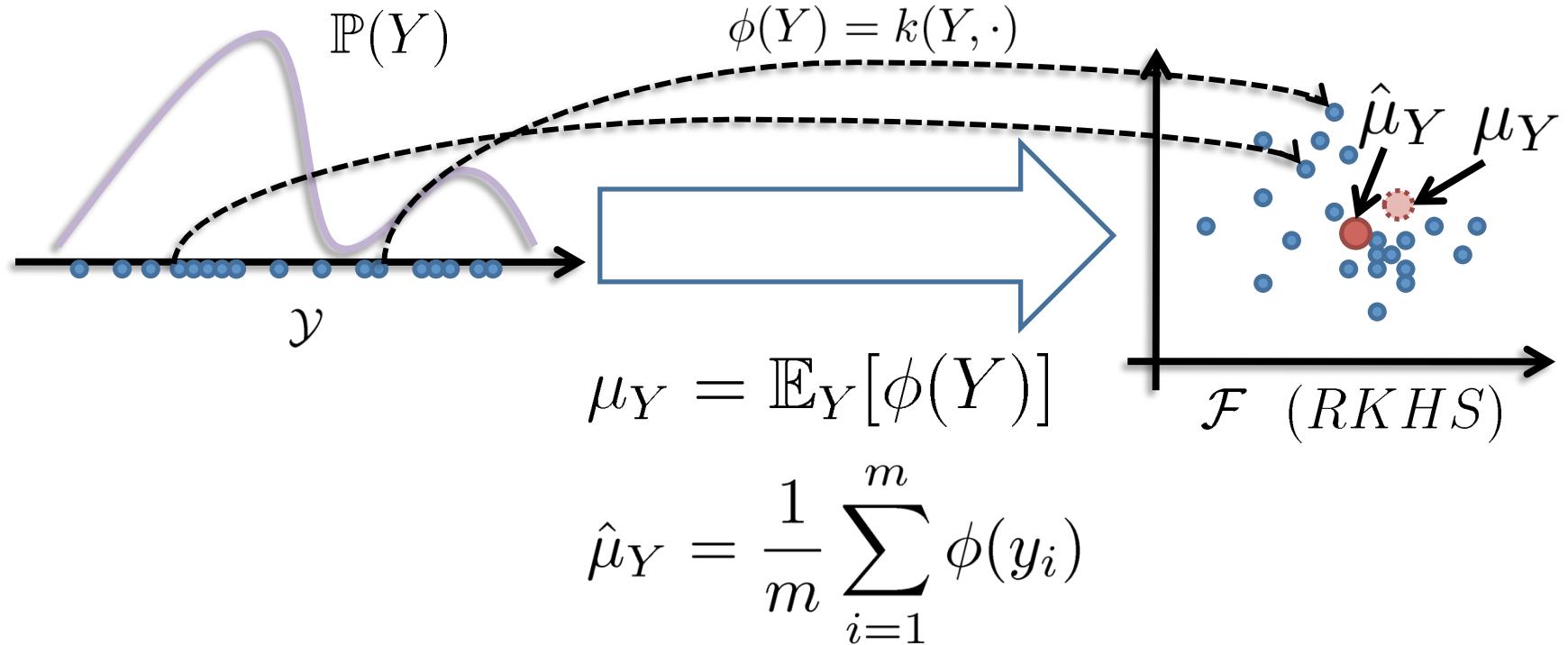
Probability $P(y_0)$

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

expected features

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

Embedding distributions



- 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

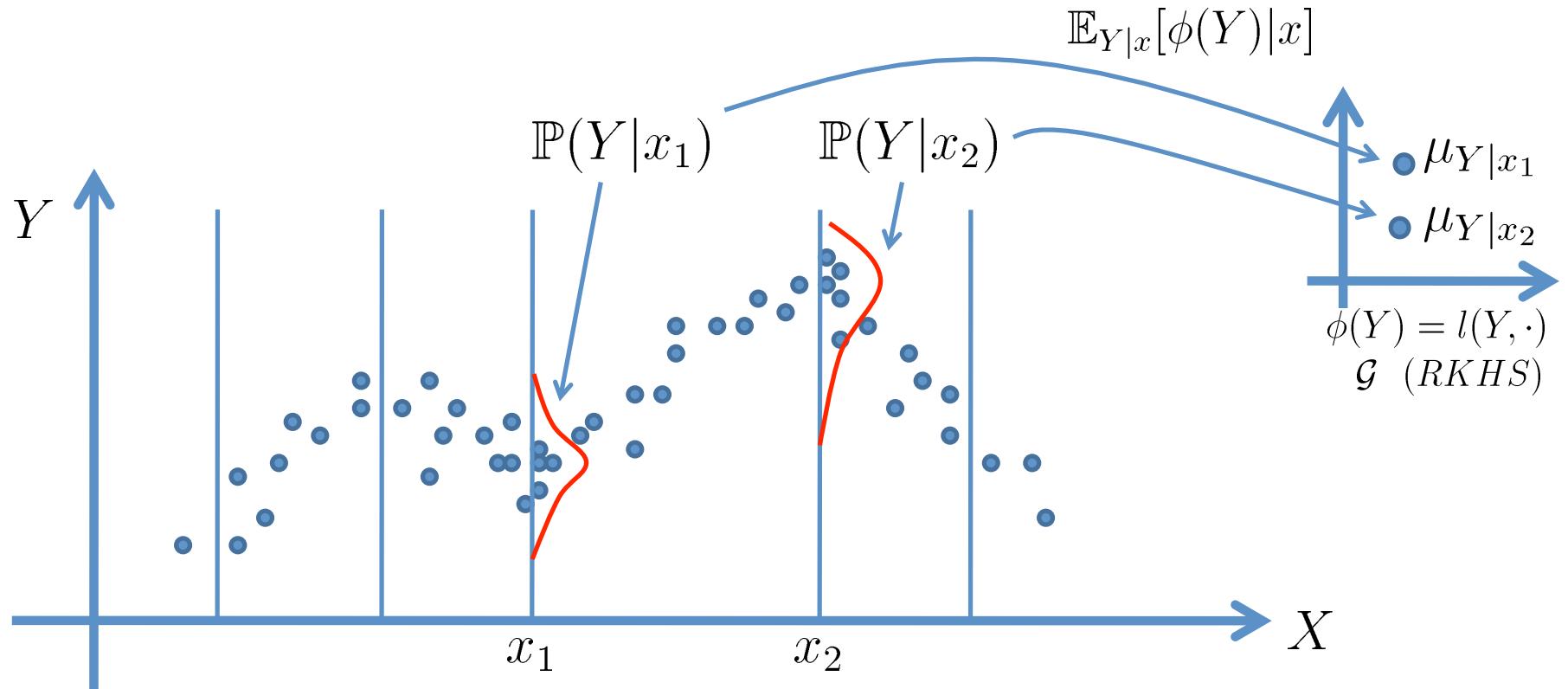
- Embedding joint distributions $\mathbb{P}(Y, X)$ using outer-product feature map $\phi(Y)\varphi(X)^\top$

$$\mu_{YX} = \mathbb{E}_{YX}[\phi(Y)\varphi(X)^\top]$$

$$\hat{\mu}_{YX} = \frac{1}{m} \sum_{i=1}^m \phi(y_i)\varphi(x_i)^\top$$

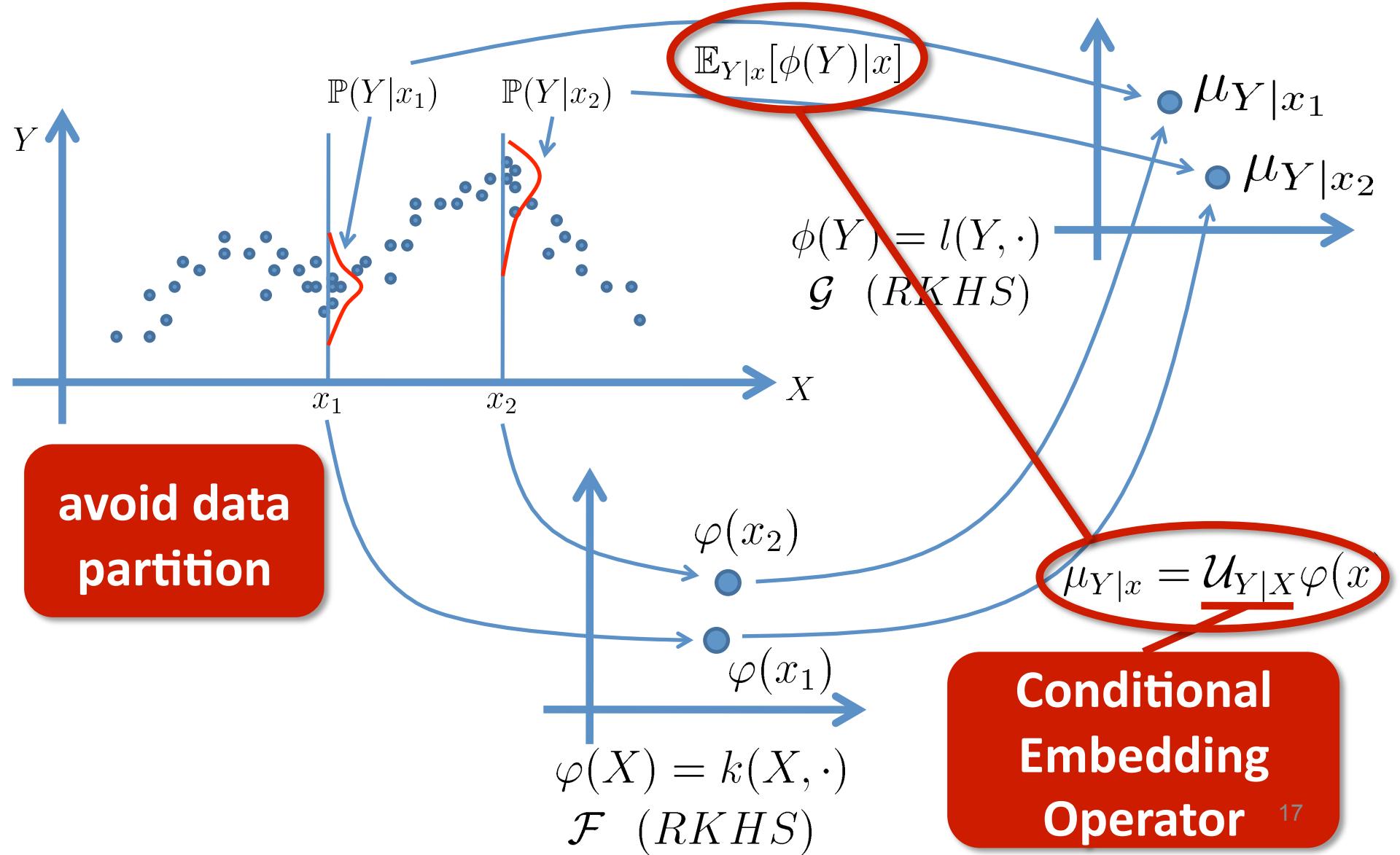
- μ_{YX} is also the covariance operator \mathcal{C}_{YX}
- Recover discrete probability with delta kernel
- Empirical estimate converges at $O_p(m^{-\frac{1}{2}})$

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]

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

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

$$\Phi := (\phi(y_1), \dots, \phi(y_m)), \quad L = \Phi^\top\Phi$$

$$\Upsilon := (\varphi(x_1), \dots, \varphi(x_2)), \quad K = \Upsilon^\top\Upsilon$$

- 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	$\mathbb{P}(X, Y) = \mathbb{P}(Y X)\mathbb{P}(X)$	$\mu_{XY} = \mathcal{U}_{Y X}\mathcal{C}_{XX}$

Hilbert Space HMMs

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

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

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

(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)

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

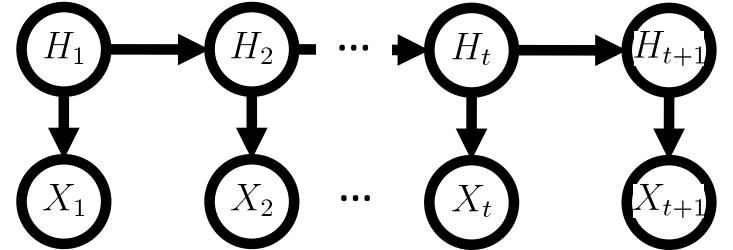
(Linearity)

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

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

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

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



Observable Representation for HMMs

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



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

$M \times M$

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

Hilbert space HMMs

$$\mu_{X_{t+1}|x_{t:1}} \propto (\mathcal{U}_{X_{t+1}|H_{t+1}} \mathcal{S}^{-1})(\mathcal{S}\mathcal{A}_{x_t} \mathcal{S}^{-1}) \dots (\mathcal{S}\mathcal{A}_{x_1} \mathcal{S}^{-1})(\mathcal{S}\mu_1)$$
$$\mathcal{C}_{2,1}(U^\top \mathcal{C}_{2,1})^\dagger \quad (U^\top \mathcal{U}_{3,1|2} \varphi(x))(U^\top \mathcal{C}_{2,1})^\dagger \quad U^\top \mu_1$$

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

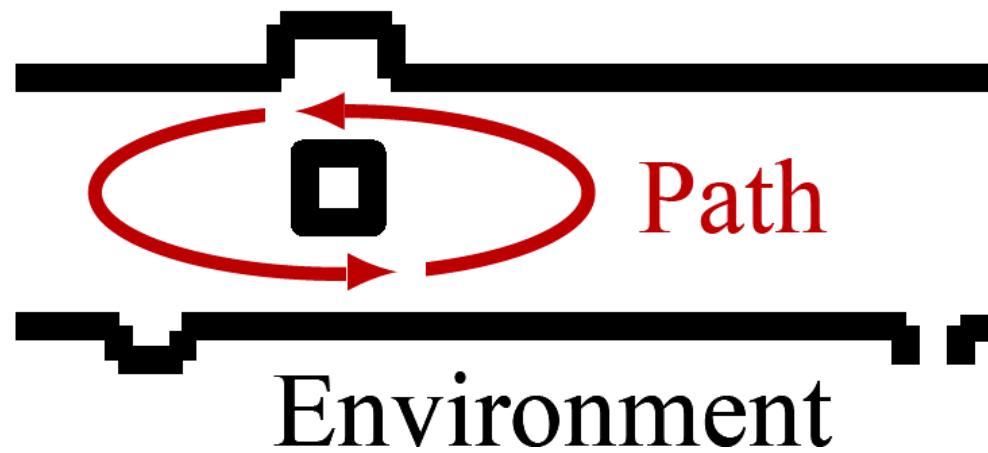
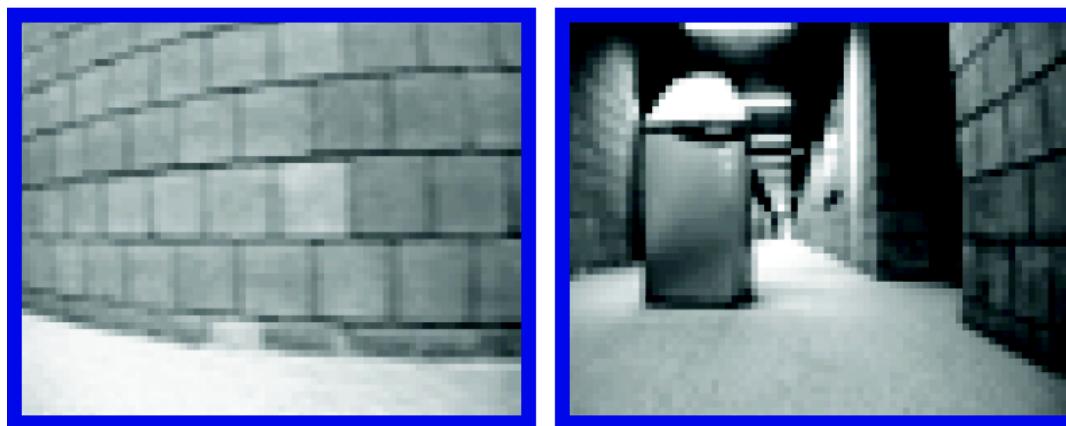
$$\mathcal{U}_{3,1|x} \varphi(x) = \mathbb{E}[\varphi(X_{t+2})\varphi(X_t)^\top | x]
for each middle x$$

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

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

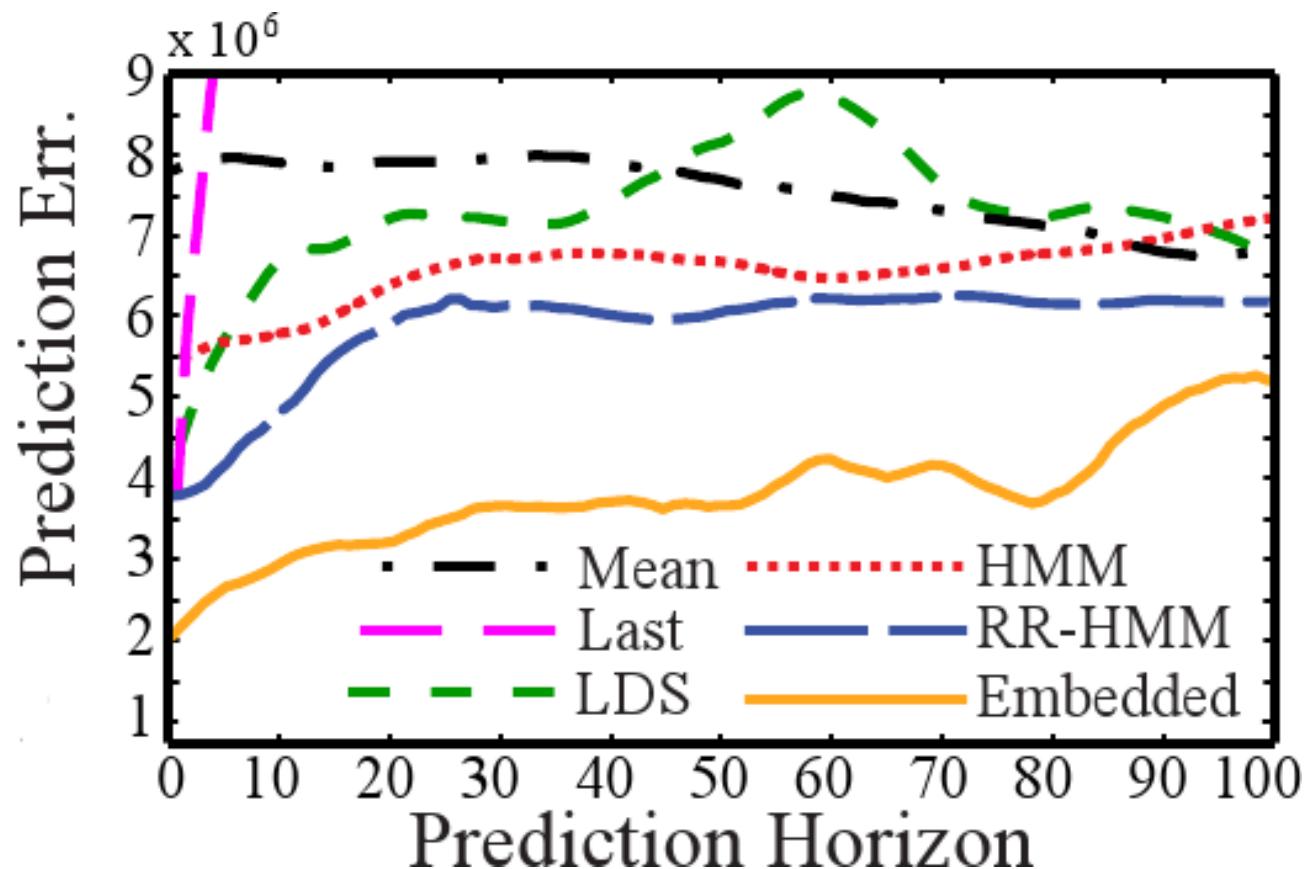
Predicting Video Sequences

Example Images



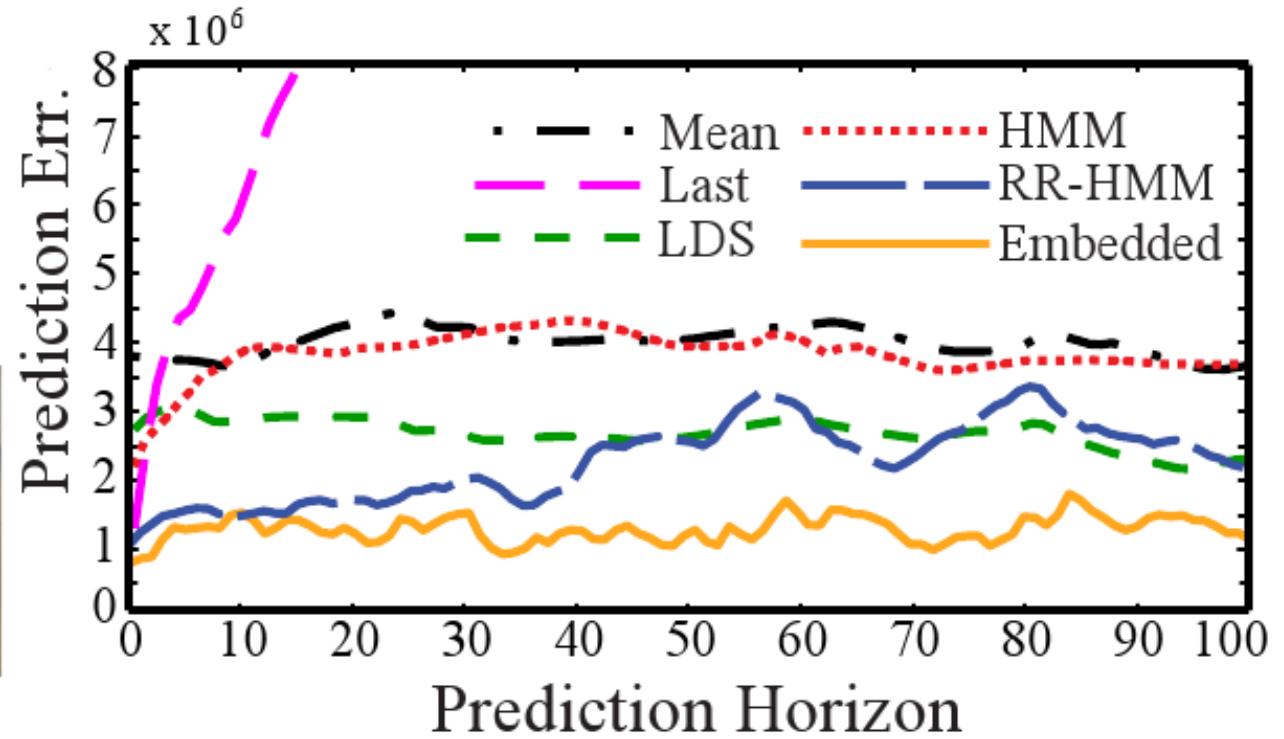
Predicting Video Sequences

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



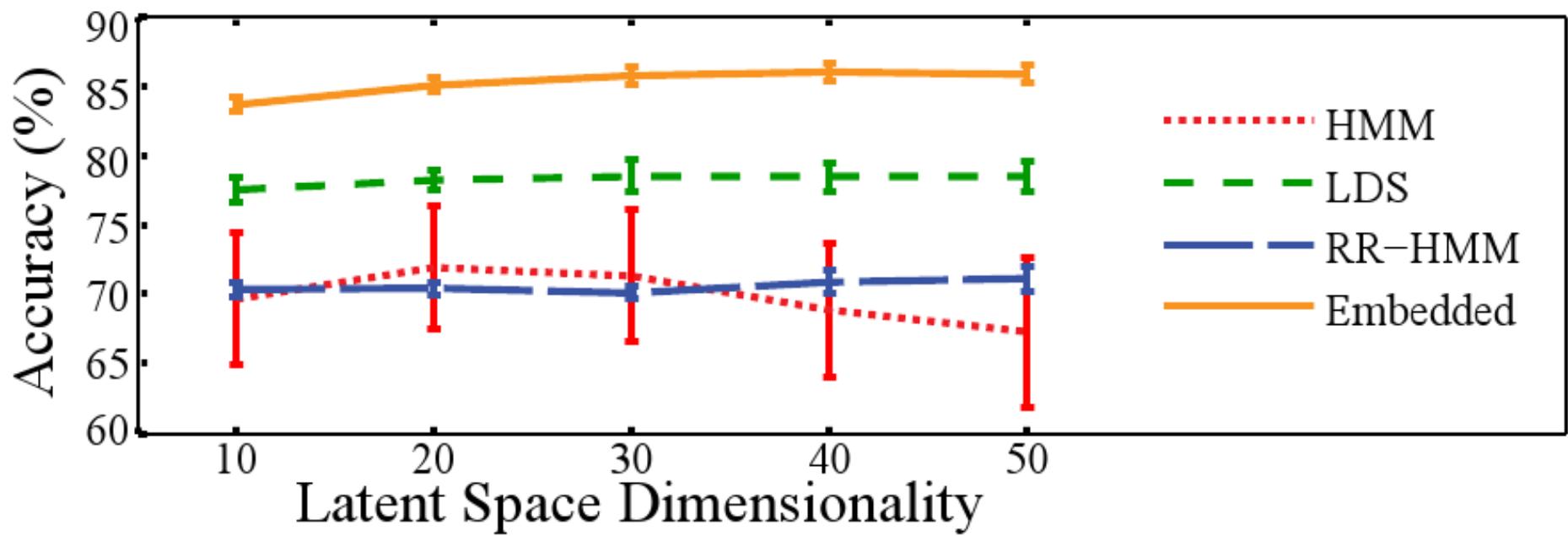
Predicting Sensor Time-series

- 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 and general graphical models with hidden variables?