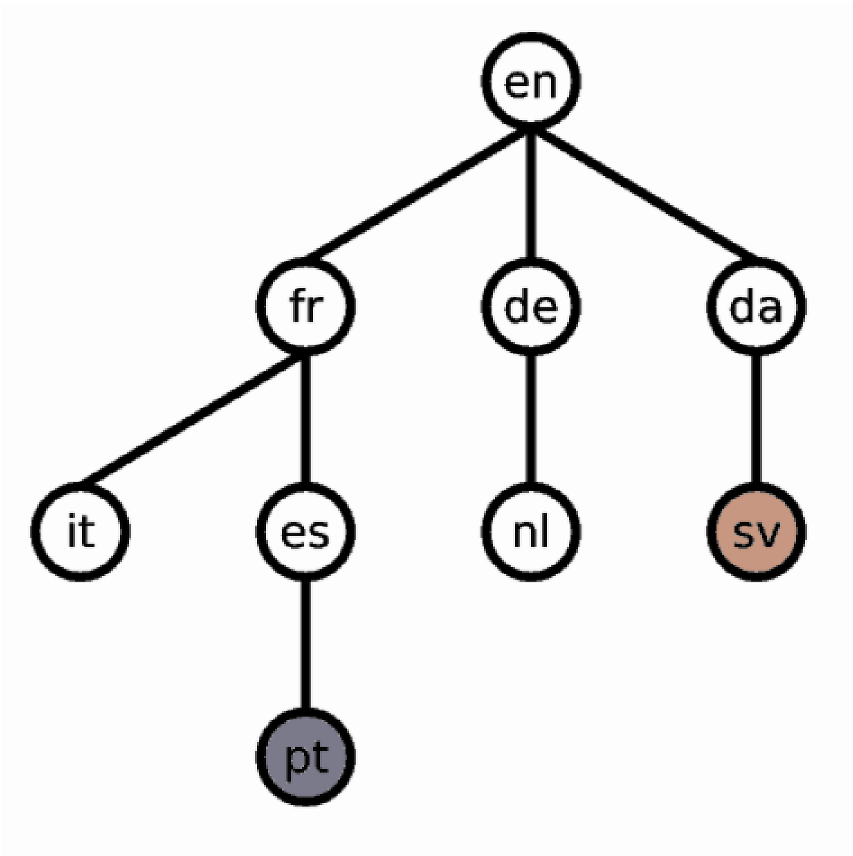
Bayesian Inference with Kernels

Arthur Gretton

Joint work with Danny Bickson, Carlos Guestrin, Yucheng Low, Le Song

Gatsby Computational Neuroscience Unit
Carnegie Mellon University
Max Planck Institute for Biological Cybernetics

A challenge: cross-language document retrieval



Cross-language document retrieval

- Many translations from "other" to English
- Few translations between unlike languages: Portuguese to Swedish

The problem: retrieve document in target language given document in source language, without examples of direct translation

Motivation and further applications

- Why use a non-parametric (kernel) algorithm?
  - Complex high-dimensional/structured data (discretization fails)
  - Non-Gaussian/multimodal (Gaussian BP fails)
  - Density estimation/integration too expensive (Parzen window approximations fail)
  - Model learned from training data

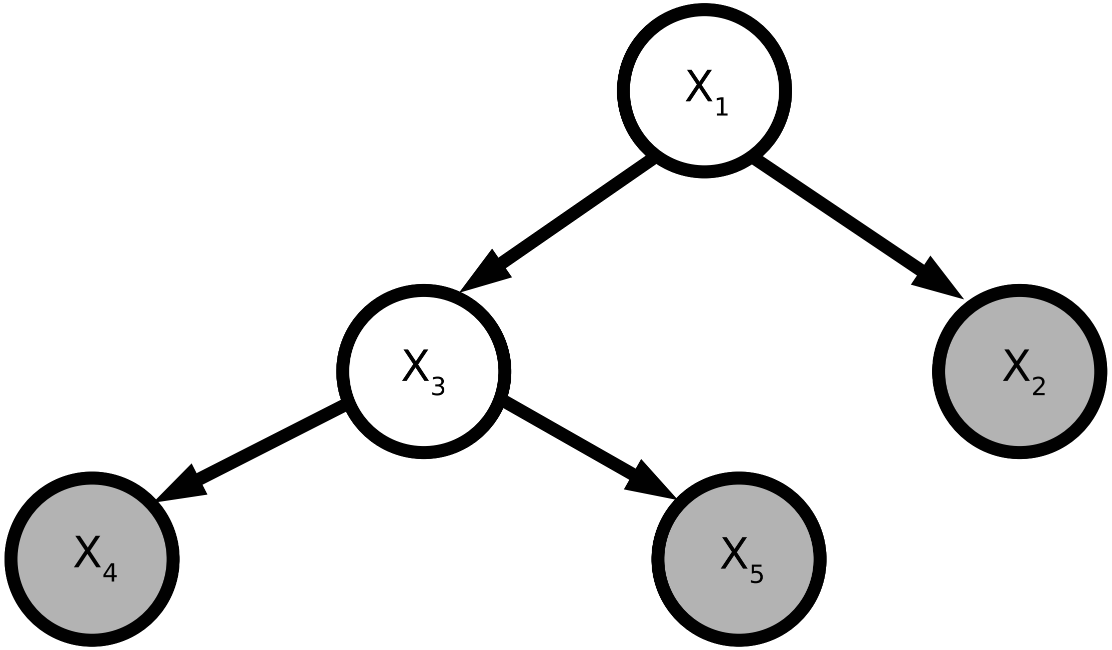
Motivation and further applications

- Why use a non-parametric (kernel) algorithm?
  - Complex high-dimensional/structured data (discretization fails)
  - Non-Gaussian/multimodal (Gaussian BP fails)
  - Density estimation/integration too expensive (Parzen window approximations fail)
  - Model learned from training data
- Exact inference on trees [Song, Gretton, and Guestrin, 2010b]
  - Cross-language document retrieval
  - Camera orientation recovery from images
- Loopy BP on pairwise MRFs [Song, Gretton, Bickson, Low, and Guestrin, 2010a]
  - Depth recovery from 2D images
  - Predicting paper categories from citation networks
  - Protein structure prediction

Motivation and further applications

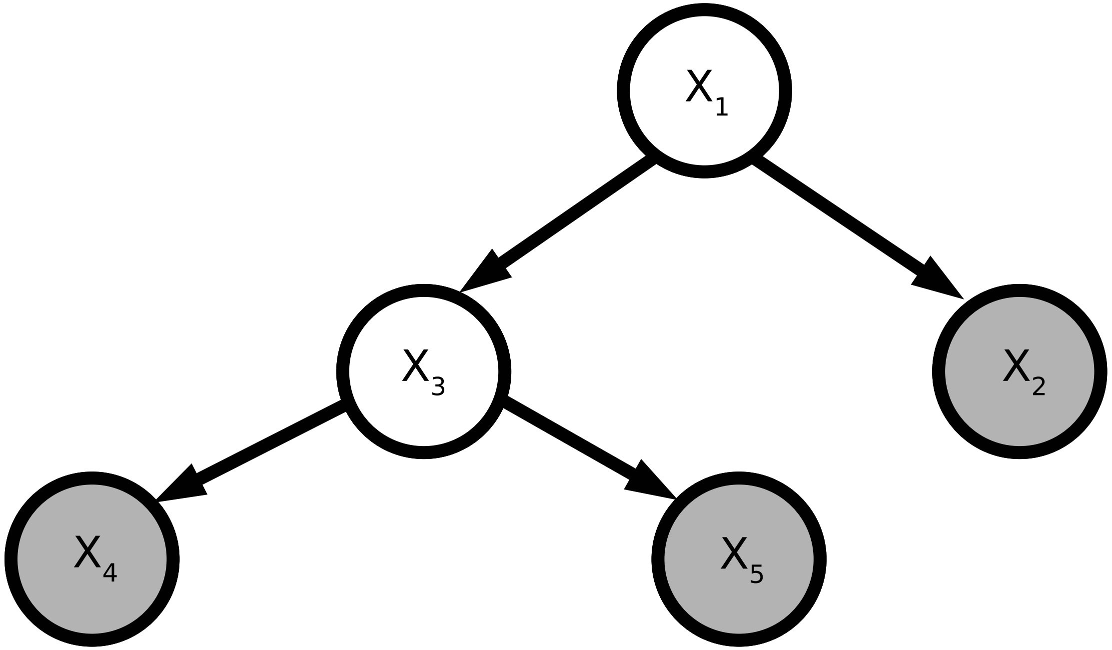
- Why use a non-parametric (kernel) algorithm?
  - Complex high-dimensional/structured data (discretization fails)
  - Non-Gaussian/multimodal (Gaussian BP fails)
  - Density estimation/integration too expensive (Parzen window approximations fail)
  - Model learned from training data
- Exact inference on trees [Song, Gretton, and Guestrin, 2010b]
  - Cross-language document retrieval
  - Camera orientation recovery from images
- Loopy BP on pairwise MRFs [Song, Gretton, Bickson, Low, and Guestrin, 2010a]
  - Depth recovery from 2D images
  - Predicting paper categories from citation networks
  - Protein structure prediction

Message passing on directed graphical models



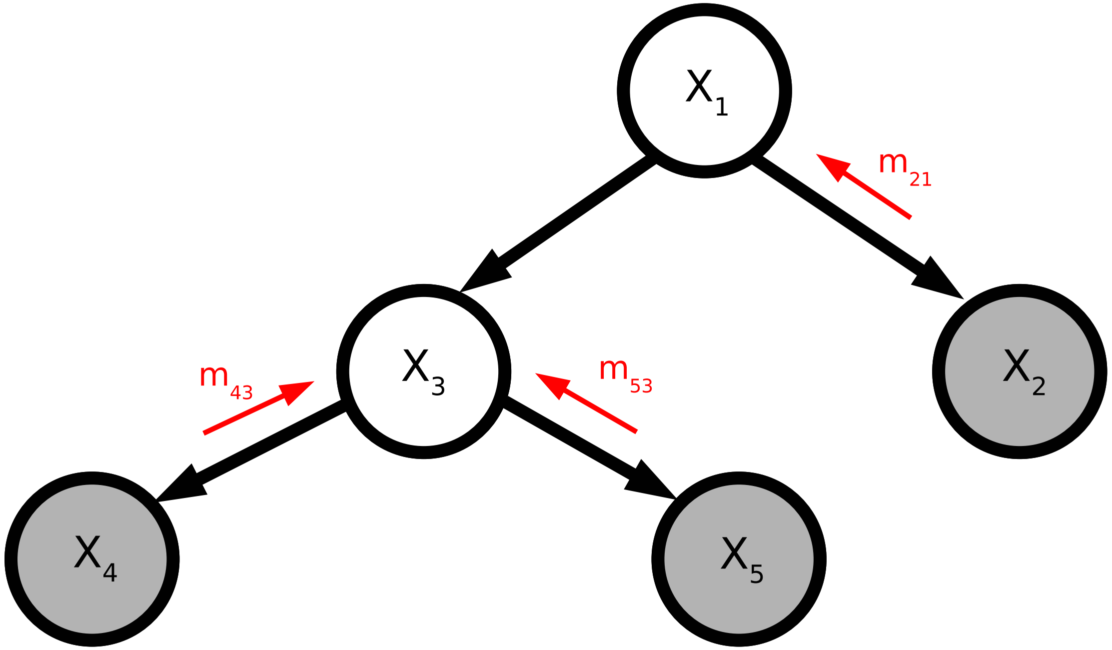
$$\mathbf{P}(X_1, x_2, x_4, x_5) = \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3)$$

Message passing on directed graphical models



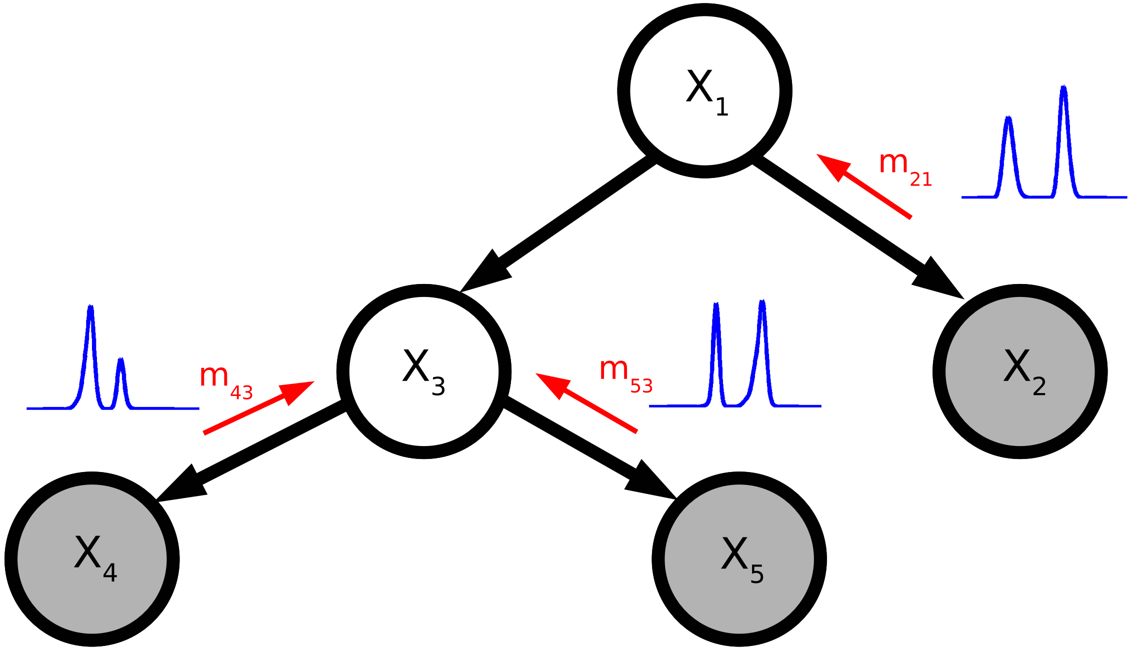
$$\begin{aligned}\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\ &= \mathbf{P}(X_1)\mathbf{P}(x_2|X_1) \int_{x_3} \mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3)\end{aligned}$$

Message passing on directed graphical models



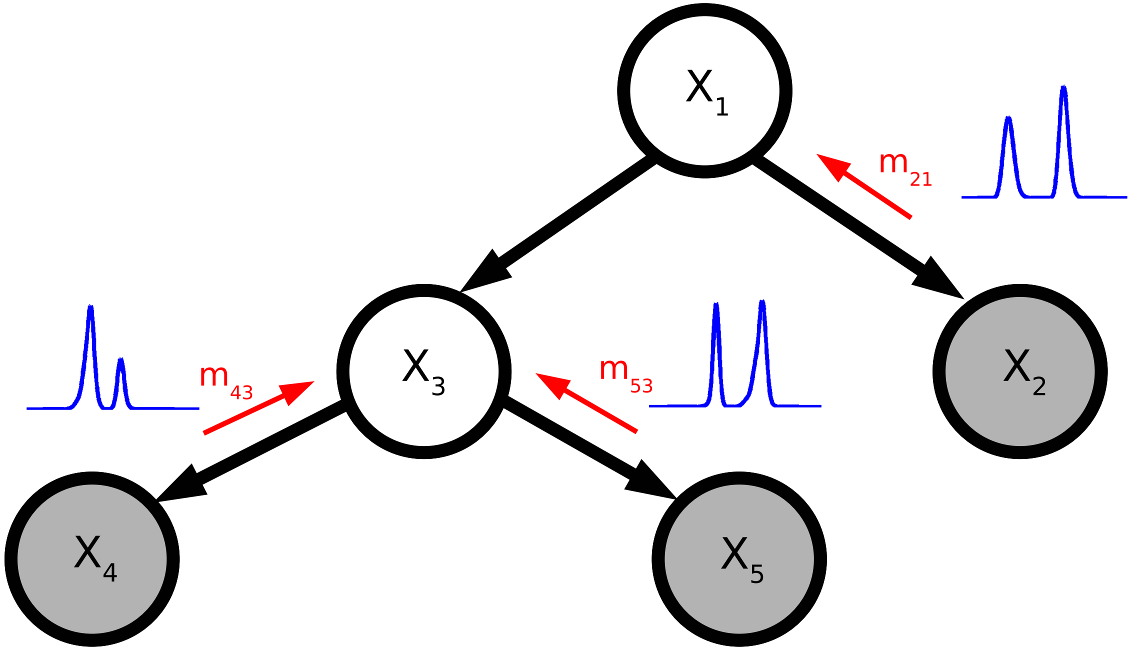
$$\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)\underbrace{\mathbf{P}(x_2|X_1)}_{m_{21}(X_1)}\int_{x_3} \mathbf{P}(X_3|X_1)\underbrace{\mathbf{P}(x_4|X_3)}_{m_{43}(X_3)}\underbrace{\mathbf{P}(x_5|X_3)}_{m_{53}(X_3)}
\end{aligned}$$

Message passing on directed graphical models



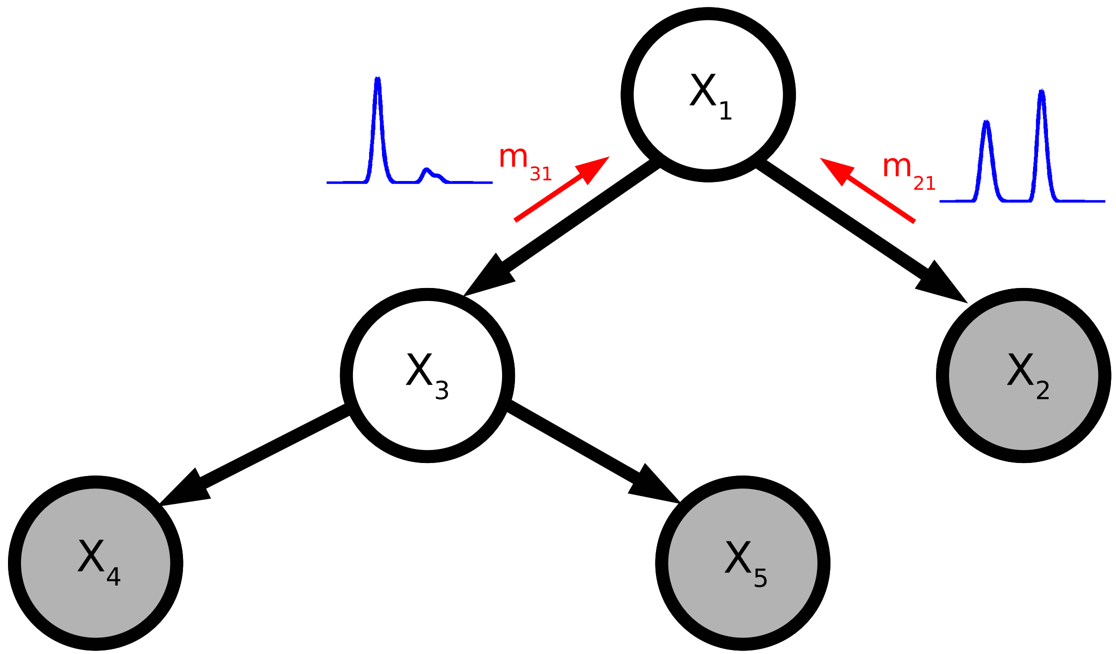
$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)\underbrace{\mathbf{P}(x_2|X_1)}_{m_{21}(X_1)} \int_{x_3} \mathbf{P}(X_3|X_1)\underbrace{\mathbf{P}(x_4|X_3)}_{m_{43}(X_3)}\underbrace{\mathbf{P}(x_5|X_3)}_{m_{53}(X_3)}
\end{aligned}
$$

Message passing on directed graphical models



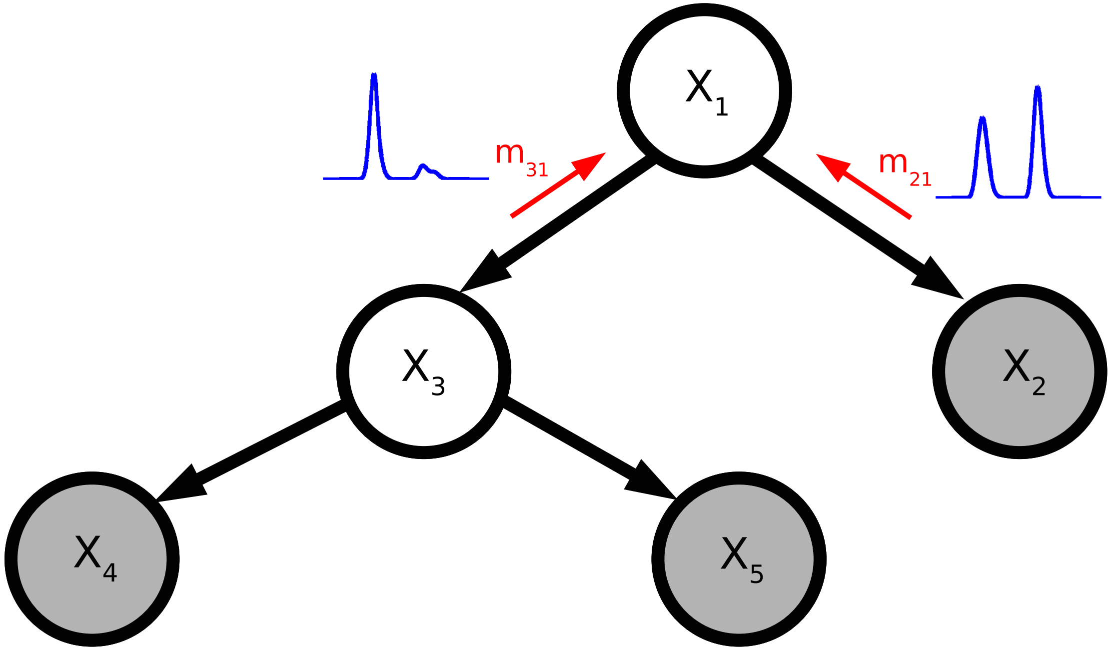
$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1){\color{red}m_{21}(X_1)} \int_{x_3} \mathbf{P}(X_3|X_1){\color{red}m_{43}(X_3)m_{53}(X_3)}
\end{aligned}
$$

Message passing on directed graphical models



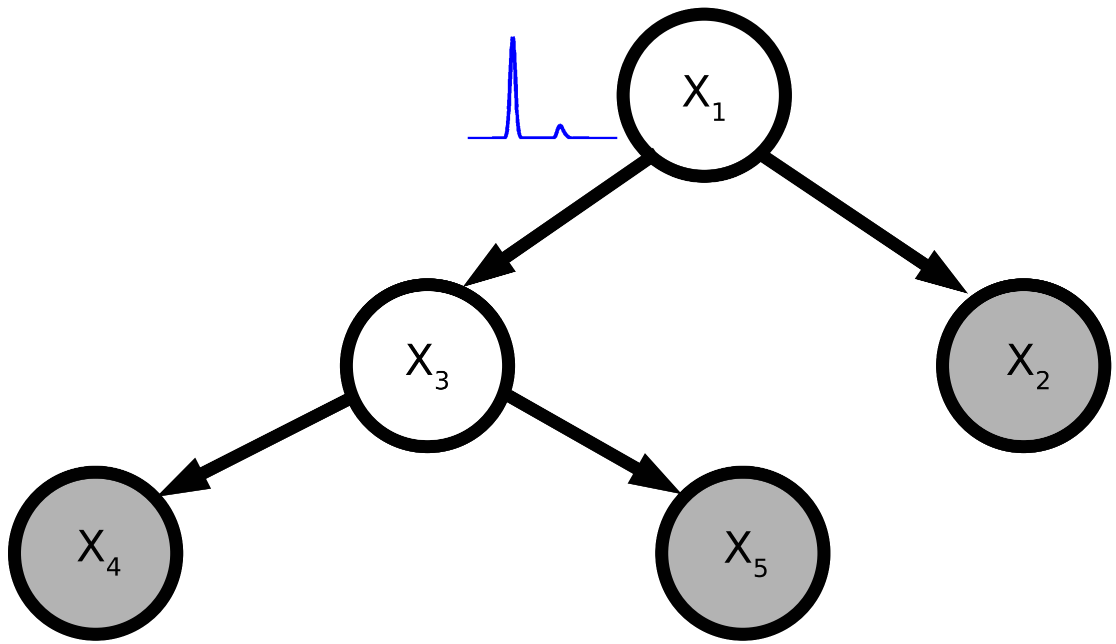
$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)\color{red}{m_{21}(X_1)}\color{black}\underbrace{\int_{x_3} \mathbf{P}(X_3|X_1)m_{43}(X_3)m_{53}(X_3)}_{\color{red}{m_{31}(X_1)}}
\end{aligned}
$$

Message passing on directed graphical models



$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)m_{21}(X_1)m_{31}(X_1)
\end{aligned}
$$

Message passing on directed graphical models



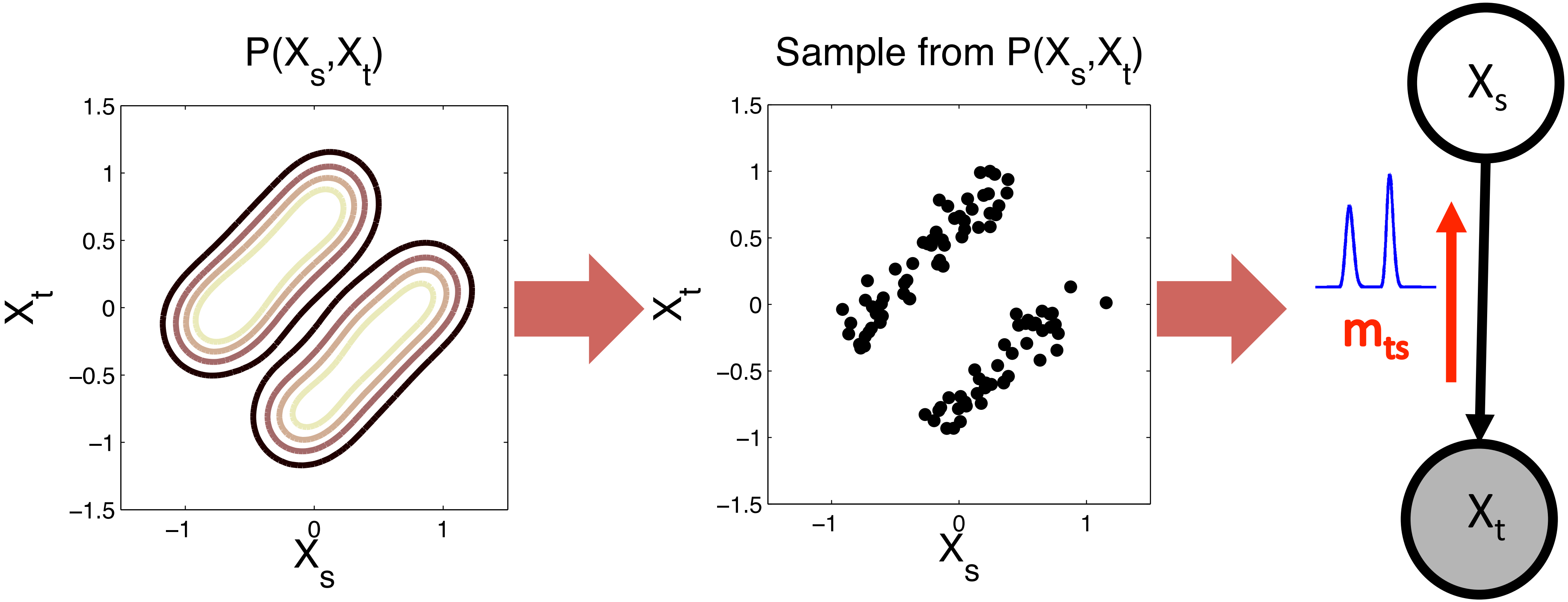
$$\mathbf{P}(X_1, x_2, x_4, x_5) = \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3)$$
$$= \mathbf{P}(X_1)m_{21}(X_1)m_{31}(X_1)$$

What's needed for learning and inference

- Learn the the messages from child nodes
  - Need to express conditional probabilities
- Combine evidence from multiple children
  - Need to marginalize

“Unusual” aspect: training phase

Model learned from training data



Conditional probabilities: gaussian case

- A hint: what would we do for the (zero mean) Gaussian?

$$p(z) \propto \left(-z^\top C^{-1} z\right),$$

- Partition

$$z = \begin{bmatrix} x \\ y \end{bmatrix} \qquad C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}.$$

- Conditional prob. of y given x:

$$\mathsf{P}(y|x) = \mathcal{N}(C_{yx}C_{xx}^{-1}x, C_{yy} - C_{yx}C_{xx}^{-1}C_{xy})$$

- Conditional expectation of y given x:

$$\begin{aligned} \mu_{y|x} &= C_{yx}C_{xx}^{-1}x \\ \mathbf{E}_{y|x}(a^\top y) &= a^\top \mu_{y|x} \end{aligned}$$

Conditional probabilities: Gaussian case

Complex functions linear in some feature space

- Nonlinear mean?

$$\begin{aligned} \mathbf{E}_X(a^\top X) &= a^\top \mu_X \\ \text{becomes} \quad \mathbf{E}_X f(X) &= \langle f, \mu_X \rangle_{\mathcal{F}} \end{aligned}$$

  in some feature space $\mathcal{F}$

- Nonlinear conditional mean?

$$\begin{aligned} \mathbf{E}_{y|x}(a^\top y) &= a^\top \mu_{y|x} &= a^\top C_{yx} C_{xx}^{-1} x \\ \text{becomes} \quad \mathbf{E}_{y|x} f(X) &= \langle f, \mu_{y|x} \rangle_{\mathcal{F}} &= ?? \end{aligned}$$

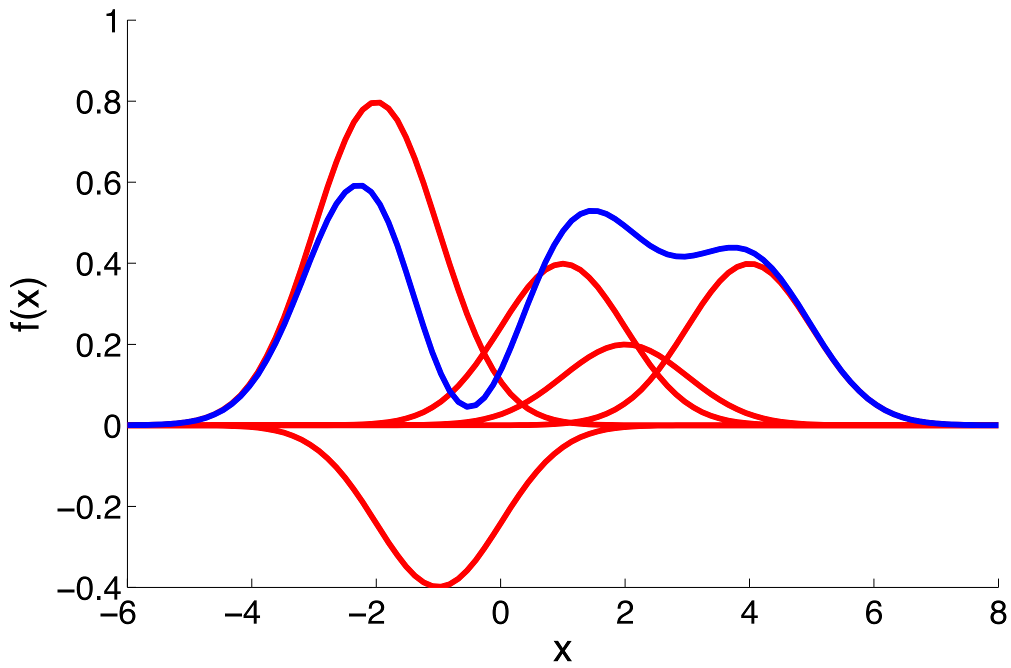
How do we do this with kernels?

Plan of attack

1. Kernelized mean
2. Kernelized covariance, leading to ...
3. ... kernel conditional mean
4. Messages from observed leaves (conditional probabilities)
5. Marginalize over internal node variables

RKHS definitions and properties

- $\mathcal{F}$ RKHS from $\mathcal{X}$ to $\mathbb{R}$ with positive definite kernel $k(x_i, x_j)$
- $\mathcal{F} = \overline{\text{span}\{k(x, \cdot) | x \in \mathcal{X}\}}$
  - Example: $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$ for arbitrary $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, $x_i \in \mathcal{X}$.



RKHS definitions and properties

- Riesz: unique representer of evaluation $\varphi_x \in \mathcal{F}$:

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}}$$

  - φ_x feature map

RKHS definitions and properties

- Riesz: unique representer of evaluation $\varphi_x \in \mathcal{F}$:

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}} = \langle f, k(x, \cdot) \rangle_{\mathcal{F}}$$

  – φ_x feature map

RKHS definitions and properties

- Riesz: unique representer of evaluation $\varphi_x \in \mathcal{F}$:

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}} = \langle f, k(x, \cdot) \rangle_{\mathcal{F}}$$

  – φ_x feature map

- Inner product between feature maps:

$$\langle \varphi_{x_1}, \varphi_{x_2} \rangle_{\mathcal{F}} = \langle k(x_1, \cdot), k(x_2, \cdot) \rangle_{\mathcal{F}} = k(x_1, x_2)$$

- Example: $f = \sum_{i=1}^{m} \alpha_i \varphi_{x_i}$

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}} = \left\langle \sum_{i=1}^{m} \alpha_i \varphi_{x_i}, \varphi_x \right\rangle_{\mathcal{F}} = \sum_{i=1}^{m} \alpha_i k(x_i, x)$$

Step 1: kernelized mean

Embedding of $\mathbf{P}_X$ to feature space

- $\mu_X \in \mathcal{F}$ such that $\forall f \in \mathcal{F}$,

$$\langle \mu_X, f \rangle = E_X f.$$

- What does mean embedding look like?

$$\begin{aligned}\mu_X(x) &= \langle \mu_X, \varphi_x \rangle \\ &= E_X k(X, x).\end{aligned}$$

Expectation of kernel!

- Empirical estimate:

$$\hat{\mu}_X(x) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, x) \qquad x_i \sim \mathbf{P}_X$$

Step 1: kernelized mean

Embedding of $\mathbf{P}_X$ to feature space

- $\mu_X \in \mathcal{F}$ such that $\forall f \in \mathcal{F}$,

$$\langle \mu_X, f \rangle = E_X f.$$

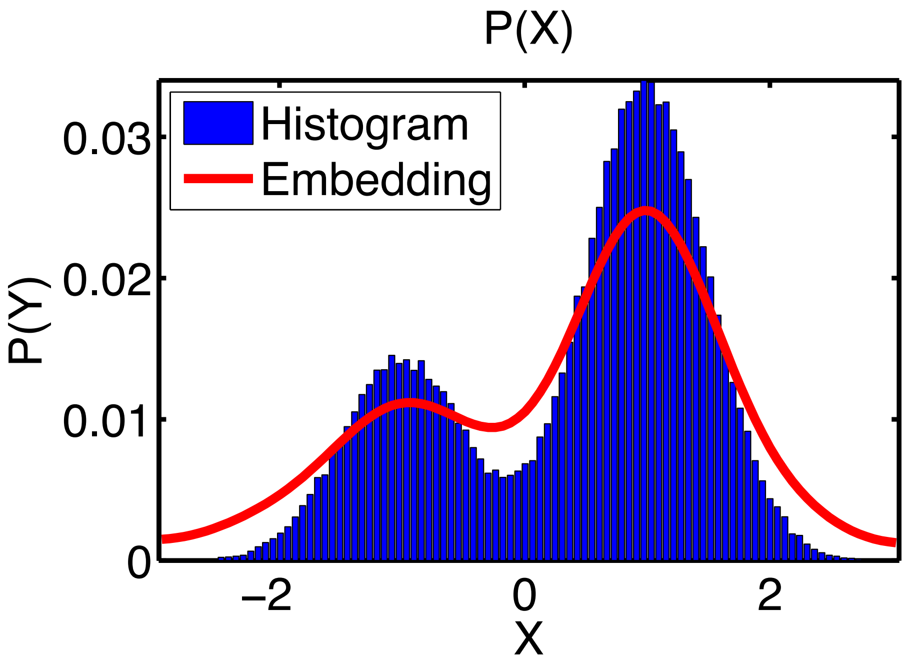
- What does mean embedding look like?

$$\begin{aligned} \mu_X(x) &= \langle \mu_X, \varphi_x \rangle \\ &= E_X k(X, x). \end{aligned}$$

Expectation of kernel!

- Empirical estimate:

$$\hat{\mu}_X(x) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, x) \qquad x_i \sim \mathbf{P}_X$$



Step 2: kernelized covariance

... in finite space

- Given $f \in \mathbb{R}^d$ and $g \in \mathbb{R}^{d'}$
- Define outer product

$$fg^\top$$

- Given $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^{d'}$,

$$(fg^\top)v = (g^\top v)f$$

and

$$\begin{aligned}\left\langle fg^\top, uv^\top \right\rangle &= \operatorname{tr}\left((fg^\top)^\top(uv^\top)\right) \\ &= (f^\top u)(g^\top v)\end{aligned}$$

... in kernel space

- Given $f \in \mathcal{F}$ and $g \in \mathcal{G}$
- Define tensor product space

$$f \otimes g \in \mathcal{F} \otimes \mathcal{G}$$

- $f \otimes g$ operator mapping $\mathcal{G} \to \mathcal{F}$: given any $v \in \mathcal{G}$,

$$f \otimes g\,(v) = \langle g, v \rangle\, f$$

- Inner product in $\mathcal{F} \otimes \mathcal{G}$:

$$\langle f \otimes g, u \otimes v \rangle_{\mathcal{F} \otimes \mathcal{G}} = \langle f, u \rangle \langle g, v \rangle$$

Step 2: kernelized covariance

- Covariance between $f \in \mathcal{F}$ and $g \in \mathcal{G}$ (uncentred)

$$\operatorname{cov}(f, g) = E_{XY}(fg)$$

- Covariance operator: mapping from $\mathcal{F} \otimes \mathcal{G} \to \mathbb{R}$.

$$\begin{aligned}
E_{XY}\, fg &= E_{XY}\ \langle f, \varphi_X \rangle \langle g, \phi_Y \rangle \\
&= E_{XY}\ \langle f \otimes g, \varphi_X \otimes \phi_Y \rangle_{\mathcal{F} \otimes \mathcal{G}} \\
&= \langle f \otimes g, E_{XY} \varphi_X \otimes \phi_Y \rangle_{\mathcal{F} \otimes \mathcal{G}} \\
&= \langle f \otimes g, C_{XY} \rangle_{\mathcal{F} \otimes \mathcal{G}} \\
&= \langle f, C_{XY}\, g \rangle_{\mathcal{F}}
\end{aligned}$$

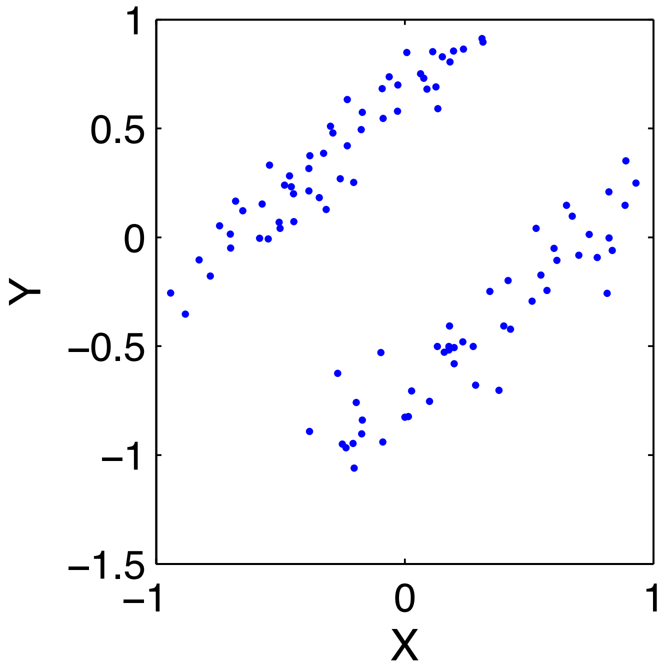
- Empirical estimate:

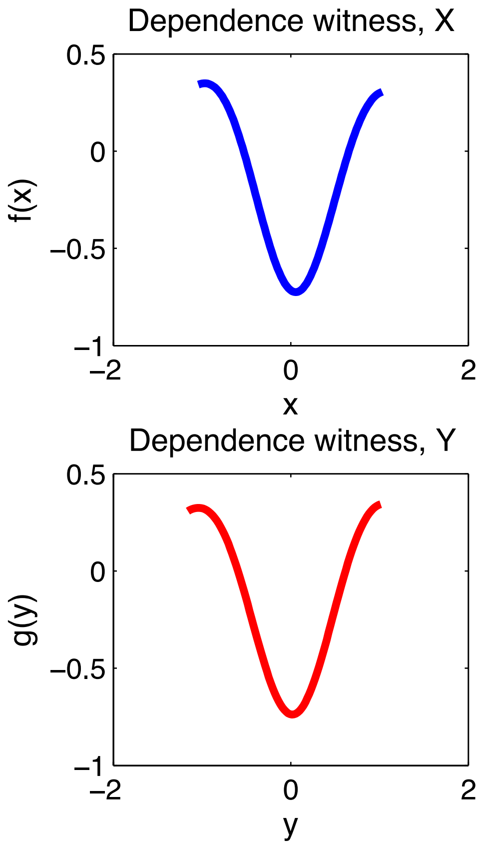
$$\widehat{C}_{XY} := \frac{1}{m} \sum_{i=1}^{m} \varphi_{x_i} \otimes \phi_{y_i} \qquad (x_i, y_i) \sim \mathbf{P}_{XY}$$

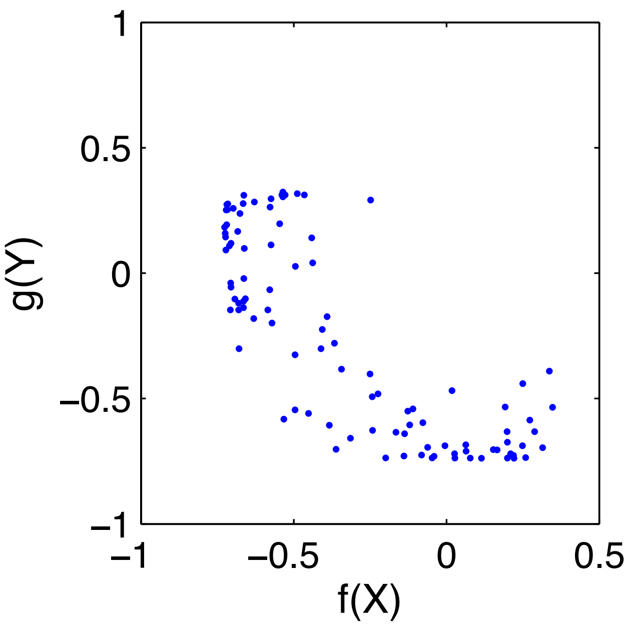
Step 2: kernelized covariance

First singular value of C_{xy}:

$$\sup_{\|f\|\leq 1, \|g\|\leq 1} \langle f, C_{xy} g \rangle_{\mathcal{F}} = \sup_{\|f\|\leq 1, \|g\|\leq 1} \mathrm{cov}(f, g)$$

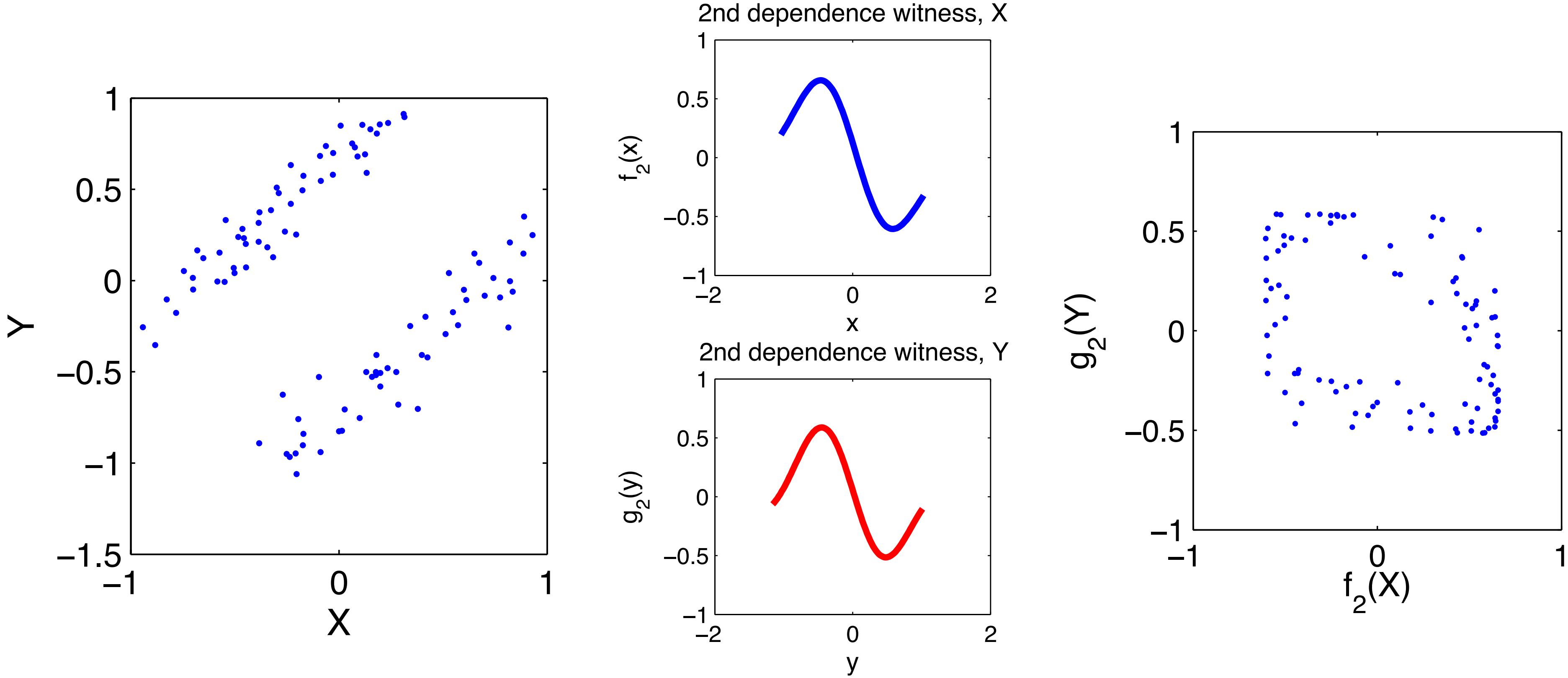






Step 2: kernelized covariance

Second singular value of C_{xy}:

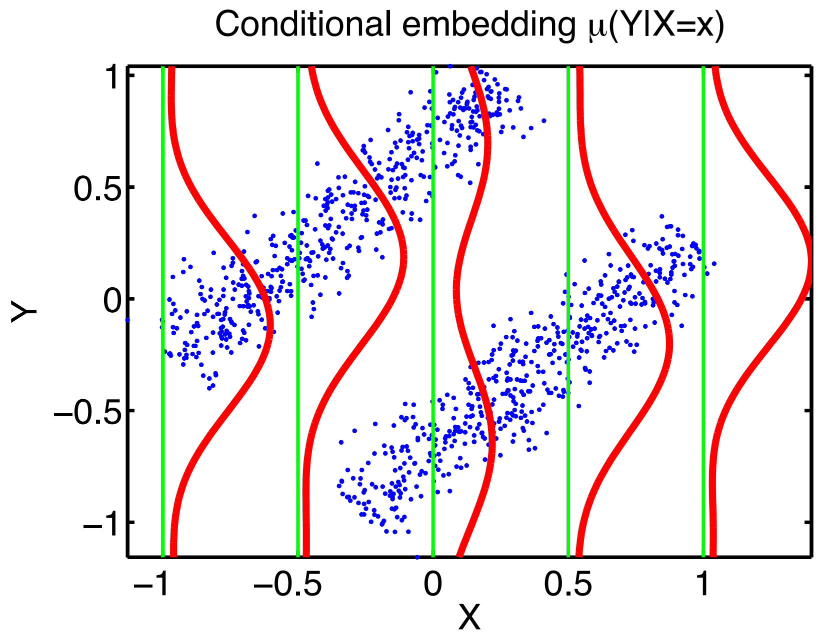


Step 3: kernelized conditional mean

- Conditional mean embedding,

$$\langle g, \mu_{Y|X=x} \rangle = E_{Y|X=x} g(Y)$$
$$\mu_{Y|X=x} := C_{YX} C_{XX}^{-1} \varphi_x$$

[Song et al., 2009]



- Reminder: Gaussian case

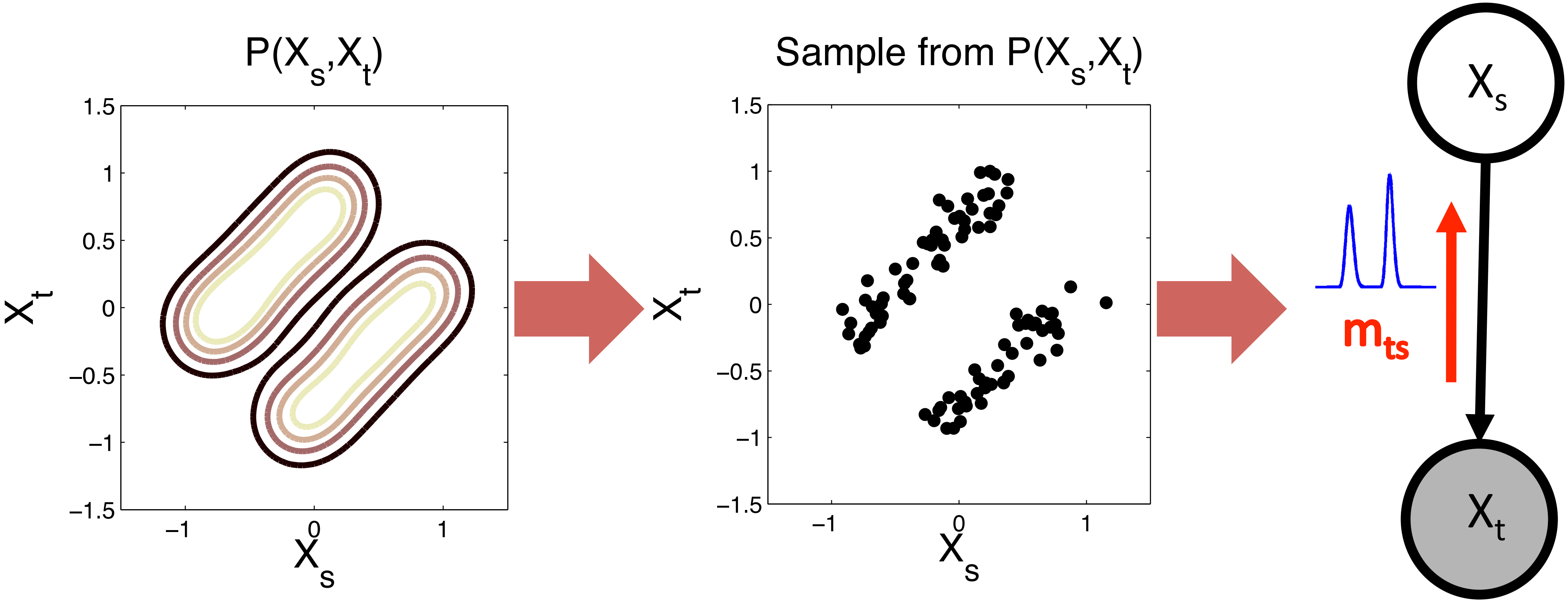
$$\mu_{Y|x} = C_{YX} C_{XX}^{-1} x$$

- Function is conditional expectation of kernel:

$$\mu_{Y|X=x}(y) = \langle \mu_{Y|X=x}, \phi_y \rangle = \mathbf{E}_{Y|x} k(Y, y)$$

Messages from observed leaves

- Goal: given leaf evidence x_t and parent X_S, want $m_{ts} := \mathbf{P}(x_t|X_s)$



Messages from observed leaves

- Goal: given leaf evidence x_t and parent X_S, want $m_{ts} := \mathbf{P}(x_t|X_s)$

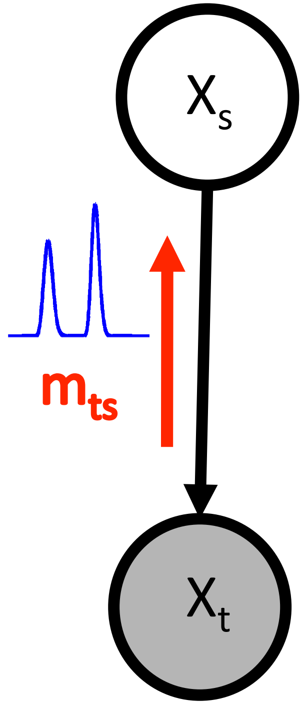
- Training data

$$(x_{s,1}, x_{t,1}), \ldots, (x_{s,m}, x_{t,m})$$

- Empirical leaf messages $m_{ts}(X_S)$

$$\begin{aligned} m_{ts}(X_s) &= \mathbf{P}(x_t|X_s) \\ &= \sum_{i=1}^{m} \beta_{ts,i} k(x_{s,i}, X_s) \end{aligned}$$

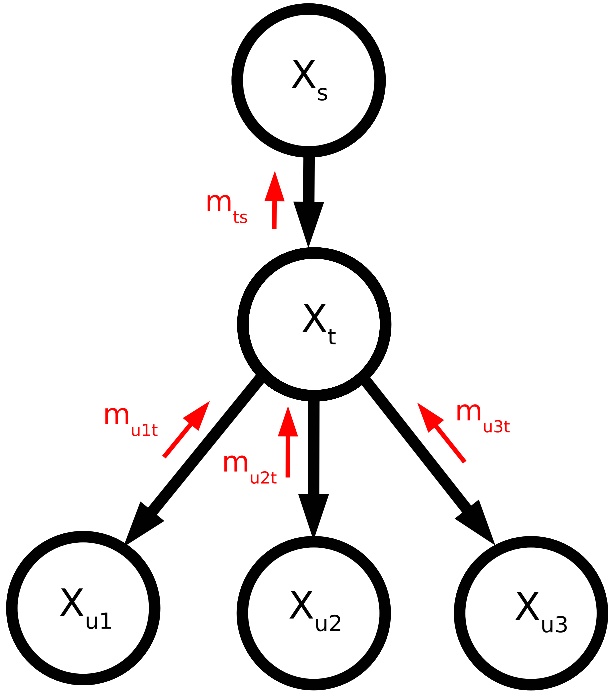
$$\beta_{ts} = ((K_t + \lambda I)(K_s + \lambda I))^{-1} k_t$$



Marginalize over internal nodes

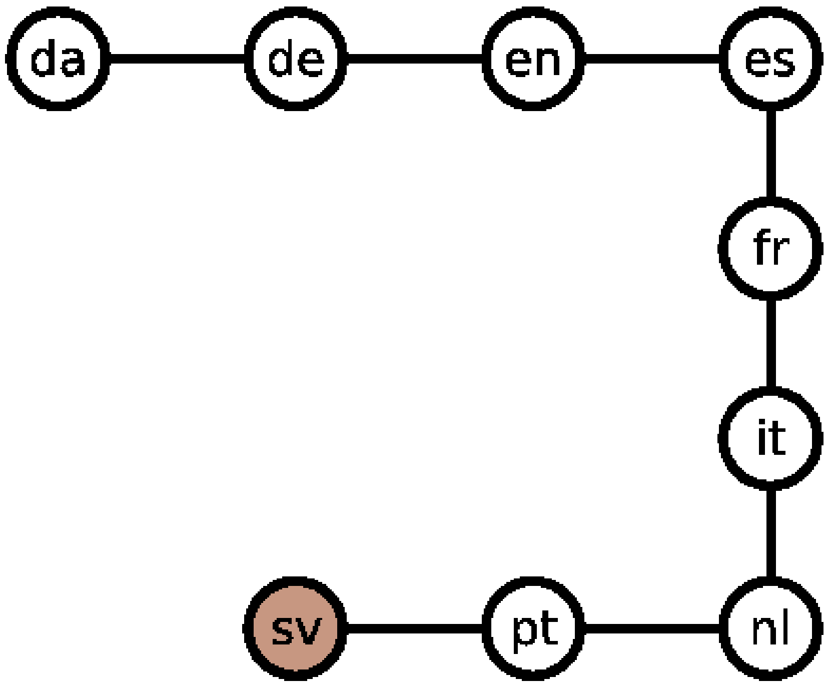
- Marginalize over X_t:

$$
\begin{aligned}
m_{ts}(X_s) &= \sum_{i=1}^{m} \beta_{ts,i} k(x_{s,i}, X_s) \\
\beta_{ts} &= (K_s + \lambda I)^{-1} \bigodot_{u \in \Gamma_t \backslash s} K_t^{(u)} \beta_{ut}
\end{aligned}
$$



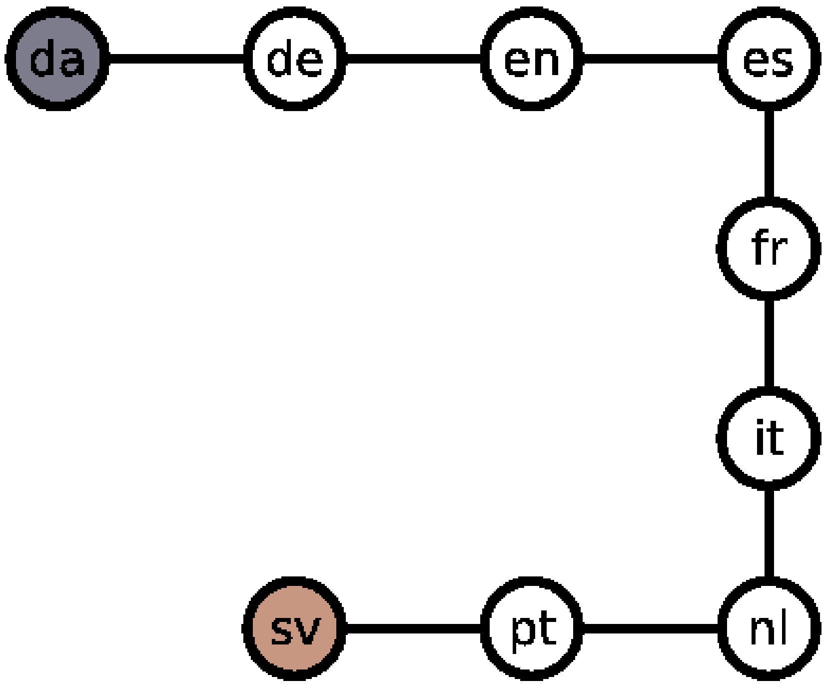
- Advantages:
  - Cost increase not exponential in depth unlike Gaussian Mixture Models (GMM) [Sudderth et al., 2003]
  - Nonparametric model learned from data unlike GMM, Particle BP [Sudderth et al., 2003, Ihler and McAllester, 2009]

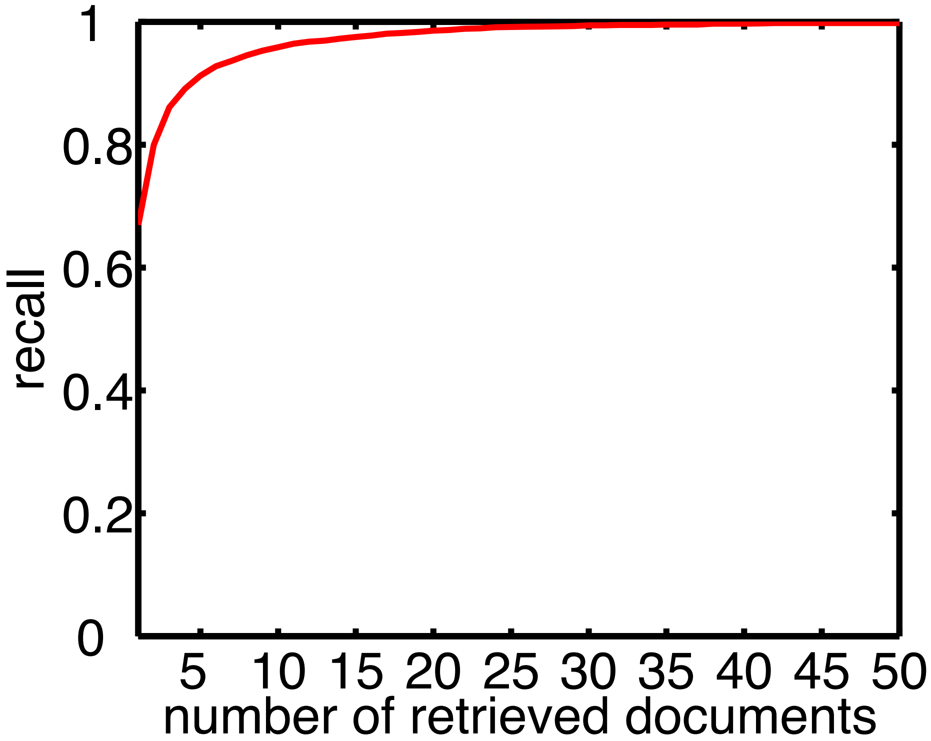
Cross-language document retrieval



- Experiment from [Song, Gretton, and Guestrin, 2010b]
- Source document one of Danish, German, English,. . .
- Target document Swedish
- Data: 300 documents from European Parliament transcripts [Koehn, 2005]

Cross-language document retrieval

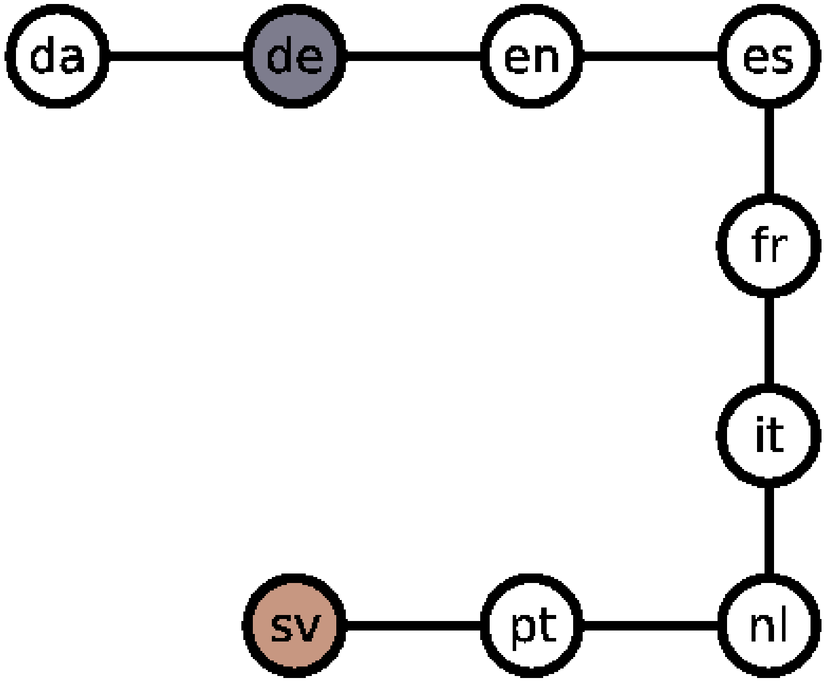


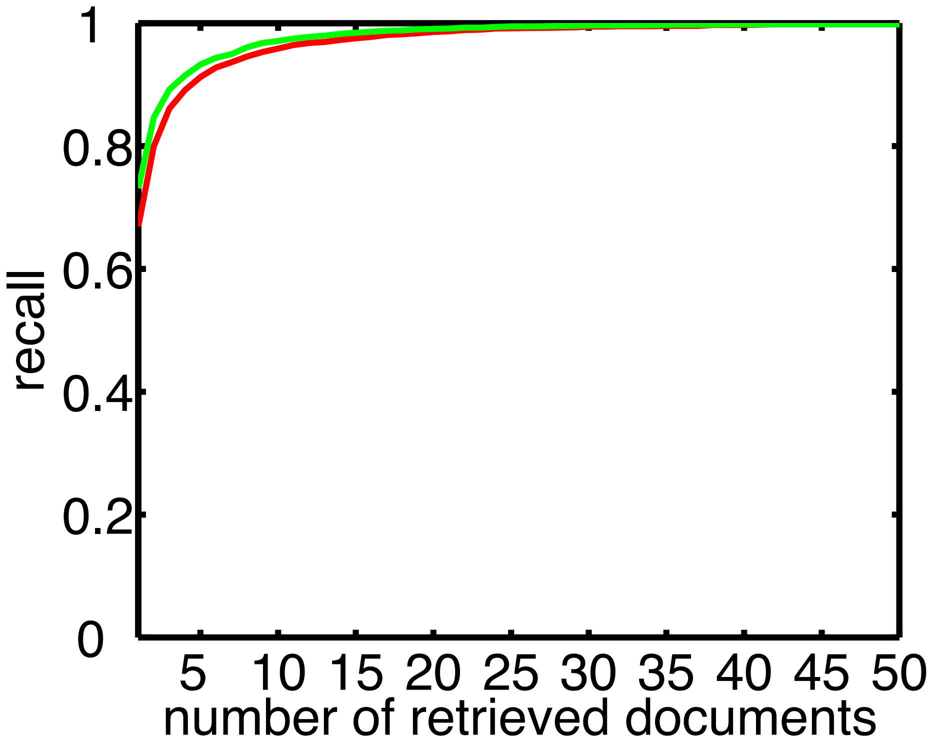


Recall score: whether target document is in set of retrieved documents

Details: TF-IDF document features, stopword removal and stemming, Gaussian RBF kernel, bandwidth at median distance between feature vectors.

Cross-language document retrieval

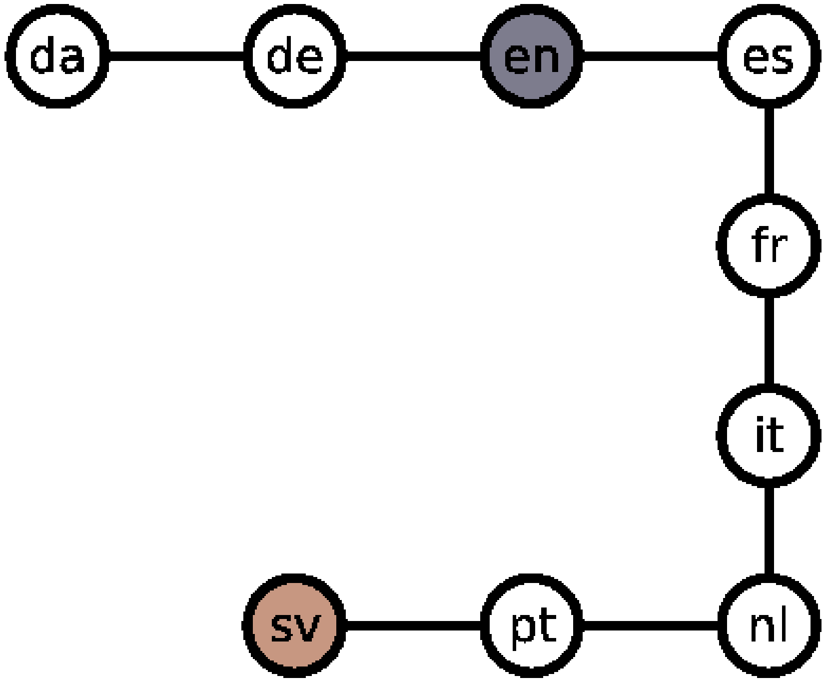


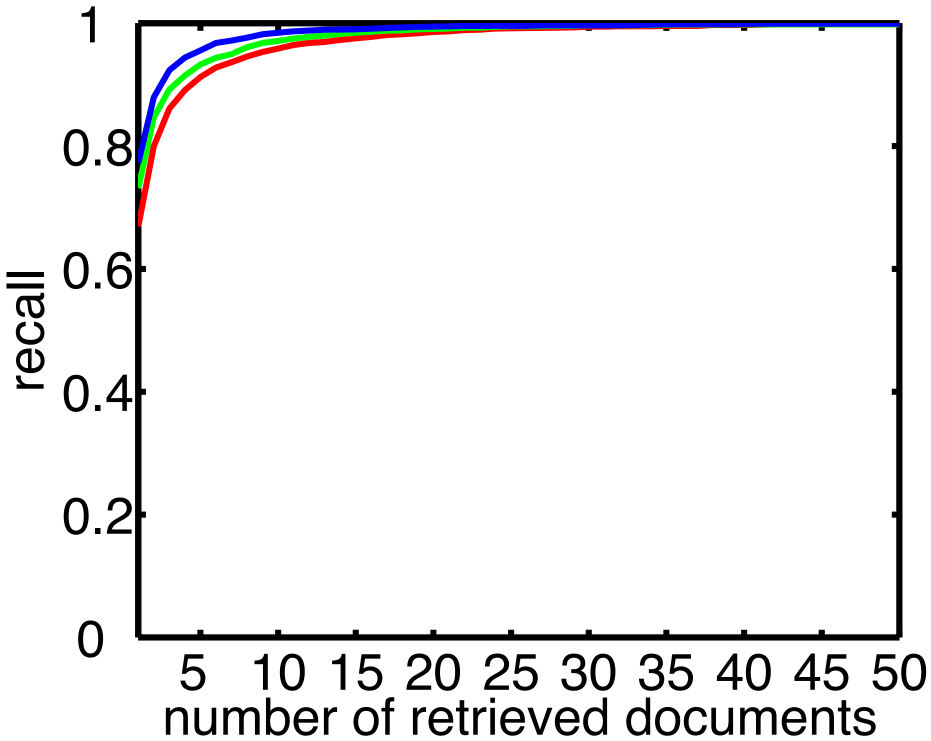


Recall score: whether target document is in set of retrieved documents

Details: TF-IDF document features, stopword removal and stemming, Gaussian RBF kernel, bandwidth at median distance between feature vectors.

Cross-language document retrieval

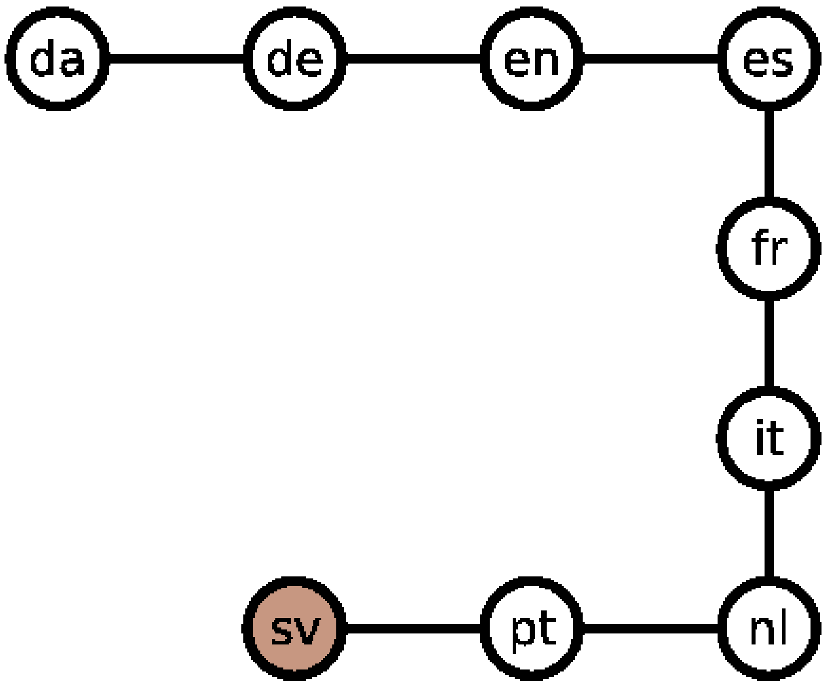


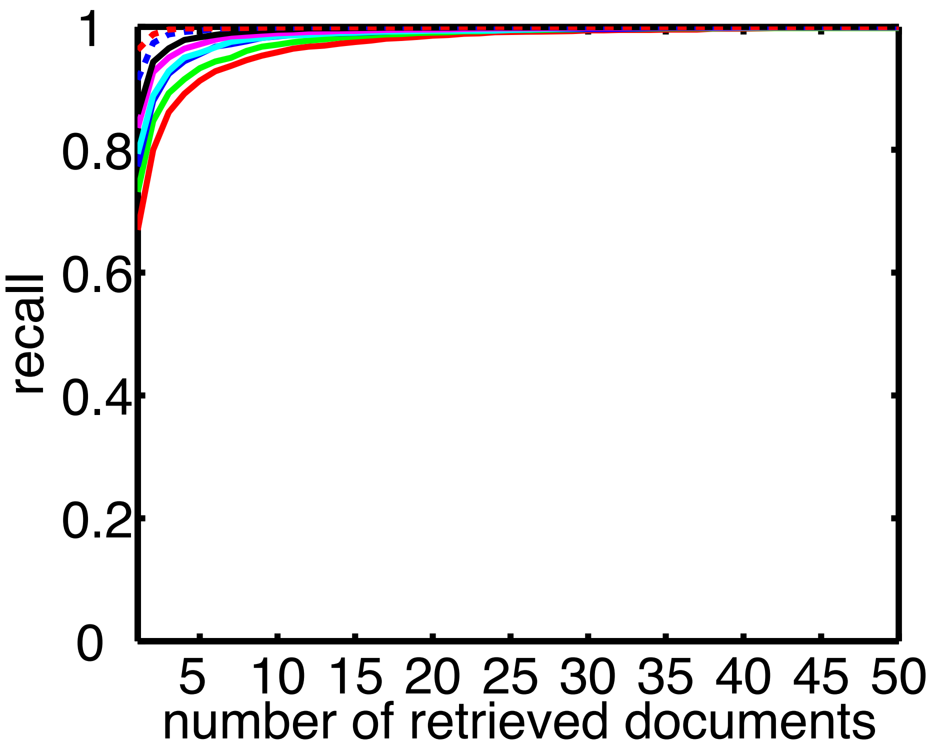


Recall score: whether target document is in set of retrieved documents

Details: TF-IDF document features, stopword removal and stemming, Gaussian RBF kernel, bandwidth at median distance between feature vectors.

Cross-language document retrieval

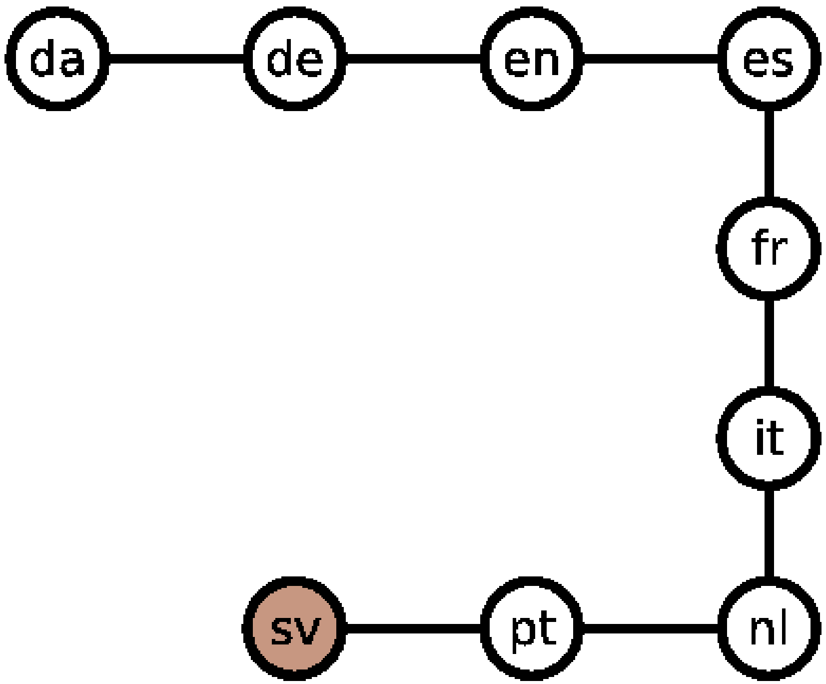


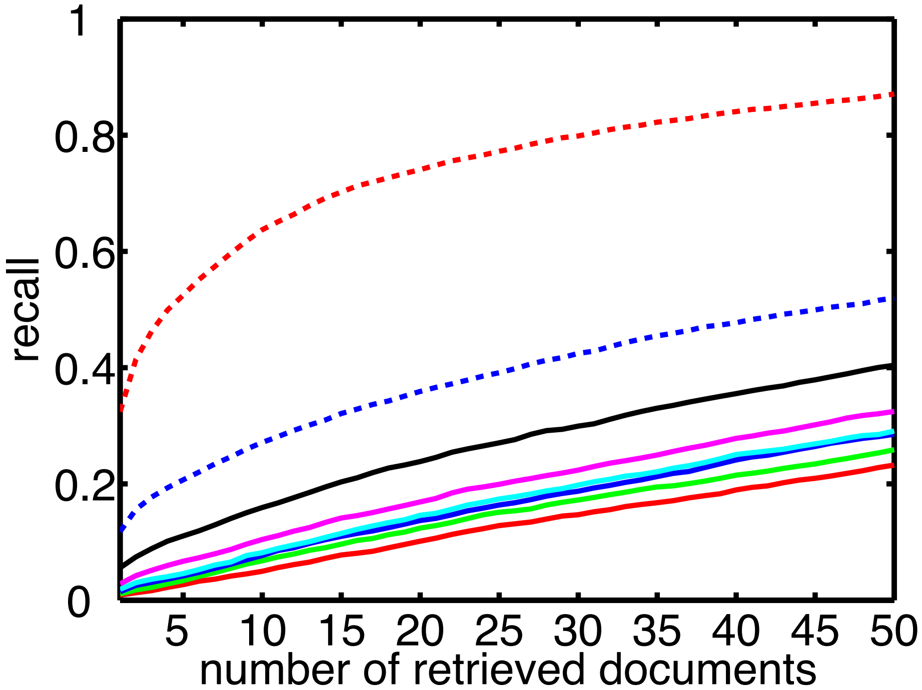


Recall score: whether target document is in set of retrieved documents

Details: TF-IDF document features, stopword removal and stemming, Gaussian RBF kernel, bandwidth at median distance between feature vectors.

Cross-language document retrieval

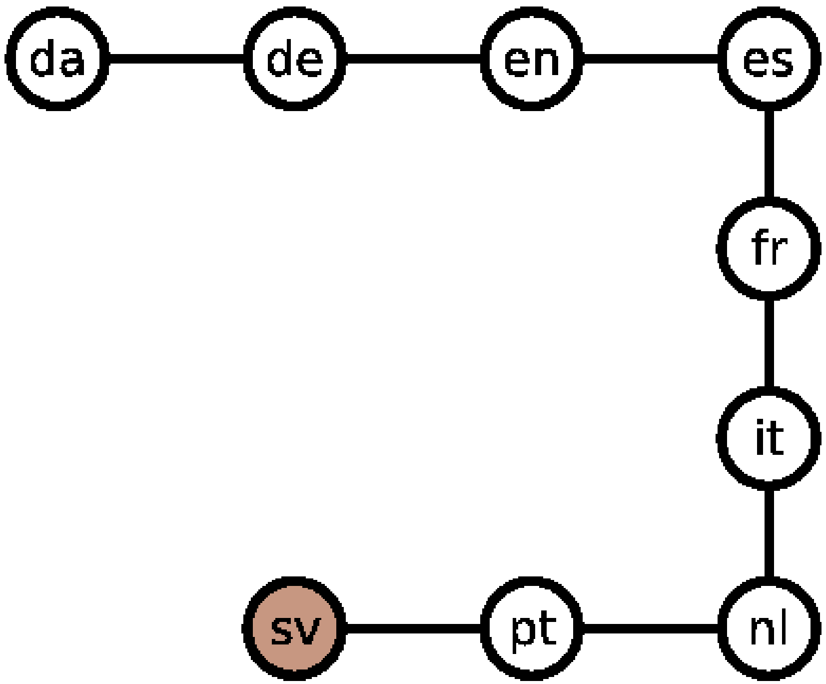


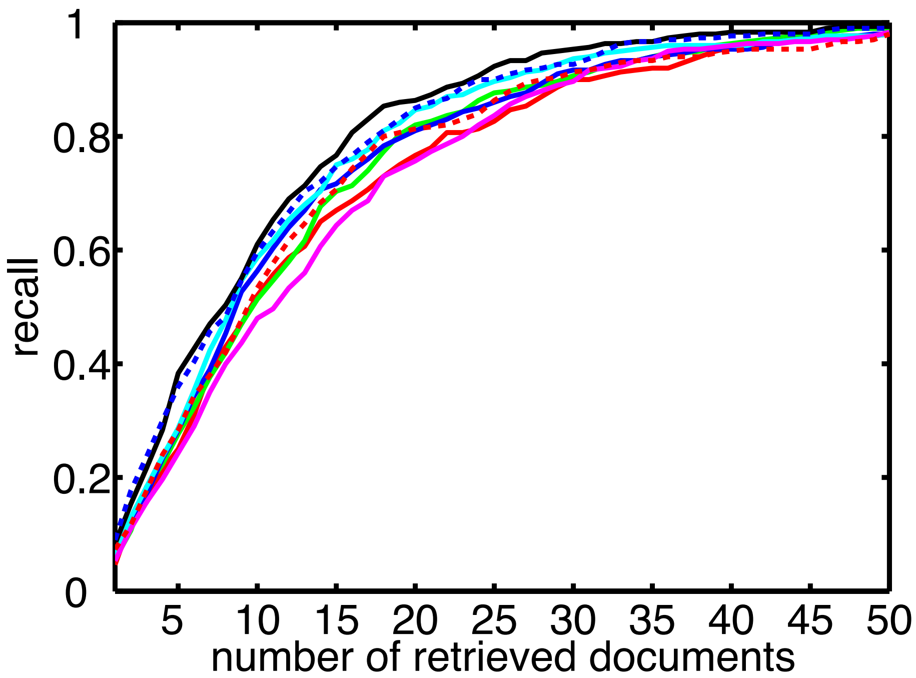


Recall score: whether target document is in set of retrieved documents

- Bilingual topic model with 50 topics for each edge [Mimno et al., 2009]
- Compare topic distribution of query in target domain with topic distributions of all target documents

Cross-language document retrieval



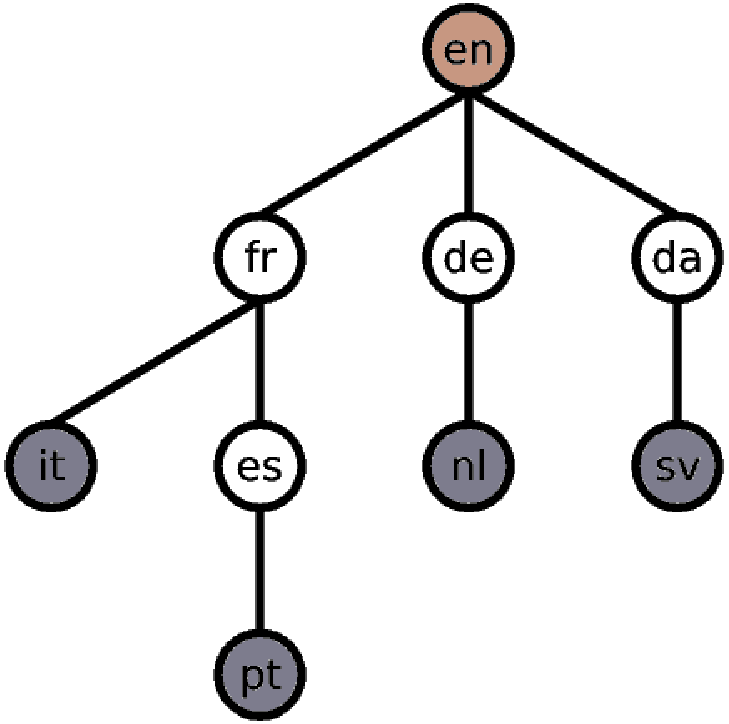


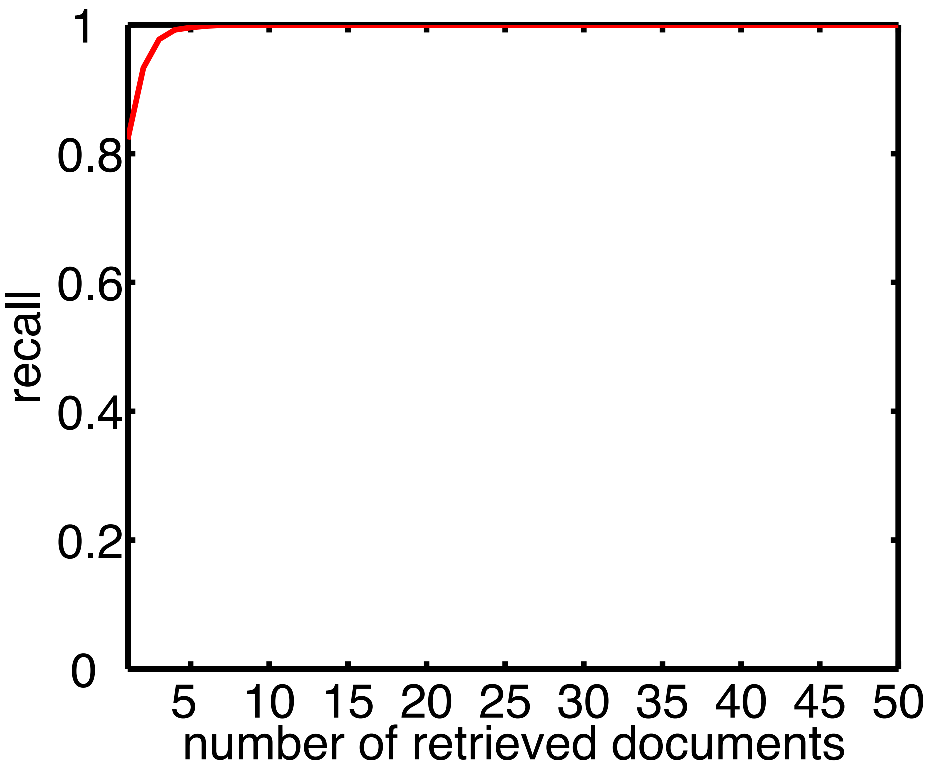
Recall score: whether target document is in set of retrieved documents

Normalized document length [Gale and Church, 1991]

- Chain length irrelevant

Cross-language document retrieval





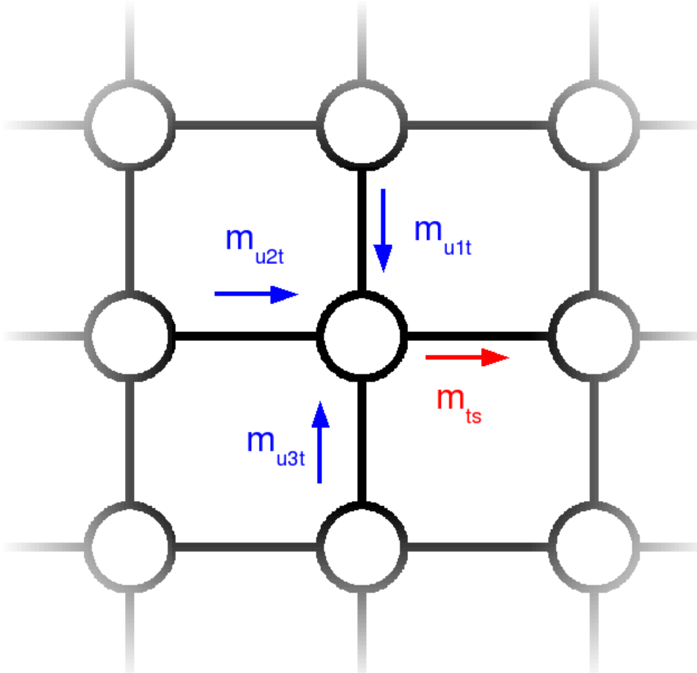
Nonparametric tree graphical model,
evidence at multiple leaves

Loopy belief propagation

- Pairwise MRF

$$\mathbf{P}(X) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \Psi_{st}(X_s, X_t) \prod_{s\in\mathcal{V}} \Psi_s(X_s),$$

- $\Psi_s(X_s)$ node potentials, $\Psi_{st}(X_s, X_t)$ edge potentials, and Z normalization.



- Loopy BP [Yedidia et al., 2001]:
  Iterate

$$m_{ts}(X_s) = \int_{X_t} \Psi_{st}(X_s, X_t)\Psi_t(X_t) \prod_{u\in\Gamma_t\backslash s} m_{ut}(X_t)\, dX_t$$

Locally consistent BP

- Locally consistent BP [Wainwright et al., 2003]

$$\Psi_s(X_s) = \mathbf{P}(X_s), \qquad \Psi(X_s, X_t) = \mathbf{P}(X_s, X_t)\mathbf{P}(X_t)^{-1}\mathbf{P}(X_t)^{-1},$$

$\mathbf{P}(X_s)$ and $\mathbf{P}(X_s, X_t)$ empirical distributions

Locally consistent BP

- Locally consistent BP [Wainwright et al., 2003]

$$\Psi_s(X_s) = \mathbf{P}(X_s), \qquad \Psi(X_s, X_t) = \mathbf{P}(X_s, X_t)\mathbf{P}(X_t)^{-1}\mathbf{P}(X_t)^{-1},$$

$\mathbf{P}(X_s)$ and $\mathbf{P}(X_s, X_t)$ empirical distributions

- Fixed point, $\mathbf{P}(X_s)$ and $\mathbf{P}(X_s, X_t)$, at empirical marginals,

$$\mathbf{P}(X_s) = \mathbf{P}(X_s) \prod_{u \in \Gamma_s} m_{us}(X_s),$$

$$\mathbf{P}(X_s, X_t) = \mathbf{P}(X_s, X_t) \left(\prod_{u \in \Gamma_s \setminus t} m_{us}(X_s) \right) \left(\prod_{u \in \Gamma_t \setminus s} m_{ut}(X_t) \right).$$

Locally consistent BP

- Locally consistent BP [Wainwright et al., 2003]

$$\Psi_s(X_s) = \mathbf{P}(X_s), \qquad \Psi(X_s, X_t) = \mathbf{P}(X_s, X_t)\mathbf{P}(X_t)^{-1}\mathbf{P}(X_t)^{-1},$$

$\mathbf{P}(X_s)$ and $\mathbf{P}(X_s, X_t)$ empirical distributions

- Fixed point, $\mathbf{P}(X_s)$ and $\mathbf{P}(X_s, X_t)$, at empirical marginals,

$$\mathbf{P}(X_s) = \mathbf{P}(X_s) \prod_{u \in \Gamma_s} m_{us}(X_s),$$

$$\mathbf{P}(X_s, X_t) = \mathbf{P}(X_s, X_t) \left(\prod_{u \in \Gamma_s \setminus t} m_{us}(X_s) \right) \left(\prod_{u \in \Gamma_t \setminus s} m_{ut}(X_t) \right).$$

- BP update: can be kernelized [Song, Gretton, Bickson, Low, and Guestrin, 2010a]

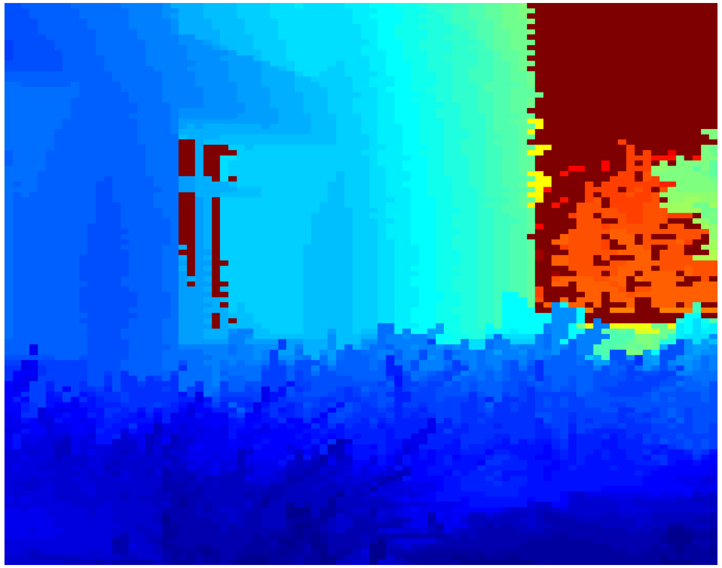
$$m_{ts}(X_s) = \int_{\mathcal{X}_t} \mathbf{P}(X_t | X_s) \prod_{u \in \Gamma_{t \setminus s}} m_{ut}(X_t) \; dX_t$$

$$= \mathbf{E}_{X_t | X_s} \left[\prod_{u \in \Gamma_{t \setminus s}} m_{ut}(X_t) \; dX_t \right].$$

Application: depth from 2D images

- 3D depth reconstruction from 2D image features.

  [Song, Gretton, Bickson, Low, and Guestrin, 2010a]
- 274 images taken on the Stanford campus [Saxena et al., 2007]
- Patches: 107 by 86, depth map using 3D laser scanners
- Patch represented by 273 dimensional feature vector:
  - local features (color and texture)
  - relative features (from adjacent patches)



Application: depth from 2D images

- Templatized model
  - Depth $y_i \in \mathbb{R}$ hidden var. for each image patch, in 2D grid
  - Depth linked to image features $x_i \in \mathbb{R}^{273}$
  - Potentials $\Psi(y_i, x_i)$ between features and depth unknown, as are $\Psi(y_i, y_k)$
- Kernels: Gaussian RBF on depth, linear on features
- Low rank QR approximation to make inference tractable

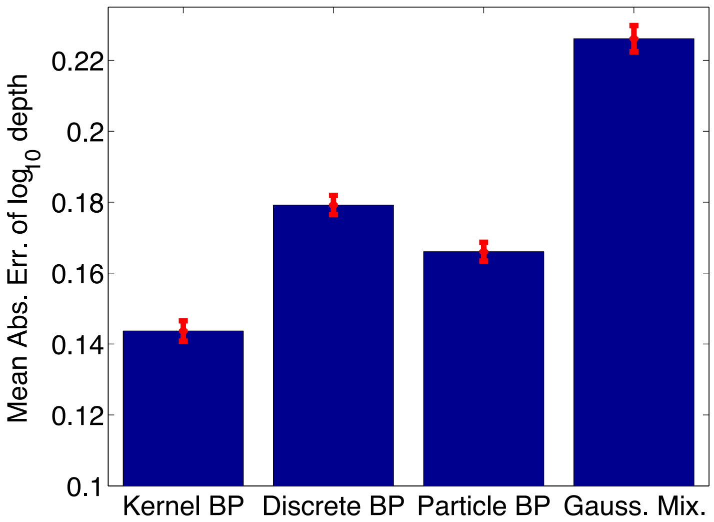
Application: depth from 2D images

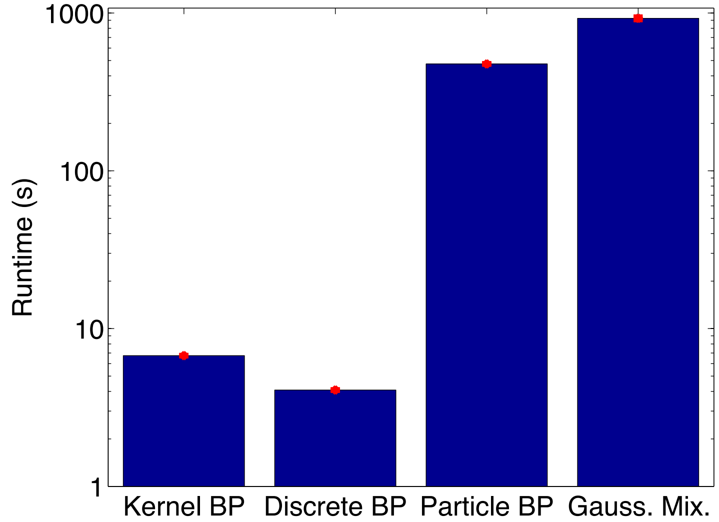
- Templatized model
  - Depth $y_i \in \mathbb{R}$ hidden var. for each image patch, in 2D grid
  - Depth linked to image features $x_i \in \mathbb{R}^{273}$
  - Potentials $\Psi(y_i, x_i)$ between features and depth unknown, as are $\Psi(y_i, y_k)$
- Kernels: Gaussian RBF on depth, linear on features
- Low rank QR approximation to make inference tractable
- Competing methods:
  - Discrete BP
  - Gaussian mixture BP [Sudderth et al., 2003]
  - Particle BP [Ihler and McAllester, 2009]
  - Conditional density learned using [Sugiyama et al., 2010]

Application: depth from 2D images

Results

- BP run for 10 iterations
- Leave-one-out error reported





Conclusions

- Kernel nonparametric message passing:
  - Exact inference on trees
  - Loopy BP on pairwise MRFs
- Advantages:
  - Complex high-dimensional/structured data
  - Non-Gaussian/multimodal
  - Density estimation/integration too expensive
  - Don't need models, just need observations!
- Experiments
  - Best performance (on all experiments)
  - Much faster than competing nonparametric methods

Questions?

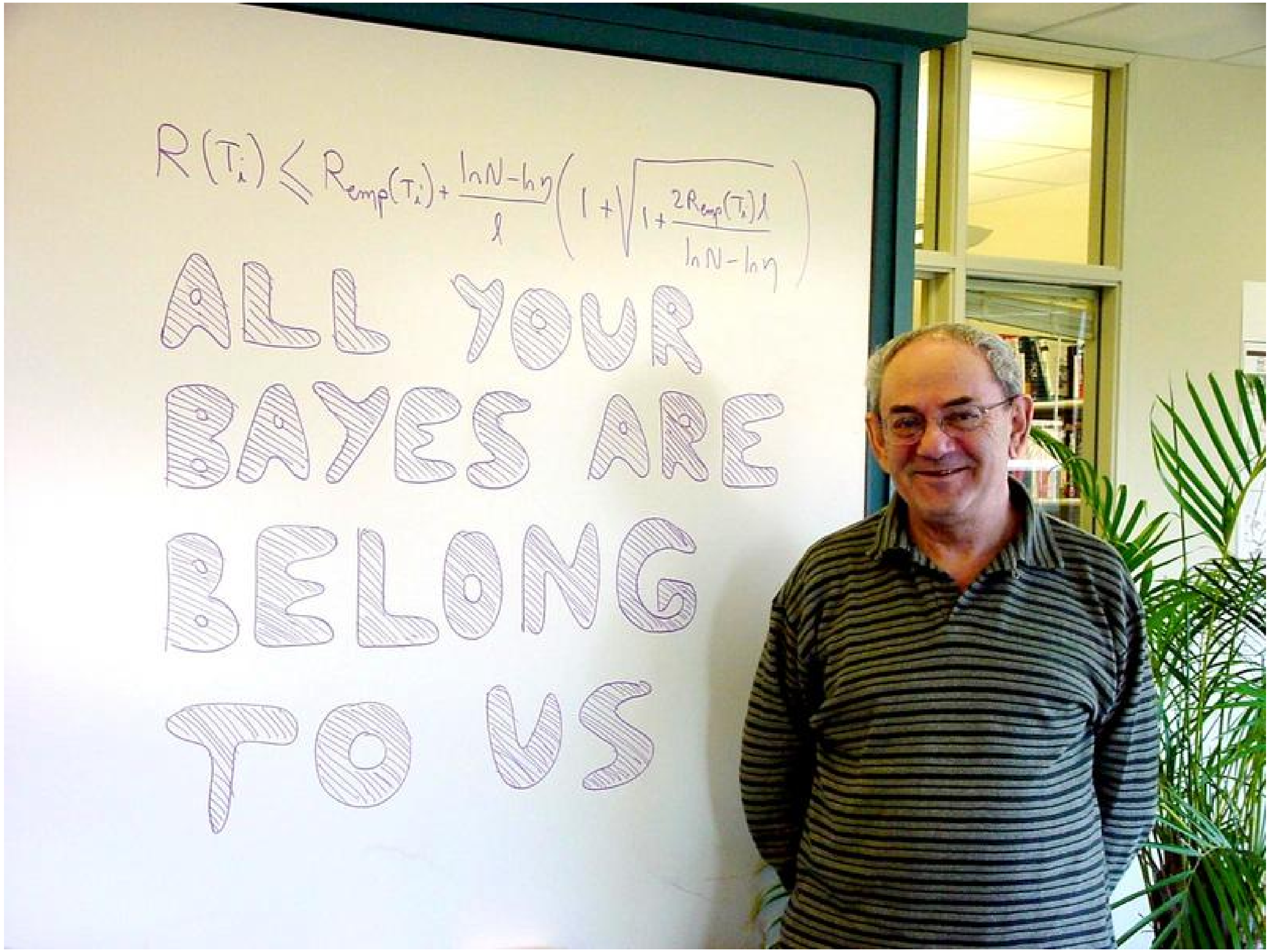


Photo credit: Yann LeCun

Bibliography

References

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, 2004.

W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184, 1991.

A. Ihler and D. McAllester. Particle belief propagation. In *AISTATS*, 2009.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86, 2005.

D. Mimno, H. Wallach, J. Naradowsky, D. Smith, and A. McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore, August 2009. ACL.

Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. 3-d depth reconstruction from a single still image. *International Journal on Computer Vision*, 76(1):53–69, 2007.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Submitted*, 2010a.

L. Song, A. Gretton, and C. Guestrin. Nonparametric tree graphical models. In *13th Workshop on Artificial Intelligence and Statistics*, volume 9 of *JMLR workshop and conference proceedings*, pages 765–772, 2010b.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions. In *Proc. Intl. Conf. Machine Learning*, 2009.

Step 3: kernelized conditional mean

Given a function $g \in \mathcal{G}$. Assume $E_{Y|X}\left[g(Y)|X=\cdot\right] \in \mathcal{F}$. Then

$$\boxed{C_{XX} E_{Y|X}\left[g(Y)|X=\cdot\right] = C_{XY}\, g.}$$

Step 3: kernelized conditional mean

Given a function $g \in \mathcal{G}$. Assume $E_{Y|X}\left[g(Y)|X = \cdot\right] \in \mathcal{F}$. Then

$$\boxed{C_{XX} E_{Y|X}\left[g(Y)|X = \cdot\right] = C_{XY} g.}$$

Proof: [Fukumizu et al., 2004]

For all $f \in \mathcal{F}$, by definition of C_{XX},

$$\begin{aligned}
&\left\langle f, C_{XX} E_{Y|X}\left[g(Y)|X = \cdot\right] \right\rangle_{\mathcal{F}} \\
&= \operatorname{cov}\left(f, E_{Y|X}\left[g(Y)|X = \cdot\right]\right) \\
&= E_X\left(f(X)\ E_{Y|X}\left[g(Y)|X\right]\right) \\
&= E_{XY}(f(X)g(Y)) \\
&= \left\langle f, C_{XY} g \right\rangle,
\end{aligned}$$

by definition of C_{XY}.

Step 3: kernelized conditional mean

- Conditional mean embedding,

$$\left\langle g, \mu_{Y|X=x} \right\rangle_{\mathcal{G}} = E_{Y|X=x} g(Y)$$

$\forall g \in \mathcal{G}$ [Song et al., 2009]

- Expression for this:

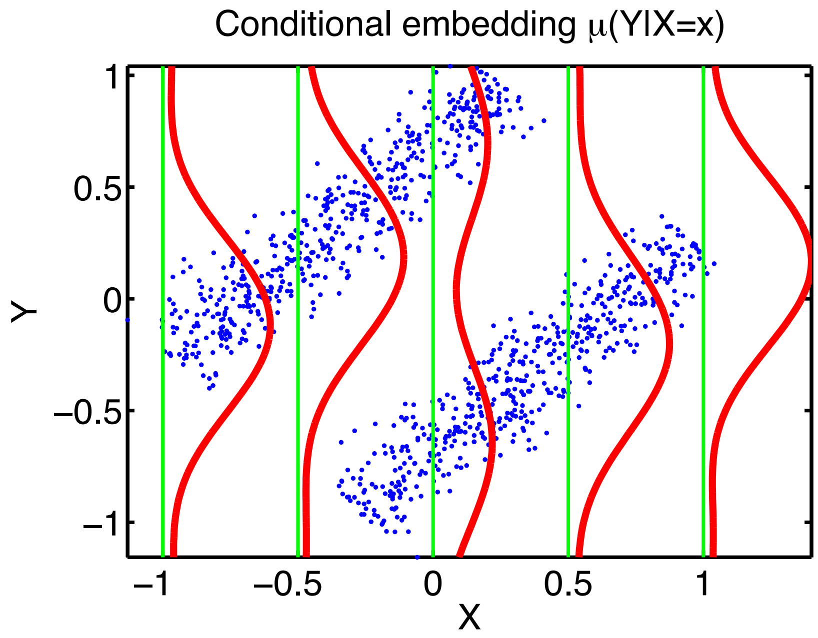
$$\begin{aligned} & E_{Y|X=x} g(Y) \\ & = \left\langle E_{Y|X}\left[g(Y) | X = \cdot \right], \varphi_x \right\rangle_{\mathcal{F}} \\ & = \left\langle C_{XX}^{-1} C_{XY} g, \varphi_x \right\rangle_{\mathcal{F}} \\ & = \left\langle g, C_{YX} C_{XX}^{-1} \varphi_x \right\rangle_{\mathcal{G}} \\ & = \left\langle g, \mu_{Y|X=x} \right\rangle_{\mathcal{G}} \end{aligned}$$

Step 3: kernelized conditional mean

- Conditional mean embedding,

$$\left\langle g, \mu_{Y|X=x} \right\rangle = E_{Y|X=x} g(Y).$$

$\forall g \in \mathcal{G}$, $\forall g \in \mathcal{G}$ [Song et al., 2009]



- Expression for this:

$$\begin{aligned} & E_{Y|X=x} g(Y) \\ & = \left\langle E_{Y|X}\left[g(Y)|X=\cdot\right], \varphi_x \right\rangle \\ & = \left\langle C_{XX}^{-1} C_{XY} g, \varphi_x \right\rangle \\ & = \left\langle g, C_{YX} C_{XX}^{-1} \varphi_x \right\rangle \\ & = \left\langle g, \mu_{Y|X=x} \right\rangle \end{aligned}$$

$\mu_{Y|X=x} := C_{YX} C_{XX}^{-1} \varphi_x.$

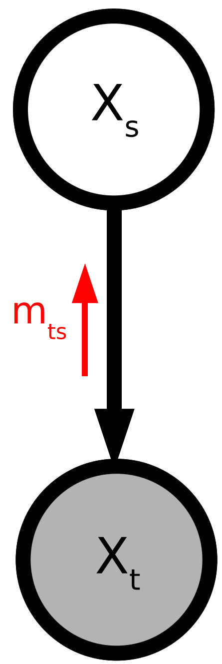
Function is conditional expectation of kernel:

$$\begin{aligned} \mu_{Y|X=x}(y) &= \langle \mu_{Y|X=x}, \phi_y \rangle \\ &= \mathbf{E}_{Y|x} l(Y, y) \end{aligned}$$

Messages from leaf nodes

- Goal: given leaf evidence x_t and parent X_S, want $m_{ts} := \mathbf{P}(x_t|X_s)$
- Assume m_{ts} an RKHS function,

$$m_{st}(x_t|x_s) := \mathbf{P}(x_t|x_s) \propto \frac{\mathbf{P}(x_s|x_t)}{\mathbf{P}(x_t)} \in \mathcal{G}_s$$



Messages from leaf nodes

- Goal: given leaf evidence x_t and parent X_S, want $m_{ts} := \mathbf{P}(x_t|X_s)$
- Assume m_{ts} an RKHS function,

$$m_{ts} := \mathbf{P}(x_t|x_s) \propto \frac{\mathbf{P}(x_t|x_s)}{\mathbf{P}(x_t)}$$

Proof: [Song, Gretton, and Guestrin, 2010b]

$$\begin{aligned}
\mu_{x_s|x_t} &= \int \mathbf{P}(x_s|x_t)\phi_{x_s} dx_s \\
&= \int \frac{\mathbf{P}(x_t|x_s)}{\mathbf{P}(x_t)} \mathbf{P}(x_s)\phi_{x_s} dx_s \\
&= \mathbf{E}_{x_s}\left[m_{ts}\phi_{x_s}\right] \\
&= \mathbf{E}_{x_s}\left[\langle m_{ts}, \phi_{x_s}\rangle \phi_{x_s}\right] \\
&= \mathbf{E}_{x_s}\left[\phi_{x_s} \otimes \phi_{x_s}\right] m_{ts} \\
&= C_{ss} m_{ts}
\end{aligned}$$