Georgia Tech

Spectral Algorithms for Latent Variable Models
Part III: Latent Tree Models

Le Song

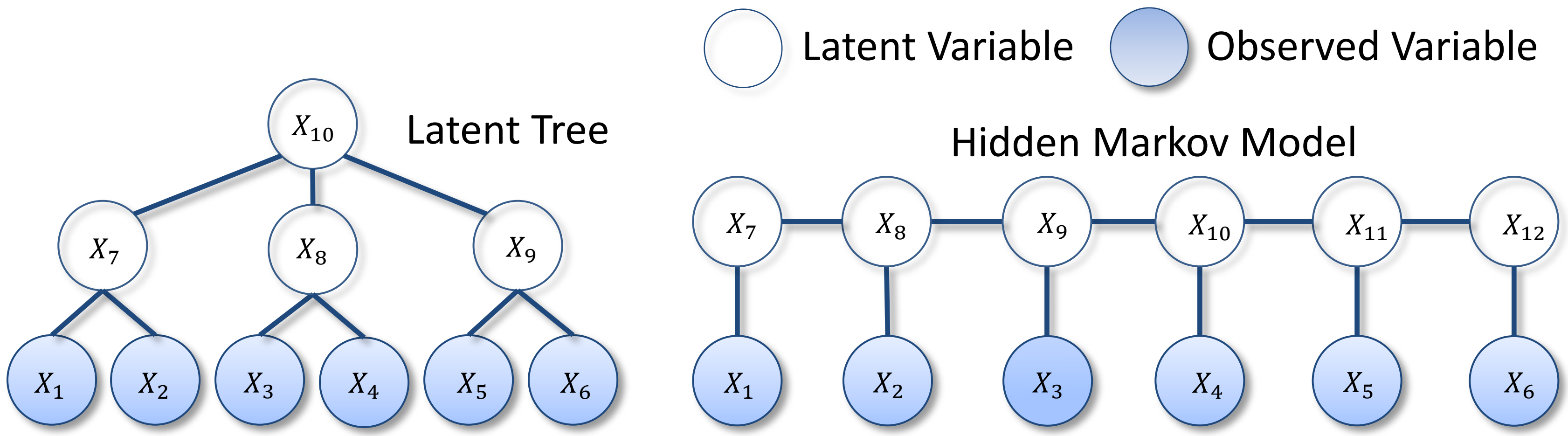
ICML 2012 Tutorial on Spectral Algorithms for Latent Variable Models, Edinburgh, UK

Joint work with Mariya Ishteva, Ankur Parikh, Eric Xing, Byron Boots , Geoff Gordon, Alex Smola and Kenji Fukumizu

Latent Tree Graphical Models

- Graphical model: nodes represent variables, edges represent conditional independence relation

- Latent tree graphical models: latent and observed variables are arranged in a tree structure



- Many real world applications, eg., time-series prediction, topic modeling

Scope of This Tutorial

- ✔ Estimating marginal probability of the observed variables
  - Spectral HMMs (Hsu et al. COLT'09)
  - Kernel spectral HMMs (Song et al. ICML'10)
  - Spectral latent tree (Parikh et al. ICML'11, Song et al. NIPS'11)
  - Spectral dimensional reduction for HMMs (Foster et al. Arxiv)
  - More recent: Cohen et al. ACL'12, Balle et al. ICML'12

- Estimating latent parameters
  - PCA approach (Mossel & Roch AOAP'06)
  - PCA and SVD approach, (Anandkumar et al. COLT'12, Arxiv)

- Estimating the structure of latent variable models
  - Recursive grouping (Choi et al. JMLR'11)
  - Spectral short quartet (Anandkumar et al. NIPS'11)

Challenge of Estimating Marginal of Observed Variables

- Exponential number of entries in $P(X_1, X_2, \ldots, X_6)$
  - Discrete variable taking n possible values, P has $O(n^6)$ entries!

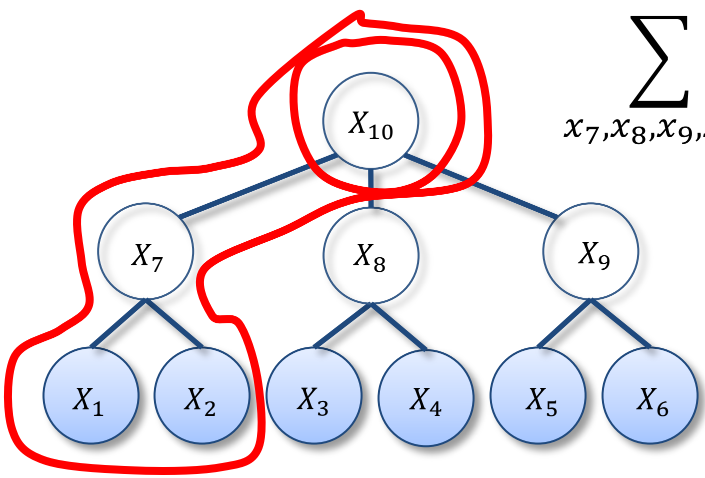
- Latent tree reduces the number of parameters

$$P(X_1, X_2, \ldots, X_6) = \sum_{x_7, x_8, x_9, x_{10}} P(X_1, X_2, \ldots, X_6, x_7, \ldots, x_{10})$$

$$\sum_{x_7, x_8, x_9, x_{10}} P(X_{10}) \, P(x_7|x_{10}) P(X_1|x_7) P(X_2|x_7) \, P(x_8|x_{10}) P(X_3|x_8) P(X_4|x_8) \, P(x_9|x_{10}) P(X_5|x_9) P(X_6|x_9)$$

$P(X_{10})$: $O(n)$ params

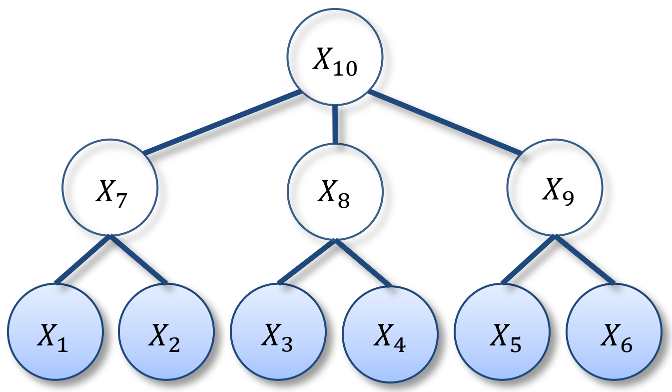
$P(x_7|x_{10})P(X_1|x_7)P(X_2|x_7)$: $O(3n^2)$ params



Latent tree has $O(9n^2)$ params
Significant saving!

EM Algorithm for Parameter Estimation

- Do not observe latent variables, need to estimate the corresponding parameters, eg., $P(X_7|X_{10})$ and $P(X_1|X_7)$



$i=1$	x_1^1	x_2^1	x_3^1	x_4^1	x_5^1	x_6^1
$\vdots$	$\vdots$					$\vdots$
$i=m$	x_1^m	x_2^m	x_3^m	x_4^m	x_5^m	x_6^m

Goal of spectral algorithm: Estimate the marginal in local-minimum-free fashion

- Expectation maximization: maximize likelihood of observations
  - $\max \prod_{i=1}^{m} P(x_1^i, \ldots, x_6^i)$
- Drawback: local maxima, slow to converge, difficult to analyze

Key Features of Spectral Algorithms

- Represent joint probability table of observed variables with low rank factorization, without using the joint table in the computation!

- Eg. $P_{\{1,\dots,d\};\{d+1,\dots,2d\}} = Reshape(P(X_1, \dots, X_{2d}), \{1, \dots, d\})$

n^d

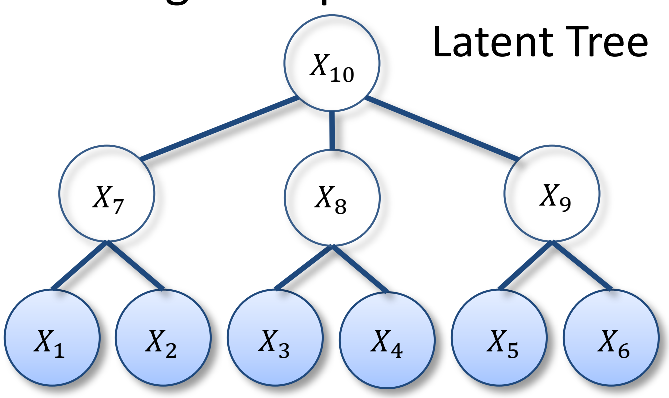
n^d $P_{\{1,\dots,d\};\{d+1,\dots,2d\}}$

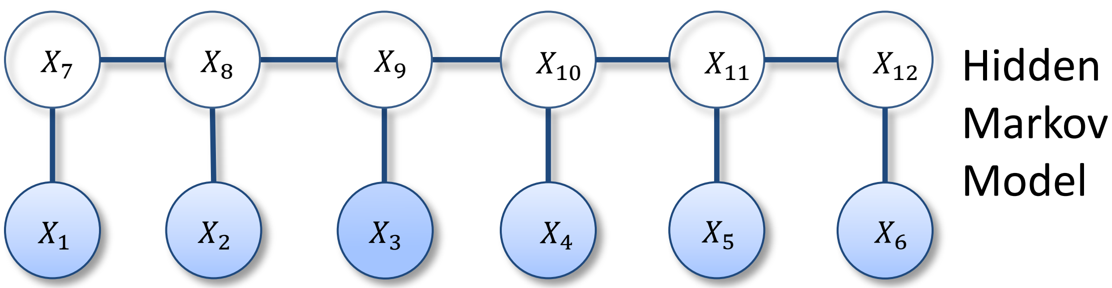
- Represent it by low rank factors to avoid exponential blowup
- Use clever decomposition technique to avoid directly using all entries from the table
- Use singular value decomposition

Tensor View of Marginal Probability

- Marginal probability table $\mathcal{T} = P(X_1, X_2, \dots, X_6)$
  - Discrete variable taking n possible values $\{1, \dots, n\}$
    - 6-way table, or 6^{th} order tensor
    - Dimension labeled by the variable
  - Value of the variable is the index to the corresponding dimension, need 6 indexes to access a single entry
    - $P(X_1 = 1, X_2 = 4, \dots, X_6 = 3)$ is the entry $\mathcal{T}[1,4, \dots, 3]$

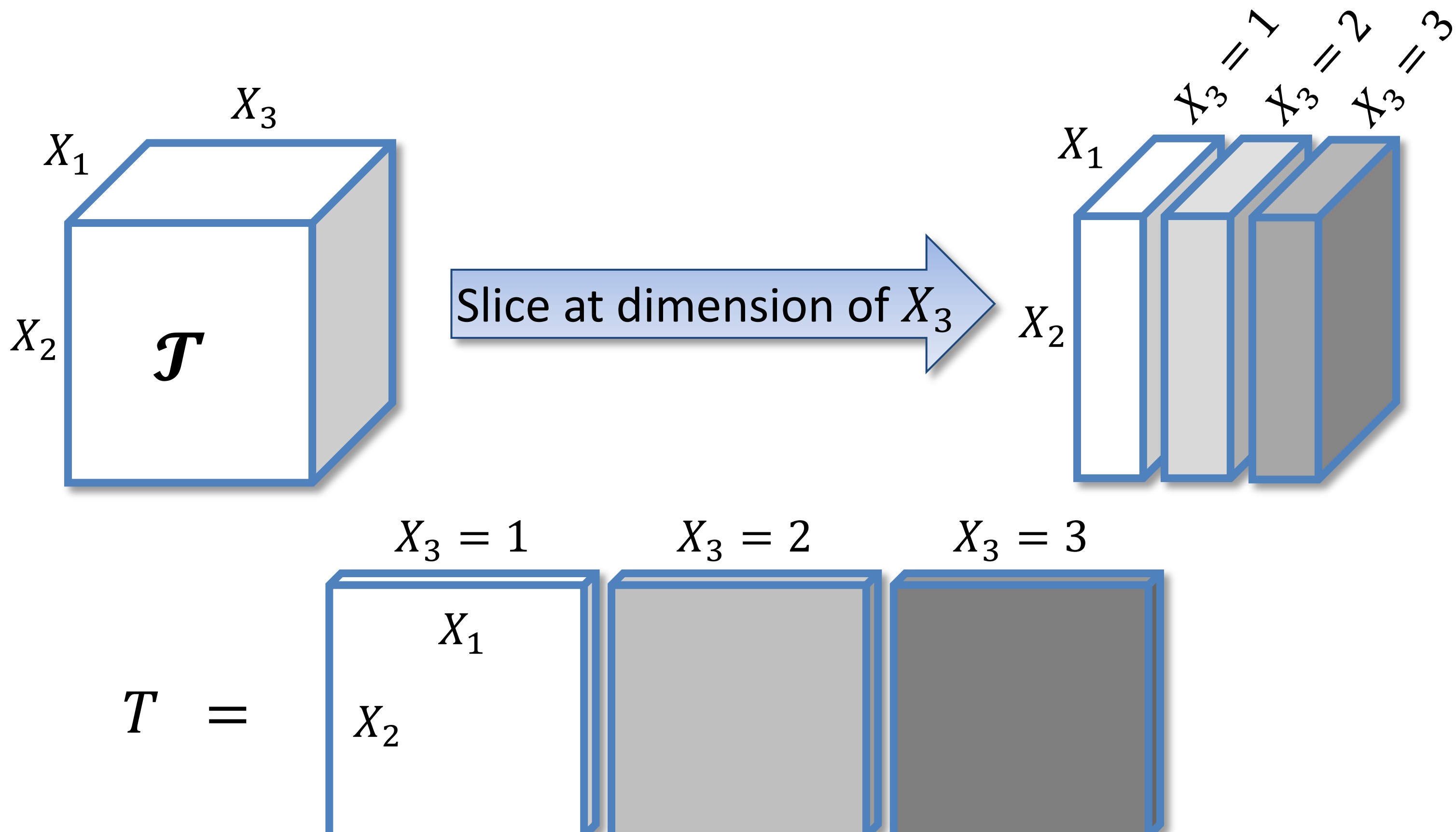
Running Examples:





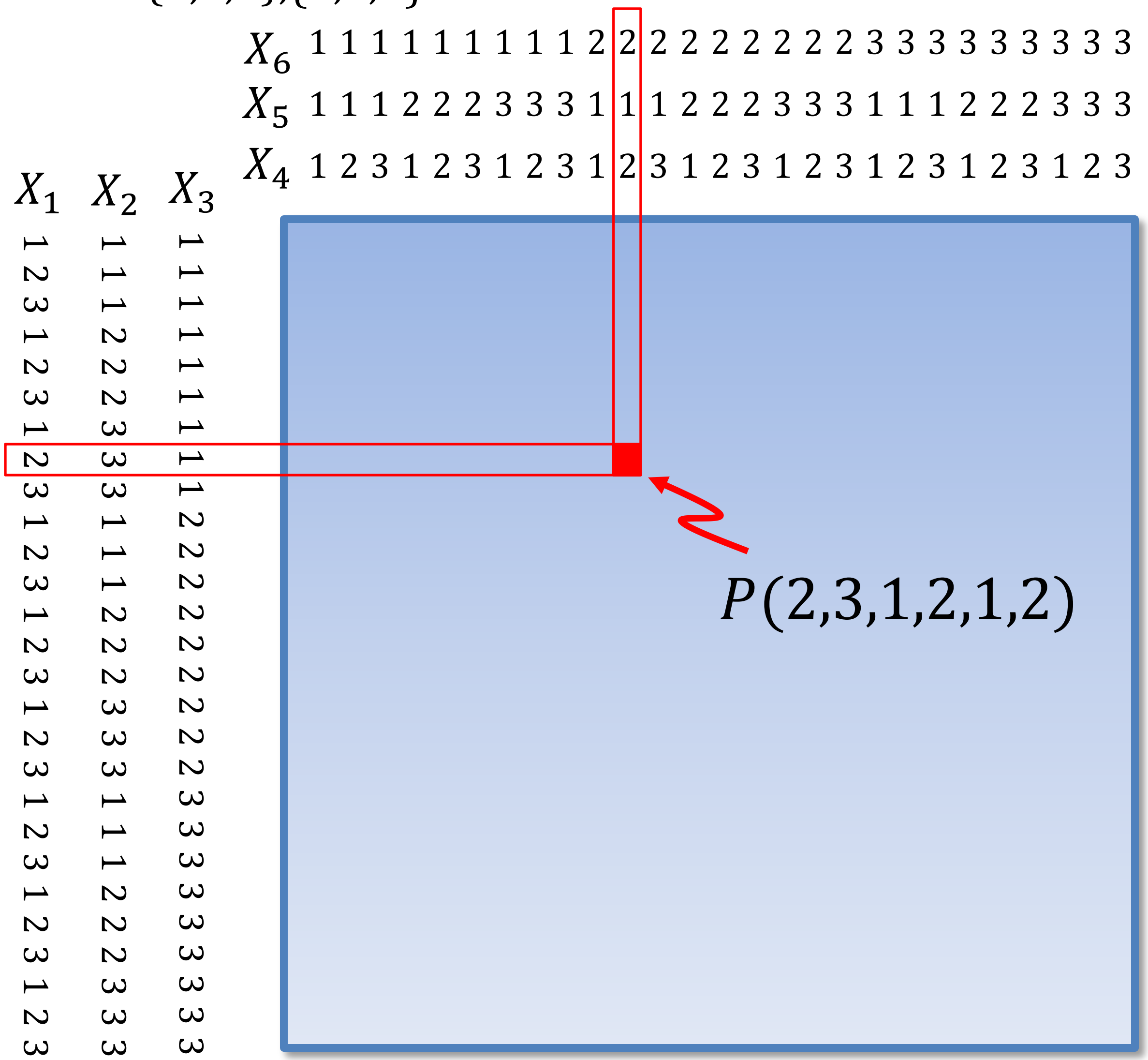
Reshaping Tensor into Matrices

- $T = Reshape(\mathcal{T}, \mathcal{C})$: multi-index $\mathcal{C}$ mapped into row index, and the remaining indexes into column index
  - Eg. $\mathcal{T} = P(X_1, X_2, X_3)$, a 3rd order tensor and $n = 3$
  - $P_{\{2\};\{1,3\}} = Reshape(\mathcal{T}, \{2\})$ turns the dimension of X_2 into row



Reshaping 6th Order Tensor

- $T = P_{\{1,2,3\};\{4,5,6\}} = Reshape(P(X_1, \dots, X_6), \{1,2,3\})$



Each entry is the probability of a unique assignment to $X_1, \dots, X_6$

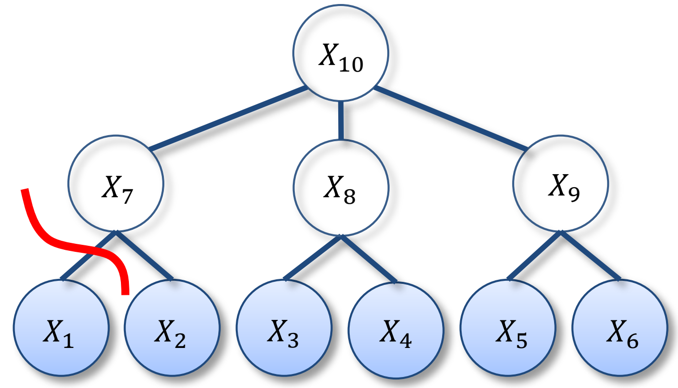
Reshaping according to Latent Tree Structure

- For marginal $\boldsymbol{\mathcal{P}} = P(X_1, X_2, \dots, X_6)$ of a latent tree model, reshape it according to the edges in the tree

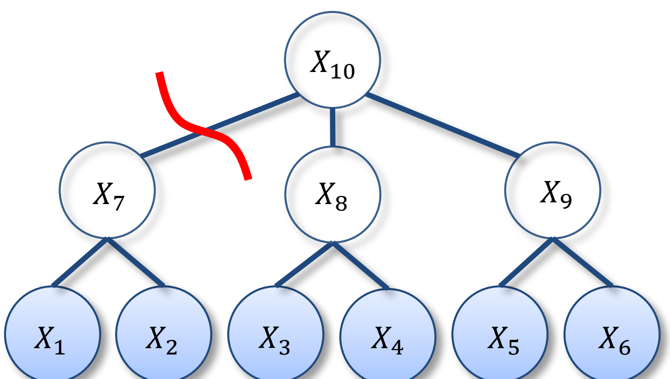
- $P_{\{1\};\{2,3,4,5,6\}} = Reshape(\boldsymbol{\mathcal{P}}, \{1\})$

- $P_{\{1,2\};\{3,4,5,6\}} = Reshape(\boldsymbol{\mathcal{P}}, \{1,2\})$

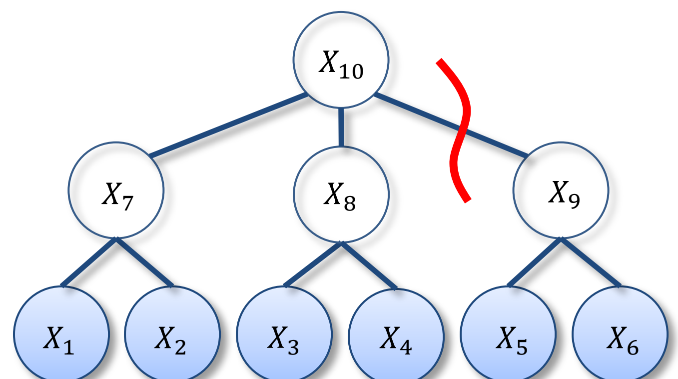
- $P_{\{1,2,3,4\};\{5,6\}} = Reshape(\boldsymbol{\mathcal{P}}, \{1,2,3,4\})$



$P_{\{1\};\{2,3,4,5,6\}}$



$P_{\{1,2\};\{3,4,5,6\}}$



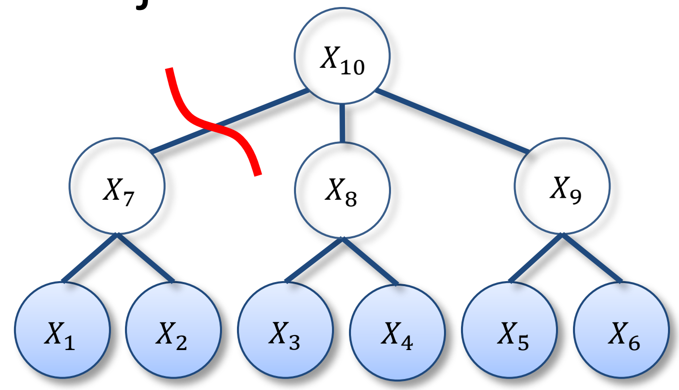
$P_{\{1,2,3,4\};\{5,6\}}$

Low Rank Structure after Reshaping

- Size of $P_{\{1,2\};\{3,4,5,6\}}$ is $n^2 \times n^4$, but its rank is just n

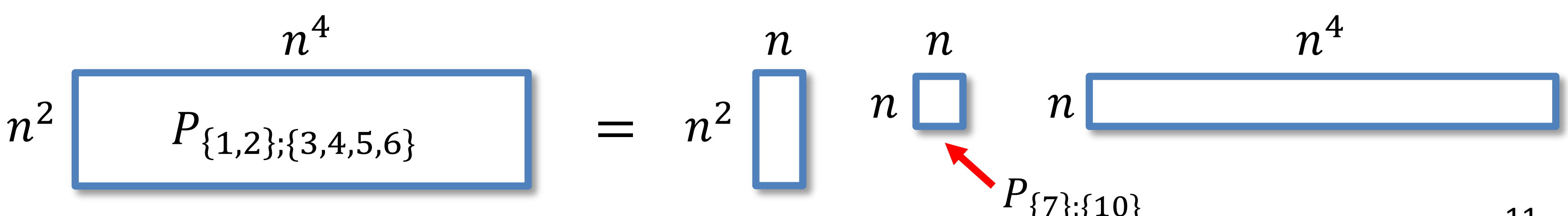
- $P(X_1, X_2, \ldots, X_6) =$

$$\sum_{x_7, x_{10}} P(X_1, X_2 | x_7)\ P(x_7, x_{10})\ P(X_3, X_4, X_5, X_6 | x_{10})$$



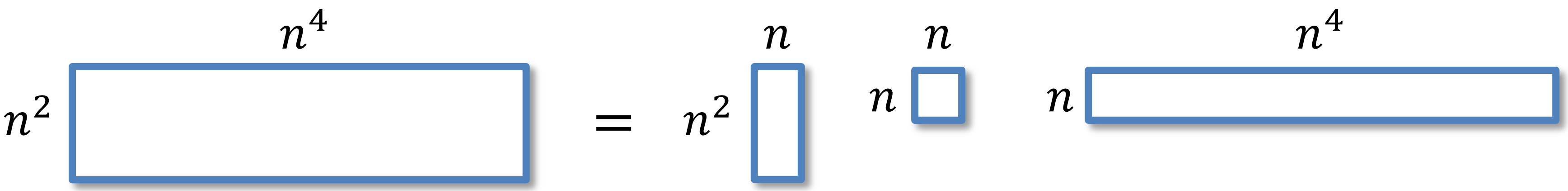
$P_{\{1,2\};\{3,4,5,6\}}$

- Use matrix multiplications to express summation over X_7, X_{10}
  - $P_{\{1,2\};\{3,4,5,6\}} = P_{\{1,2\}|\{7\}}\ P_{\{7\};\{10\}}\ P^{\top}_{\{3,4,5,6\}|\{10\}}$
  - $P_{\{1,2\}|\{7\}} := Reshape(P(X_1, X_2 | X_7), \{1,2\})$
  - $P_{\{3,4,5,6\}|\{10\}} := Reshape(P(X_3, X_4, X_5, X_6 | X_{10}), \{3,4,5,6\})$

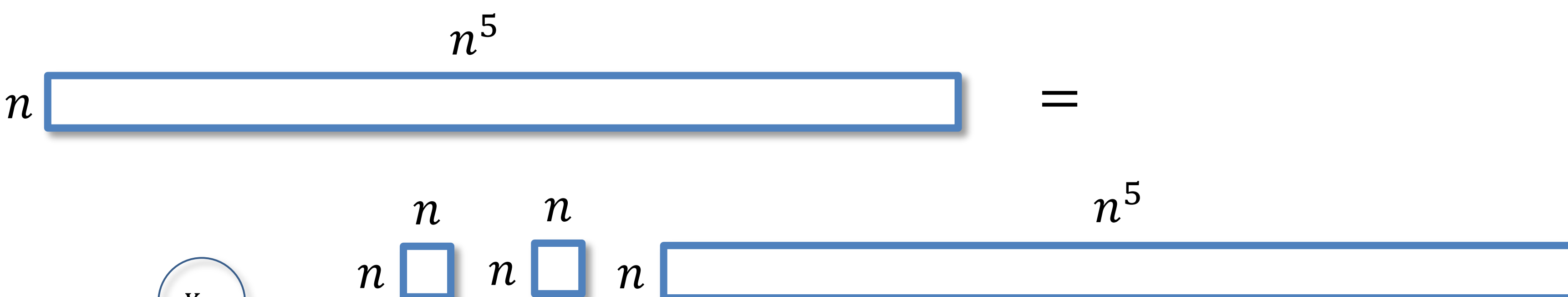


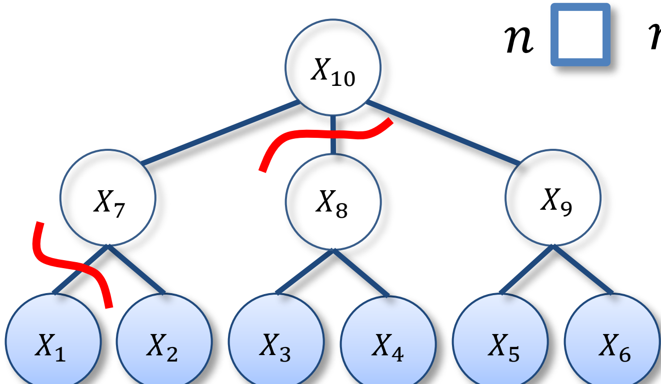
Low Rank Structure of Latent Tree Model

- $P_{\{3,4\};\{1,2,5,6\}} = P_{\{3,4\}|\{8\}}\, P_{\{8\};\{10\}}\, P^{\top}_{\{1,2,5,6\}|\{10\}}$

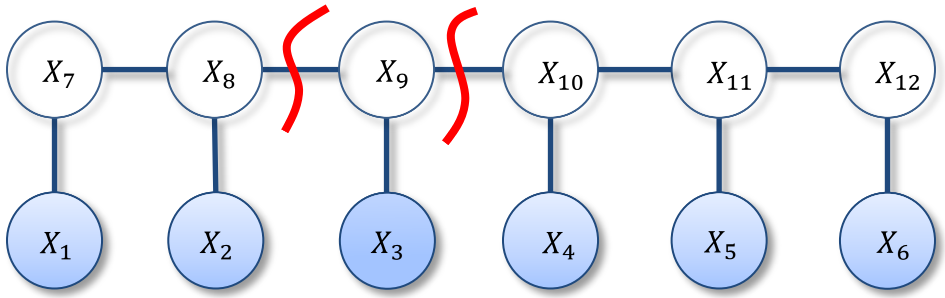


- $P_{\{1\};\{2,3,4,5,6\}} = P_{\{1\}|\{7\}}\, P_{\{7\};\{7\}}\, P^{\top}_{\{2,3,4,5,6\}|\{7\}}$

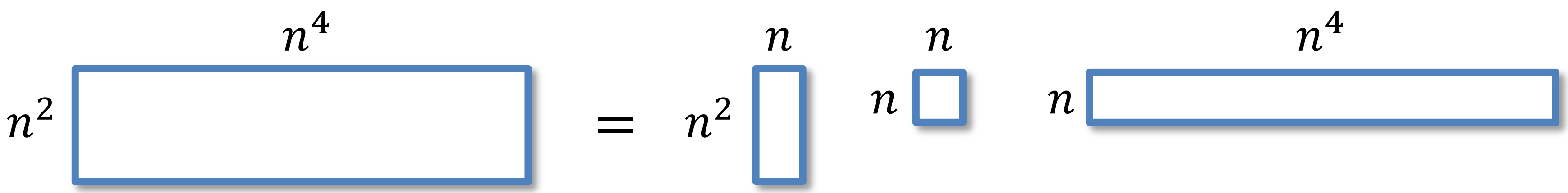




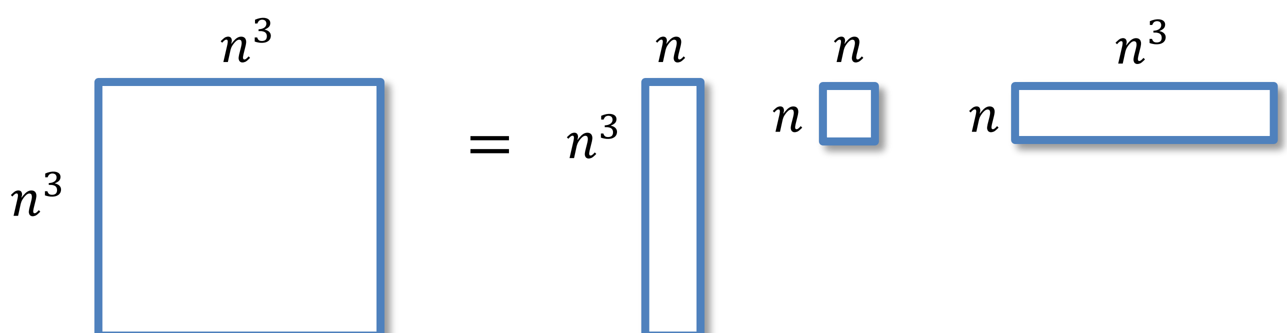
All these reshapings are low rank, and with rank n



- $P_{\{1,2\};\{3,4,5,6\}} = P_{\{1,2\}|\{8\}} \, P_{\{8\};\{9\}} \, P^{\top}_{\{3,4,5,6\}|\{9\}}$



- $P_{\{1,2,3\};\{4,5,6\}} = P_{\{1,2,3\}|\{9\}} \, P_{\{9\};\{10\}} \, P^{\top}_{\{4,5,6\}|\{10\}}$



Key Features of Spectral Algorithms

- Represent joint probability table of observed variables with low rank factorization, without using the joint table in the computation!

- Eg. $P_{\{1,\ldots,d\};\{d+1,\ldots,2d\}} = Reshape(P(X_1, \ldots, X_{2d}), \{1, \ldots, d\})$

n^d

n^d $P_{\{1,\ldots,d\};\{d+1,\ldots,2d\}}$

- Represent it by low rank factors to avoid exponential blowup
- Use clever decomposition technique to avoid directly using all entries from the table
- Use singular value decomposition

Key Theorem

Theorem 1:

$$P: \ size\ m \times n, rank\ k$$
$$A: \ size\ n \times k, rank\ k$$
$$B: \ size\ k \times m, rank\ k$$
$$If\ (BPA)\ invertible, then\ P = (PA)(BPA)^{-1}(BP)$$

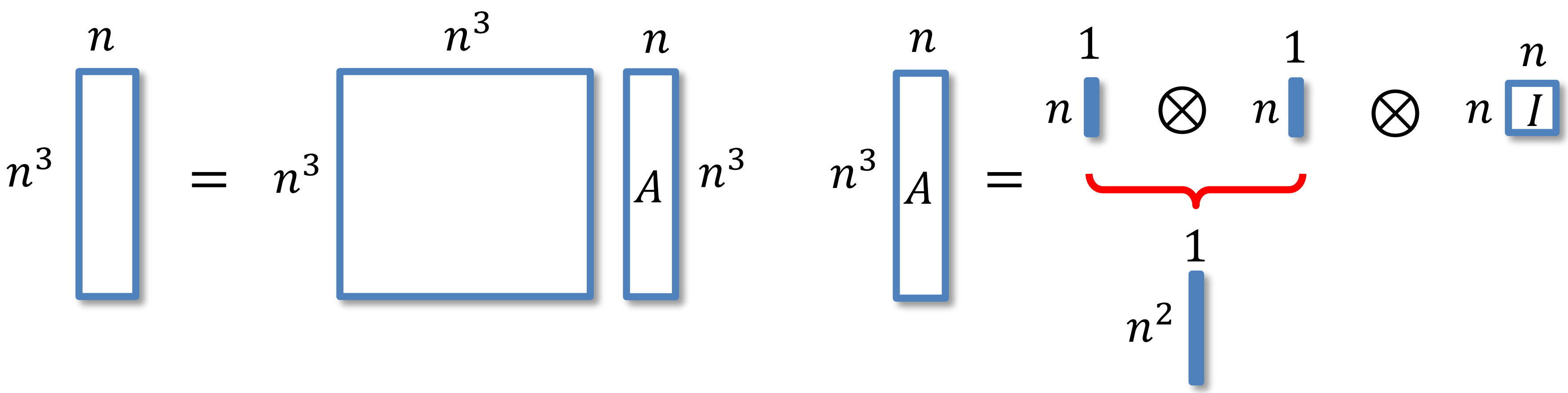
- P will be the reshaped joint probability table
- A and B will be marginalization operator
- Theorem 1 will be applied recursively
- Recover several existing spectral algorithms as special cases

Marginalization Operator A and B

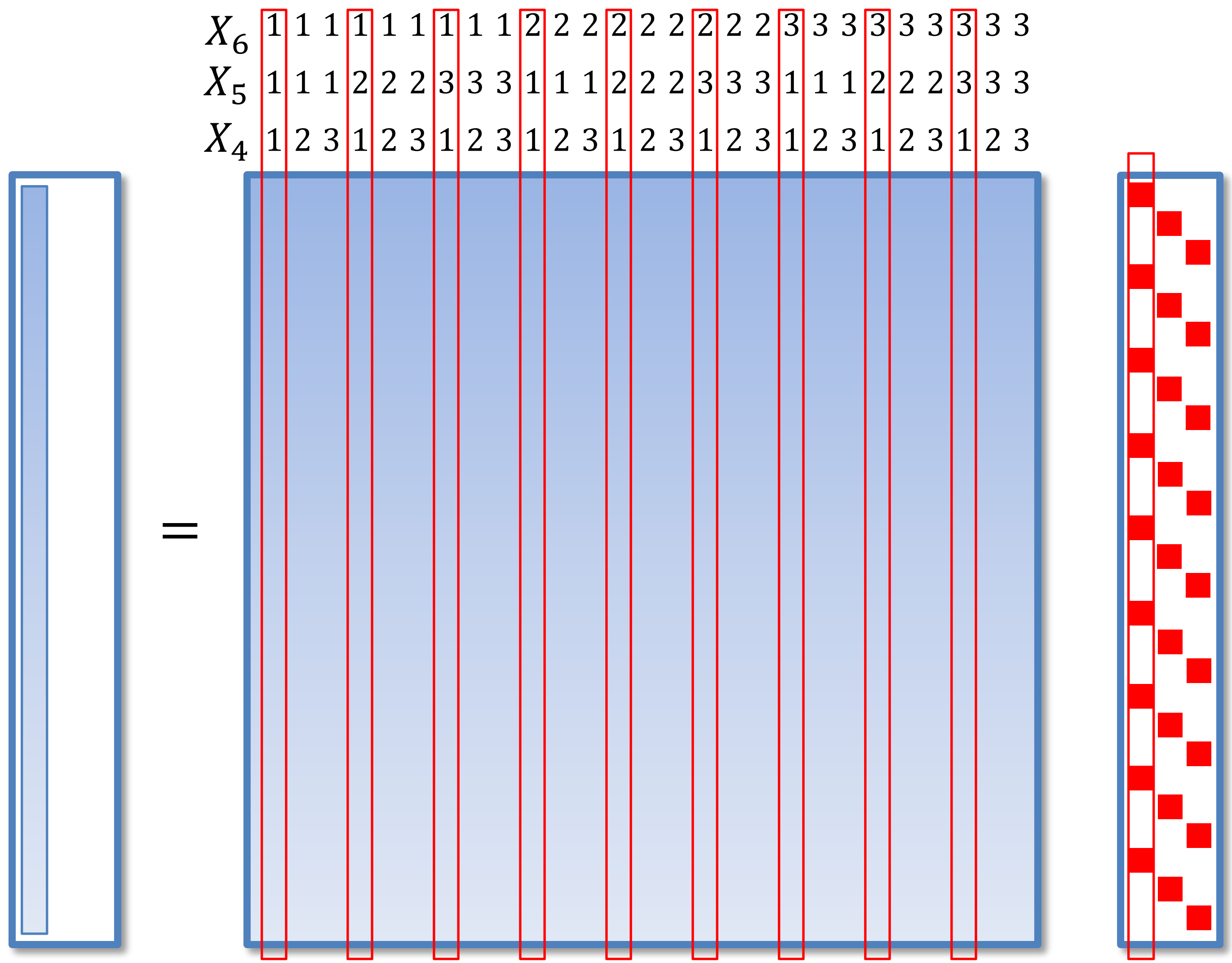
- Compute the marginal probability of a subset of variables can be expressed as matrix product

- $P(X_1, X_2, X_3, X_4) = \sum_{x_5, x_6} P(X_1, X_2, X_3, X_4, x_5, x_6)$

- $P_{\{1,2,3\};\{4\}} = P_{\{1,2,3\};\{4,5,6\}} A$, where $A = 1_n \otimes 1_n \otimes I_n$



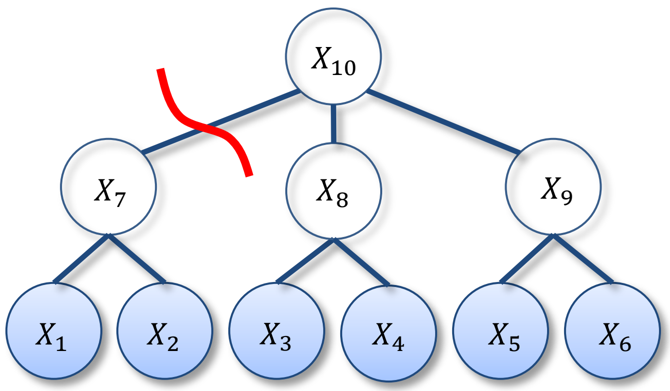
Zoom into Marginalization Operation



$P_{\{1,2,3\};\{4\}}$ $\quad$ $P_{\{1,2,3\};\{4,5,6\}}$ $\quad$ $1_3 \otimes 1_3 \otimes I_3$

Apply Theorem 1 to Latent Tree Model

- Let
  - $P = P_{\{1,2\};\{3,4,5,6\}}$
  - $A = 1_n \otimes 1_n \otimes 1_n \otimes I_n$
  - $B = (I_n \otimes 1_n)^\top$



$P_{\{1,2\};\{3,4,5,6\}}$

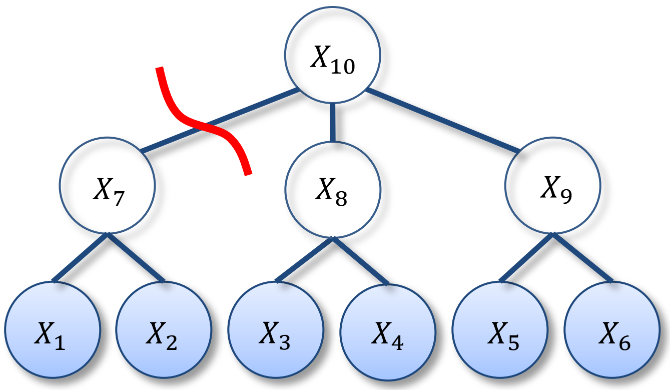
- Then
  - $P_{\{1,2\};\{3,4,5,6\}}A = P_{\{1,2\};\{3\}}$
  - $BP_{\{1,2\};\{3,4,5,6\}} = P_{\{2\};\{3,4,5,6\}}$
  - $BP_{\{1,2\};\{3,4,5,6\}}A = P_{\{2\};\{3\}}$

- Finally use $P = (PA)(BPA)^{-1}(BP)$
  - $P_{\{1,2\};\{3,4,5,6\}} = P_{\{1,2\};\{3\}}P^{-1}_{\{2\};\{3\}}P_{\{2\};\{3,4,5,6\}}$

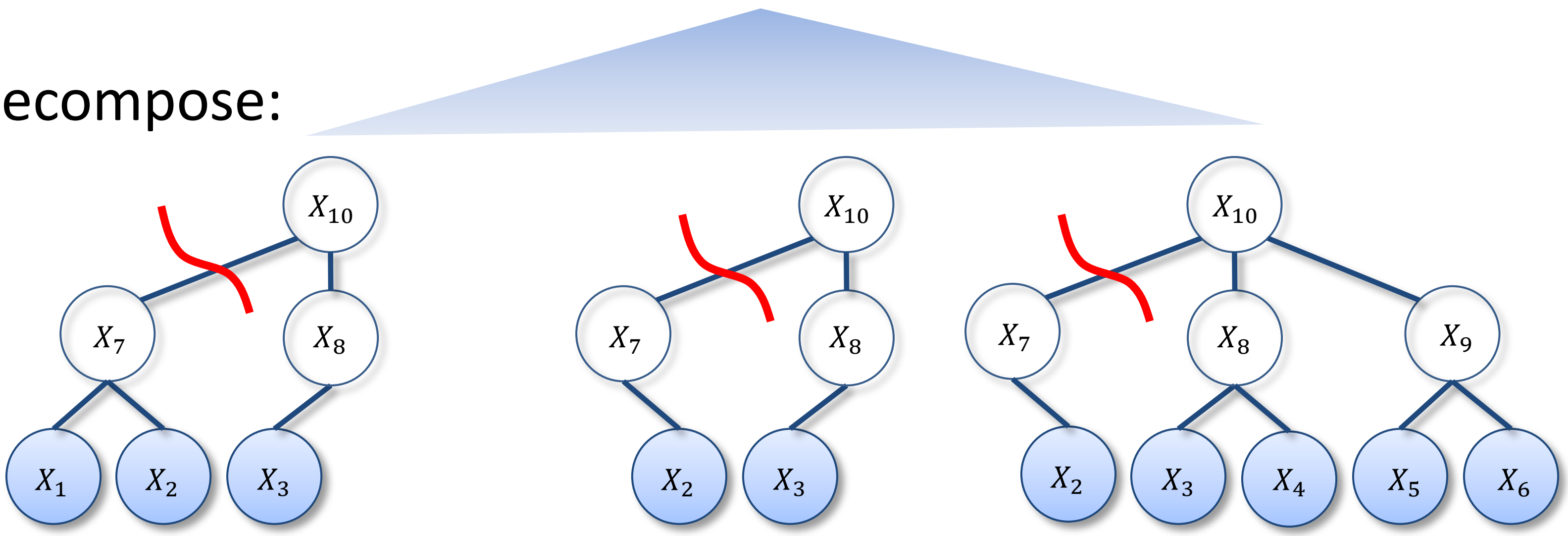
Latent Tree Decomposition

- $P_{\{1,2\};\{3,4,5,6\}} = P_{\{1,2\};\{3\}} P^{-1}_{\{2\};\{3\}} P_{\{2\};\{3,4,5,6\}}$

$P_{\{1,2\};\{3,4,5,6\}}$



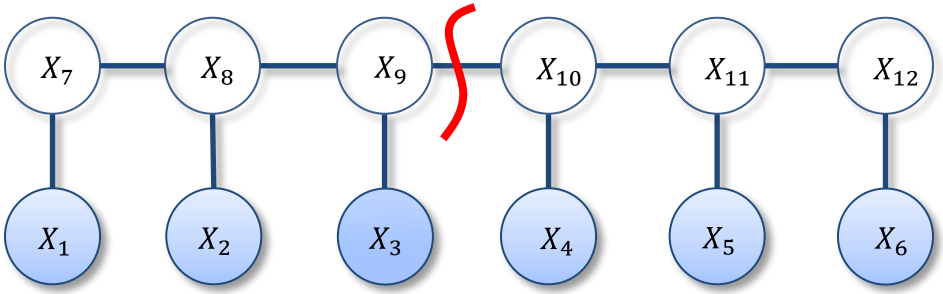
Decompose:



$P_{\{1,2\};\{3\}}$ $P^{-1}_{\{2\};\{3\}}$ $P_{\{2\};\{3,4,5,6\}}$

Apply Theorem 1 to Hidden Markov Models

- Let
  - $P = P_{\{1,2,3\};\{4,5,6\}}$
  - $A = 1_n \otimes 1_n \otimes I_n$
  - $B = (I_n \otimes 1_n \otimes 1_n)^\top$



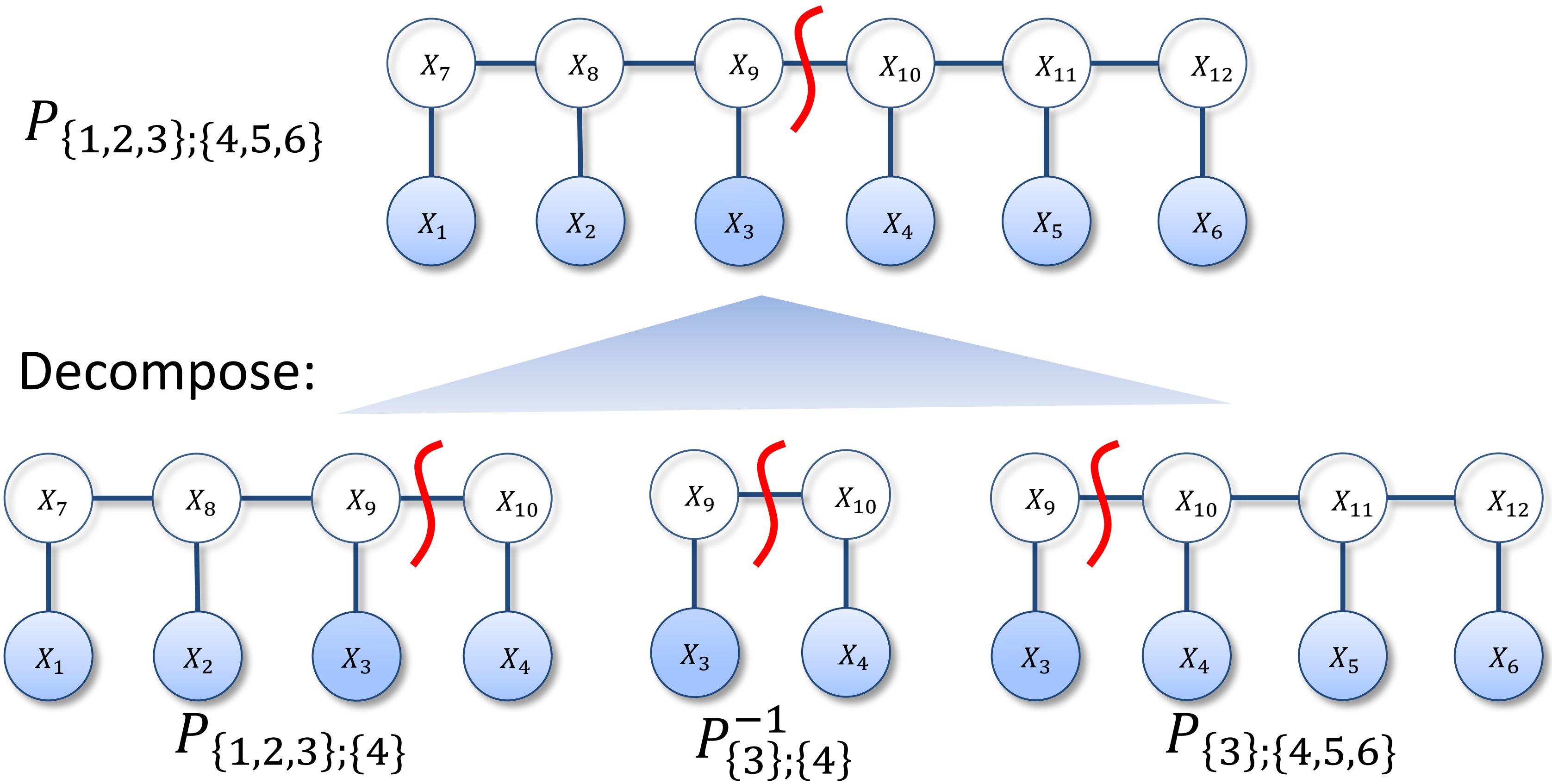
$P_{\{1,2,3\};\{4,5,6\}}$

- Then
  - $P_{\{1,2,3\};\{4,5,6\}}A = P_{\{1,2,3\};\{4\}}$
  - $BP_{\{1,2,3\};\{4,5,6\}} = P_{\{3\};\{4,5,6\}}$
  - $BP_{\{1,2,3\};\{4,5,6\}}A = P_{\{3\};\{4\}}$

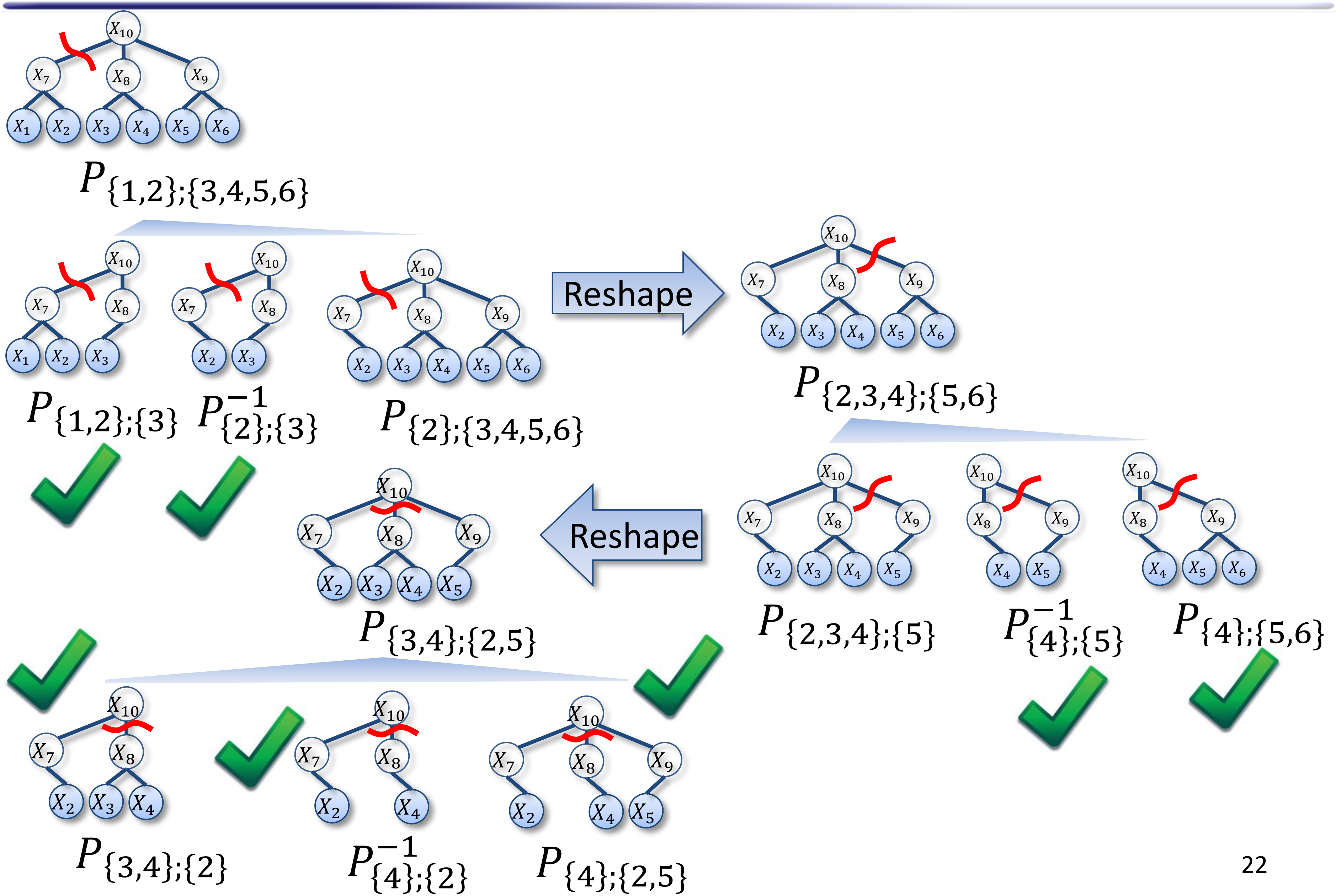
- Finally use $P = (PA)(BPA)^{-1}(BP)$
  - $P_{\{1,2,3\};\{4,5,6\}} = P_{\{1,2,3\};\{4\}}P_{\{3\};\{4\}}^{-1}P_{\{3\};\{4,5,6\}}$

Hidden Markov Model Decomposition

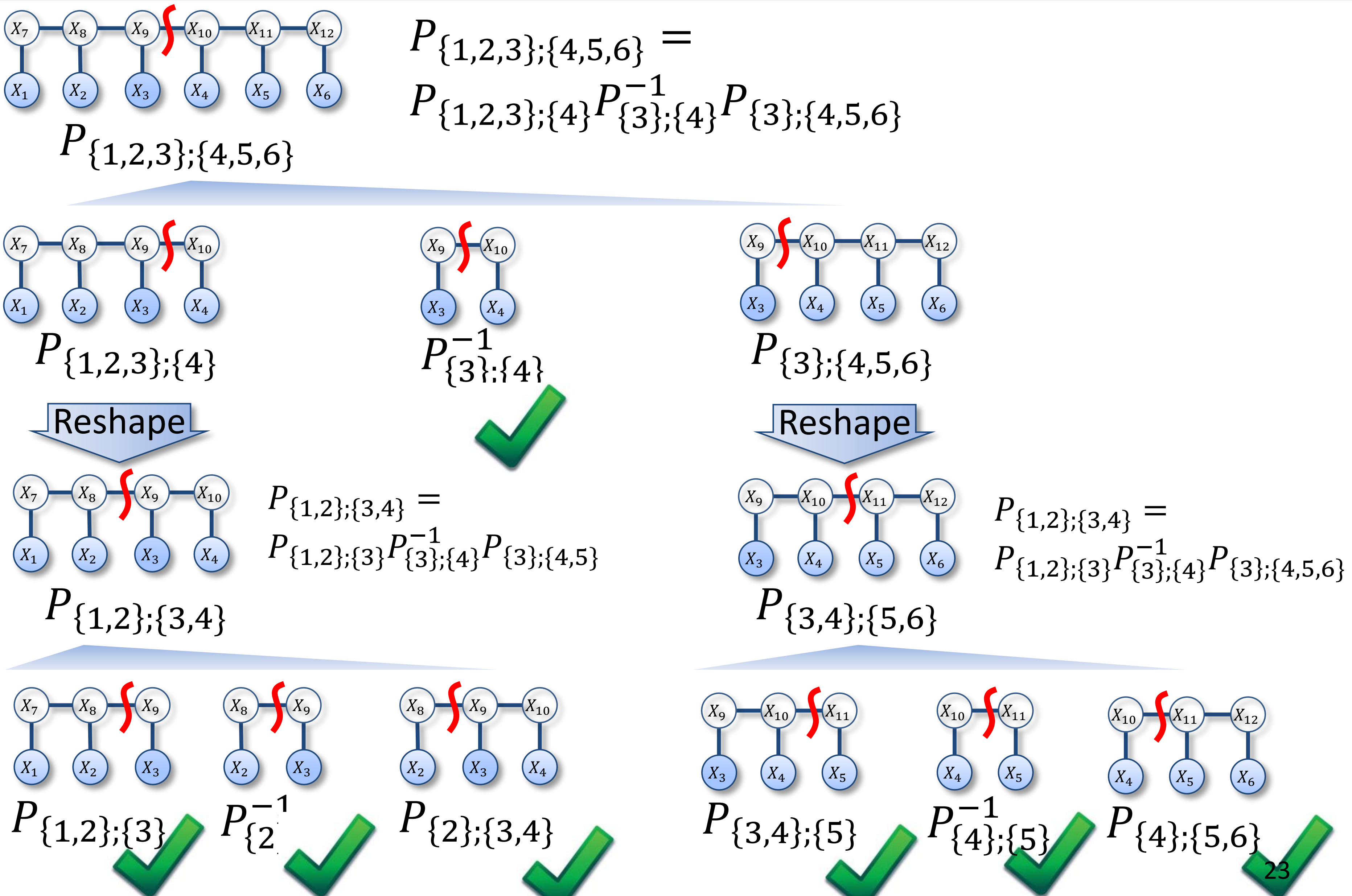
- $P_{\{1,2,3\};\{4,5,6\}} = P_{\{1,2,3\};\{4\}} P^{-1}_{\{3\};\{4\}} P_{\{3\};\{4,5,6\}}$



Recursive Decomposition of Latent Tree



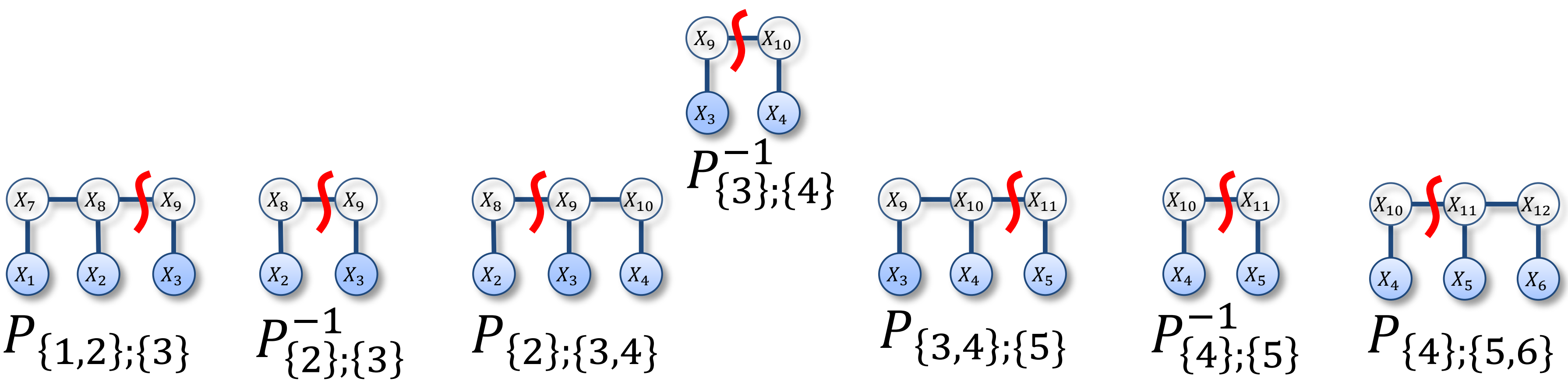
Recursive Decomposition of HMM



One Entries in Joint Probability Table of HMM

- Fix some observations
  - Fix $X_3 = x_3$, $P_{\{2\};x_3;\{4\}} := P(X_2, x_3, X_4)$ is a matrix
  - Fix $X_2 = x_2$, $X_2 = x_2$, $P_{x_1;x_2;\{3\}} := P(x_1, x_2, X_3)$ is a vector

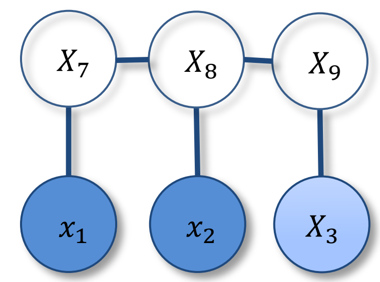
- $P(x_1, x_2, x_3, x_4, x_5, x_6)$
  $= P_{x_1;x_2;\{3\}}\ P^{-1}_{\{2\};\{3\}} P_{\{2\};x_3;\{4\}} P^{-1}_{\{3\};\{4\}} P_{\{3\};x_4;\{5\}} P^{-1}_{\{4\};\{5\}} P_{\{4\};x_5;x_6}$

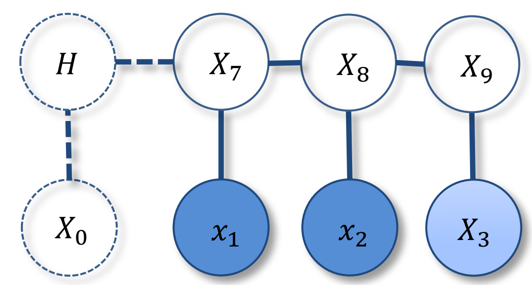


Connection to Foster et al.

- $P(x_1, x_2, x_3, x_4, x_5, x_6)$
  $= P_{x_1;x_2;\{3\}}\, P^{-1}_{\{2\};\{3\}} P_{\{2\};x_3;\{3\}} P^{-1}_{\{3\};\{4\}} P_{\{4\};x_4;\{5\}} P^{-1}_{\{4\};\{5\}} P_{\{4\};x_5;x_6}$

- Introduce variable X_0 into $P_{x_1;x_2;\{3\}}$
  - $= 1^\top P_{\{0\};x_1;\{1\}} P^{-1}_{\{1\};\{2\}} P_{\{1\};x_2;\{3\}}$
  - $= 1^\top P_{\{0\};\{1\}} P^{-1}_{\{0\};\{1\}} P_{\{0\};x_1;\{1\}} P^{-1}_{\{1\};\{2\}} P_{\{1\};x_2;\{3\}}$
  - $= P^\top_{\{1\}} P^{-1}_{\{0\};\{1\}} P_{\{0\};x_1;\{1\}} P^{-1}_{\{1\};\{2\}} P_{\{1\};x_2;\{3\}}$





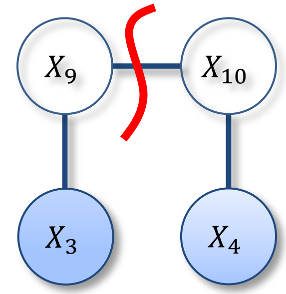
- Do similar things to $P_{\{6\};x_5;x_6}$

- Assume time homogeneous
  - $P^{-1}_{\{0\};\{1\}} = P^{-1}_{\{1\};\{2\}}, P_{\{1,2\};\{3\}} = P_{\{2,3\};\{4\}}$

What if Hidden State $k <$ Observed State n

- Let $P = P_{\{1,2,3\};\{4,5,6\}}, A = 1_n \otimes 1_n \otimes I_n, B = (I_n \otimes 1_n \otimes 1_n)^\top$.

- Use $P = (PA)(BPA)^{-1}(BP)$
  - $P_{\{1,2,3\};\{4,5,6\}} = P_{\{1,2,3\};\{4\}} P_{\{3\};\{4\}}^{-1} P_{\{3\};\{4,5,6\}}$



$P_{\{3\};\{4\}}^{-1}$

- $P_{\{3\};\{4\}}$ of size $n \times n$ has rank k and not invertible!
  - Singular Value decomposition of $P_{\{3\};\{4\}} = U_k \Sigma_k V_k^\top$

- Solution: Use further projection such that (BPA) is invertible
  - Let $A = (1_n \otimes 1_n \otimes I_n)V_k, B = U_k^\top (I_n \otimes 1_n \otimes 1_n)^\top$
  - $P_{\{1,2,3\};\{4,5,6\}} = P_{\{1,2,3\};\{4\}} V_k \left(U_k^\top P_{\{3\};\{4\}} V_k\right)^{-1} U_k^\top P_{\{3\};\{4,5,6\}}$

Connection to Hsu et al.

- Two equivalent forms of applying further projection U_k and V_k
  - $P_{\{3\};\{4\}} = U_k \Sigma_k V_k^\top$

- $\left(U_k^\top P_{\{3\};\{4\}} V_k\right)^{-1} U_k^\top P_{\{3\};\{4,5,6\}} = \left(P_{\{3\};\{4\}} V_k\right)^{\dagger} P_{\{3\};\{4,5,6\}}$

- $P(x_1, x_2, x_3, x_4, x_5, x_6)$

$$= P_{x_1;x_2;\{3\}} V_k \quad \left(P^{-1}_{\{2\};\{3\}} V_k\right)^{\dagger} P_{\{2\};x_3;\{4\}} V_k \left(P^{-1}_{\{3\};\{4\}} V_k\right)^{\dagger} P_{\{4\};x_4;\{5\}} V_k \quad \left(P^{-1}_{\{4\};\{5\}} V_k\right)^{\dagger} P_{\{4\};x_5;x_6}$$

- $b_1^\top B_{x_1} \ldots B_{x_6} b_\infty$

Proof of Theorem 1

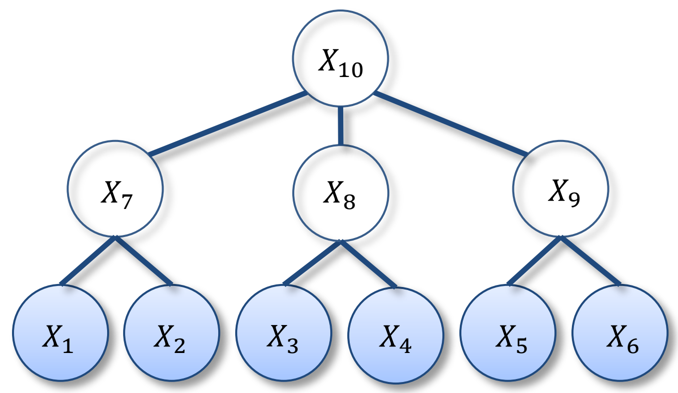
Georgia Tech

Theorem 1:
$Let\ P\ be\ a\ rank\ k\ matrix\ of\ size\ m \times n,$
$A\ be\ a\ rank\ k\ matrix\ of\ size\ n \times k,$
$B\ be\ a\ rank\ k\ matrix\ of\ size\ k \times m,$
$then\ P = (PA)(BPA)^{-1}(BP)$

- SVD: $P = U_k \Sigma_k V_k^\top + U_\perp 0 V_\perp^\top$
- Assume
  - $A = (V_k, V_\perp)\begin{pmatrix} C \\ D \end{pmatrix}$, C of size $k \times k$ and invertible
  - $B = (U_k, U_\perp)\begin{pmatrix} E \\ F \end{pmatrix}$, E of size $k \times k$ and invertible
- Plug the above A and B into $(PA)(BPA)^{-1}(BP)$

Finite Sample Estimator

- Given m iid samples, estimate pairwise and triplet marginals
- One-of-n encoding, e.g., $\phi(x=1) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, $\phi(x=2) = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$
- $\hat{P}_{\{1\};\{2\};\{3\}} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_1^i) \otimes \phi(x_2^i) \otimes \phi(x_3^i)$
- $\hat{P}_{\{1\};\{2\}} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_1^i)\, \phi(x_2^i)^{\top}$
- $\hat{P}_{\{1\}} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_1^i)$



$i=1$	x_1^1	x_2^1	x_3^1	x_4^1	x_5^1	x_6^1
$\vdots$	$\vdots$					$\vdots$
$i=m$	x_1^m	x_2^m	x_3^m	x_4^m	x_5^m	x_6^m

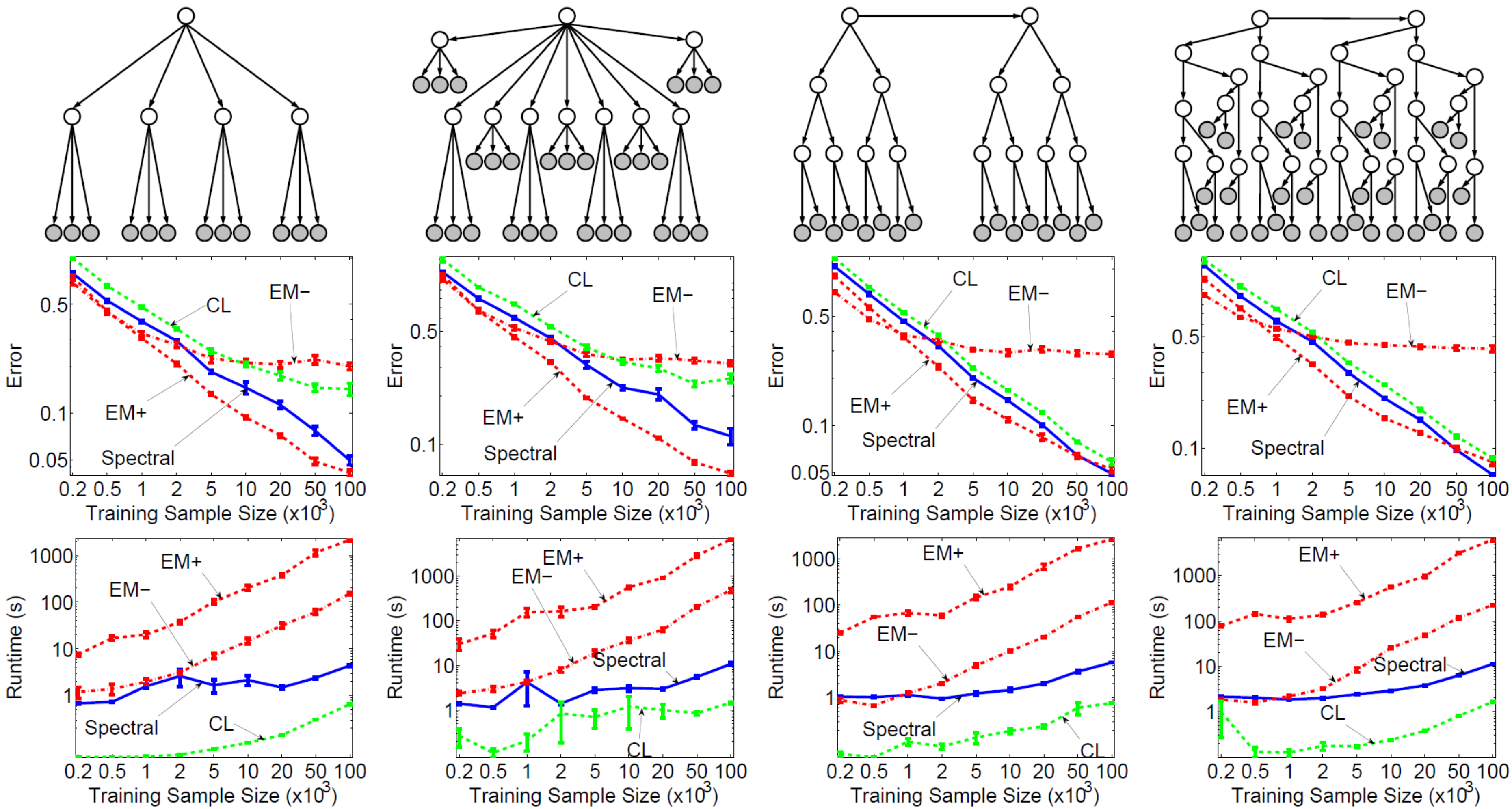
Sample Complexity Analysis

- Error in estimate $\widehat{P}_{\{1\};\{2\};\{3\}}, \widehat{P}_{\{1\};\{2\}}$
- Error propagation in the recursive decomposition
  - $P(x_1, x_2, x_3, x_4, x_5, x_6)$

$$= P_{x_1;x_2;\{3\}} P_{\{2\};\{3\}}^{-1} P_{\{2\};x_3;\{3\}} P_{\{3\};\{4\}}^{-1} P_{\{4\};x_4;\{5\}} P_{\{4\};\{5\}}^{-1} P_{\{4\};x_5;x_6}$$

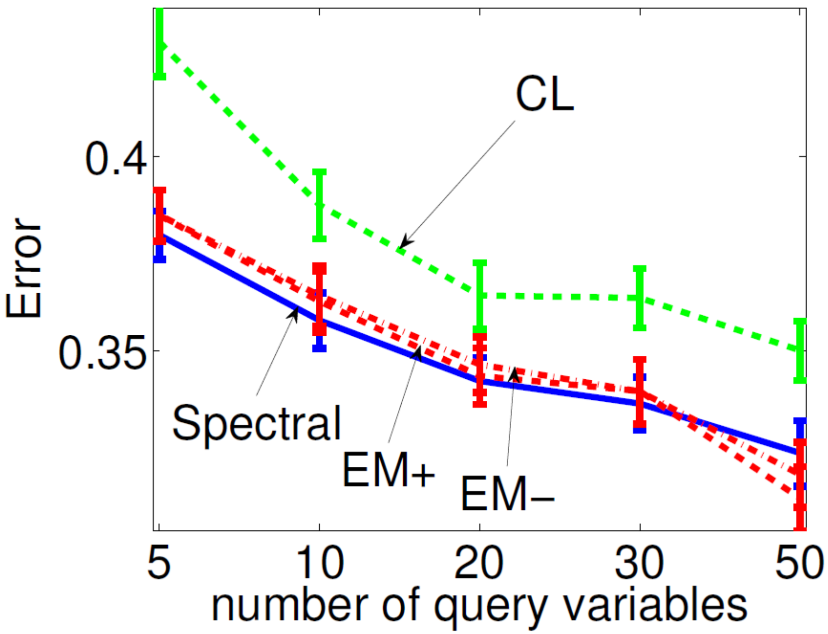
  - Depends on the smallest singular value of the invesion terms eg., $P_{\{1\};\{2\}}$
- Spectral algorithms
  - Use SVD for further projection
  - Error depends on singular value

Synthetic Data



Stock Trend Prediction Data

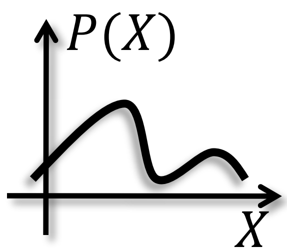
- 59 stocks, 6800 samples, learn latent structure first and then estimate the marginal
- MAP prediction task $x_i = argmax\ P(X_i | x_1, x_2, \dots, x_{i-1})$ (query i variables)
- Absolute error $|x_i - x_i^{\star}|$
- Also compared with Chow-Liu tree (fully observed model)

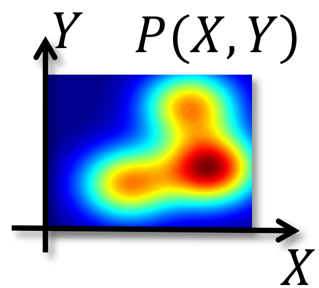


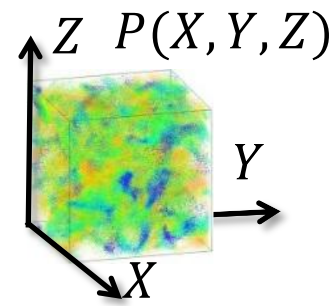
Non-discrete, Non-Gaussian Case

- Previous approach all about discrete variables

- Real world data can be continuous, and have multimodal distribution and other rich statistical features



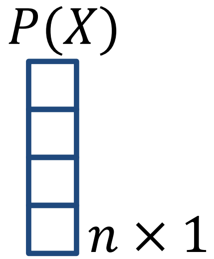


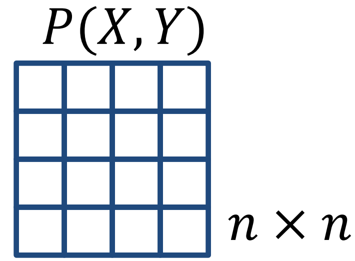


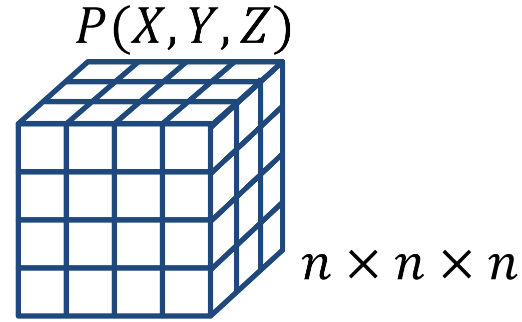
- Replace discrete probabilities by kernel embedding of distributions
  - $k(x,x') = \langle \phi(x), \phi(x') \rangle$, eg., $\exp(-s\|x - x'\|^2)$
  - Expected feature of distribution $\mu_X = \mathbb{E}_X[\phi(X)]$ (can be infinite dimensional feature)
  - One-of-n feature of discrete case is a special case

Kernel Embedding and Covariance Operator

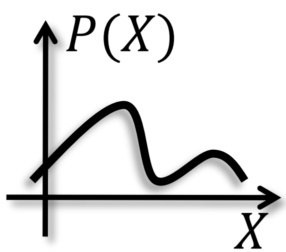
Discrete



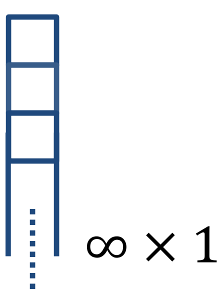


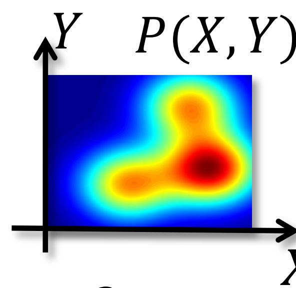


Kernel Embedding

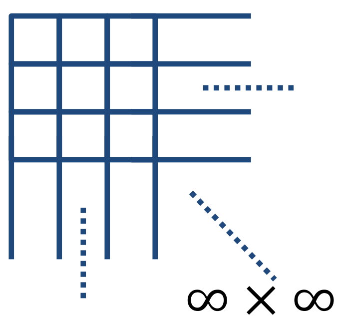


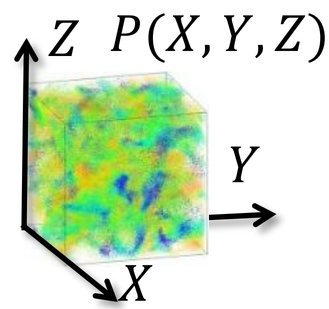
$\mu_X :=$
$\mathbb{E}_X[\phi(X)]$



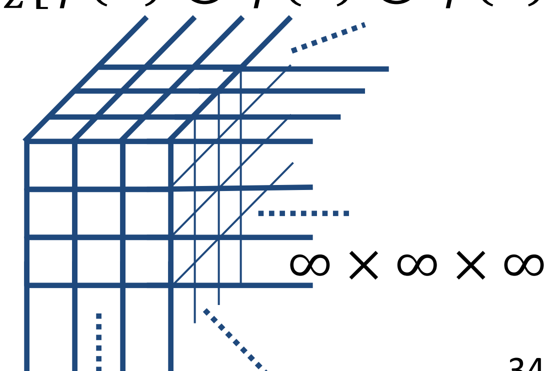


$\mathcal{C}_{XY} :=$
$\mathbb{E}_{XY}[\phi(X) \otimes \phi(Y)]$

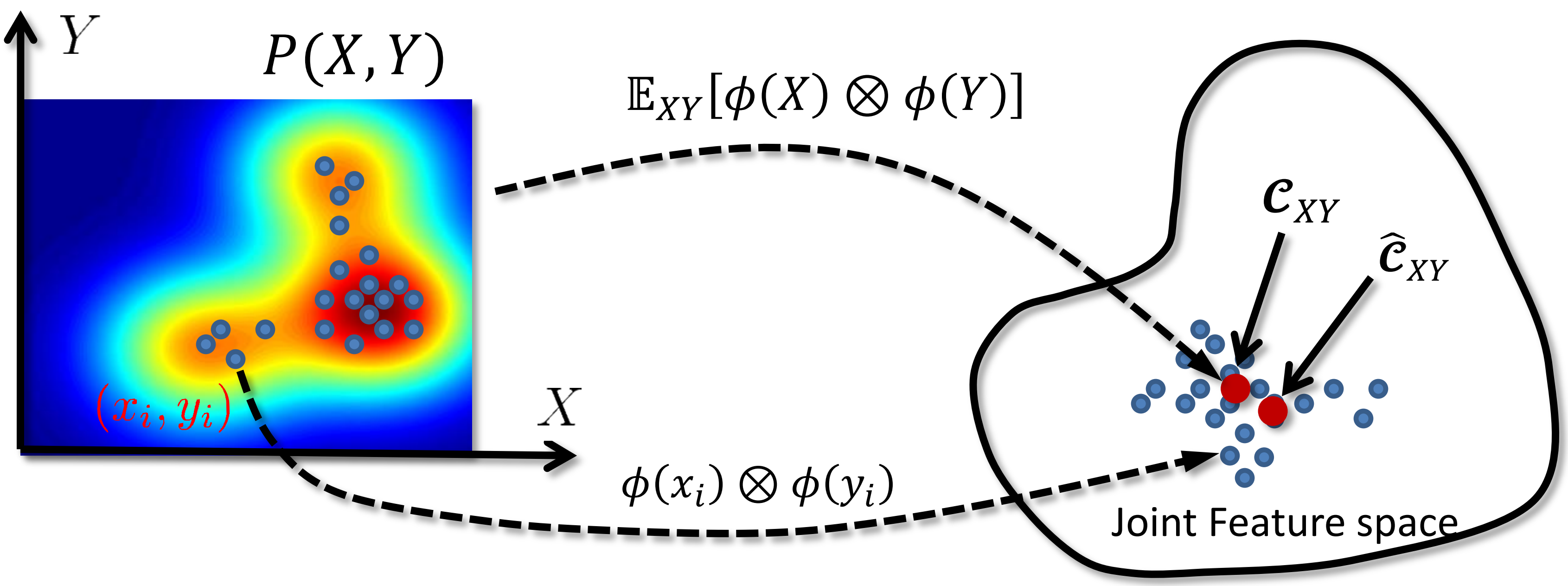




$\mathcal{C}_{XYZ} :=$
$\mathbb{E}_{XYZ}[\phi(X) \otimes \phi(Y) \otimes \phi(Z)]$



Kernel Embedding with Finite Sample



$$\mathcal{C}_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \phi(Y)] \approx \widehat{\mathcal{C}}_{XY} = \frac{1}{m}\sum_{i=1}^{m} \phi(x_i) \otimes \phi(y_i)$$

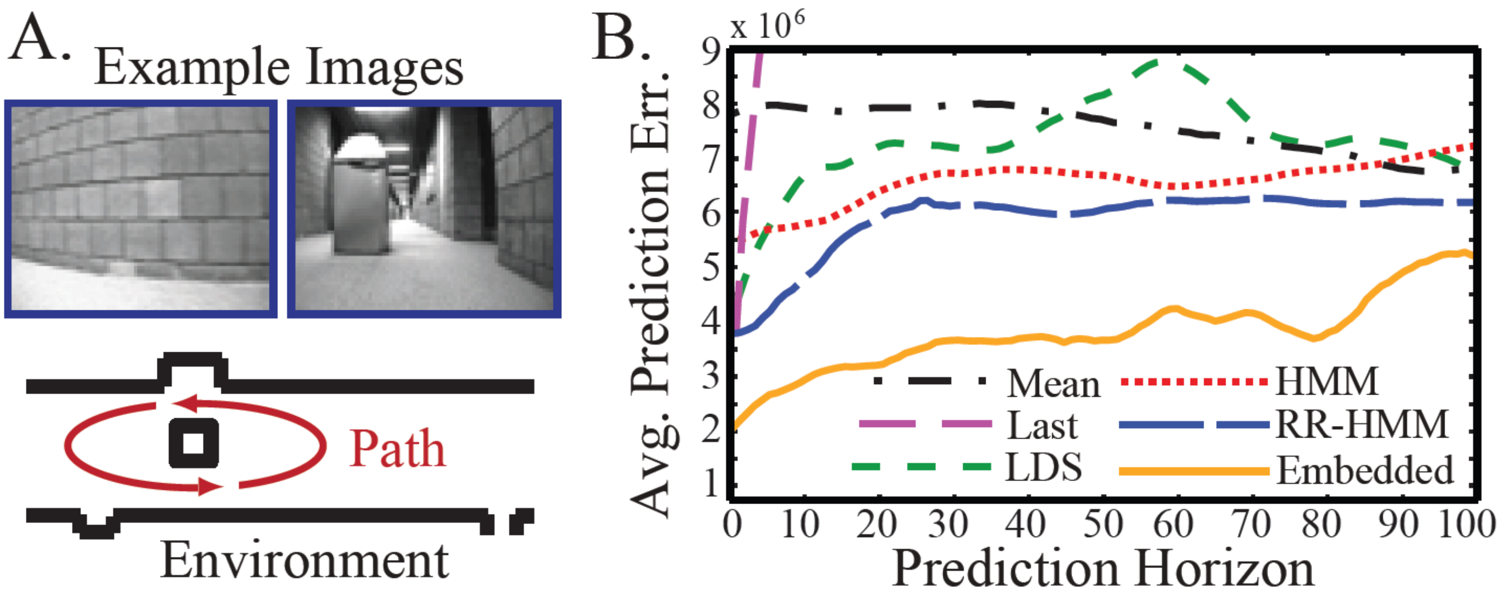
Use finite sample mean to approximate expectation,
Then apply the recursively low rank decomposition

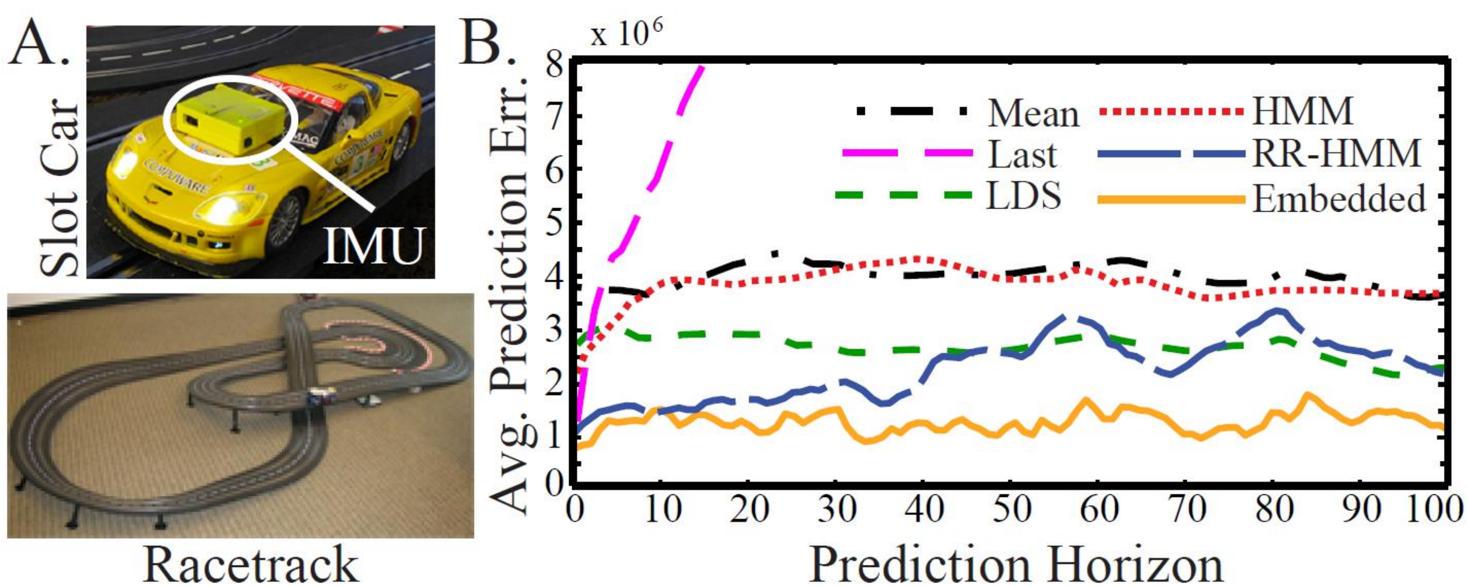
How to Deal with Infinite Features?

- Kernel trick: never explicitly compute features, always turn it into inner product $k(x, x') = \langle \phi(x), \phi(x') \rangle$

- Eg. kernel Singular Value Decomposition
  - $\widehat{\mathcal{C}}_{XY} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \otimes \phi(y_i) = U \Sigma V^\top$
  - Run kernel principal component analysis on $\widehat{\mathcal{C}}_{XY} \widehat{\mathcal{C}}_{XY}^\top$
  - Eigenvector lies the span of data $U = \sum_{i=1}^{m} \alpha_i \phi(x_i)$
  - Solve a generalized eigenvalue problem
    - $KGK\alpha = \lambda K \alpha$
    - Kernel matrix $K_{ij} = k(x_i, x_j)$ and $G_{ij} = k(y_i, y_j)$

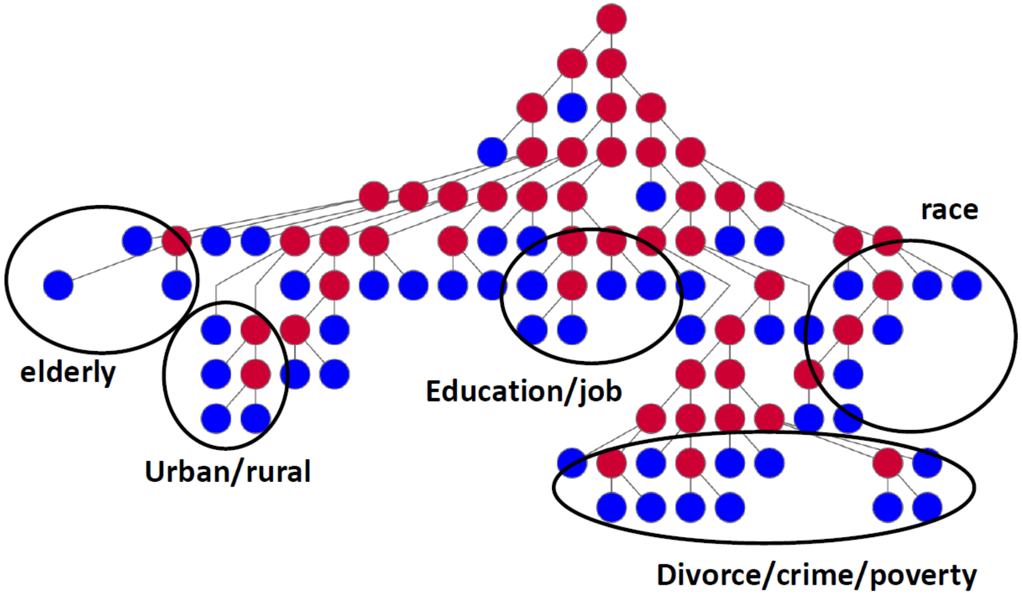
Video and Slot Car Senor Prediction

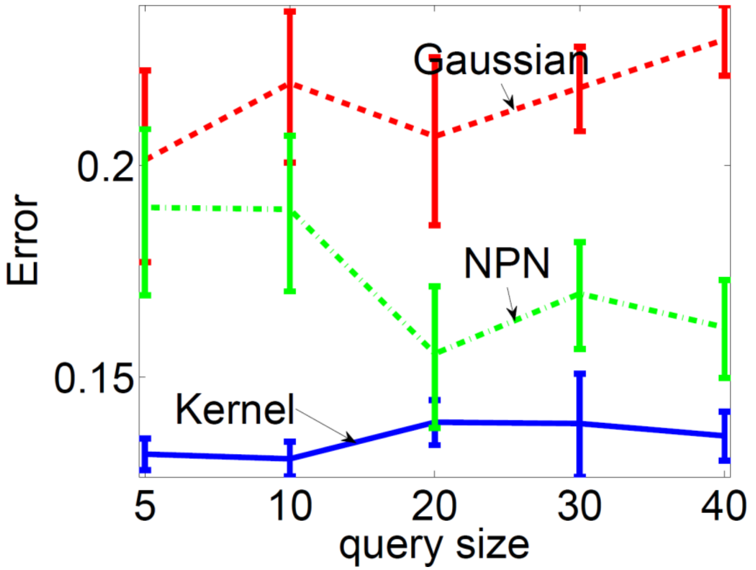




Demographic Feature Prediction

- 50 variables, 1400 samples, learn the latent structure first and then run spectral algorithms
- Compare to Gaussian latent variable model and Gaussian copula model (NPN), absolute error $|x - x^{\star}|$





Summary and Future direction (more)

- Spectral algorithm is the consequence of low rank structure of latent variable model
  - $P = (PA)(BPA)^{-1}(BP)$
  - Recursively decomposition
  - Better low rank approximation?

- What if the latent variable model is the wrong model?
  - Estimating latent parameters
    - PCA approach (Mossel & Roch AOAP'06), PCA and SVD approach, (Anandkumar et al. COLT'12, Arxiv)
  - Estimating the structure of latent variable models
    - Recursive grouping (Choi et al. JMLR'11), Spectral short quartet (Anandkumar et al. NIPS'11)

Questions?

- Thanks