Kernel Distribution Embeddings: Theory and Applications

Arthur Gretton

Gatsby Computational Neuroscience Unit

Oxford, February 2012

First motivating question

- How do you detect dependence. . .
- . . .in a discrete domain? [Read and Cressie, 1988]

First motivating question

- How do you detect dependence…
- …in a discrete domain? [Read and Cressie, 1988]

First motivating question

- How do you detect dependence...
- ...in a discrete domain? [Read and Cressie, 1988]

P(A,T)	On time	Late
Alarm	0.27	0.03
No alarm	0.07	0.63

First motivating question

- How do you detect dependence. . .
- . . .in a discrete domain? [Read and Cressie, 1988]

P(A,T)	On time	Late
Alarm	0.10	0.20
No alarm	0.24	0.46

First motivating question

- How do you detect dependence...
- ...in a discrete domain? [Read and Cressie, 1988]

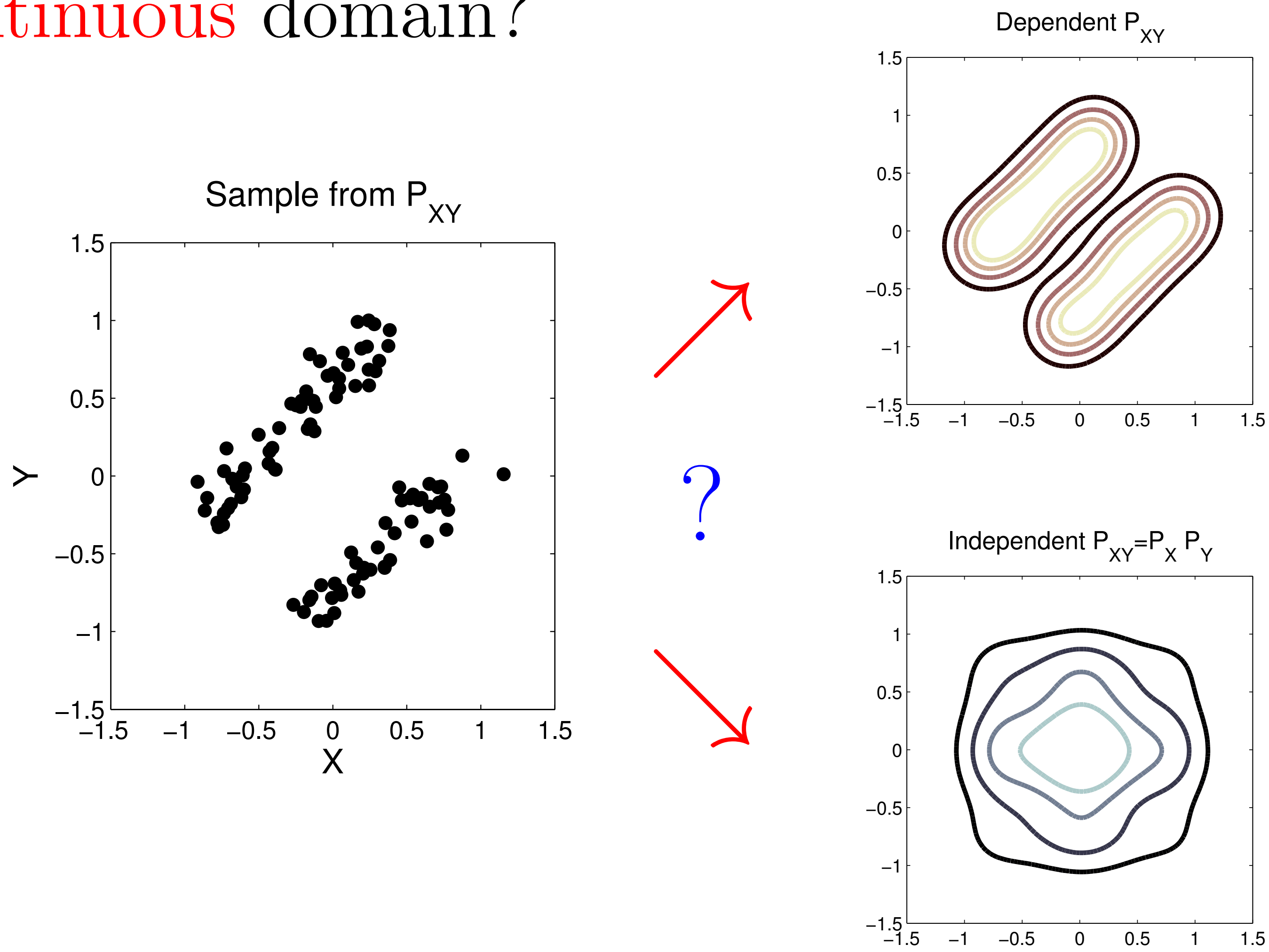
... no doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development...

$\overset{?}{\Longleftrightarrow}$

... il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants...

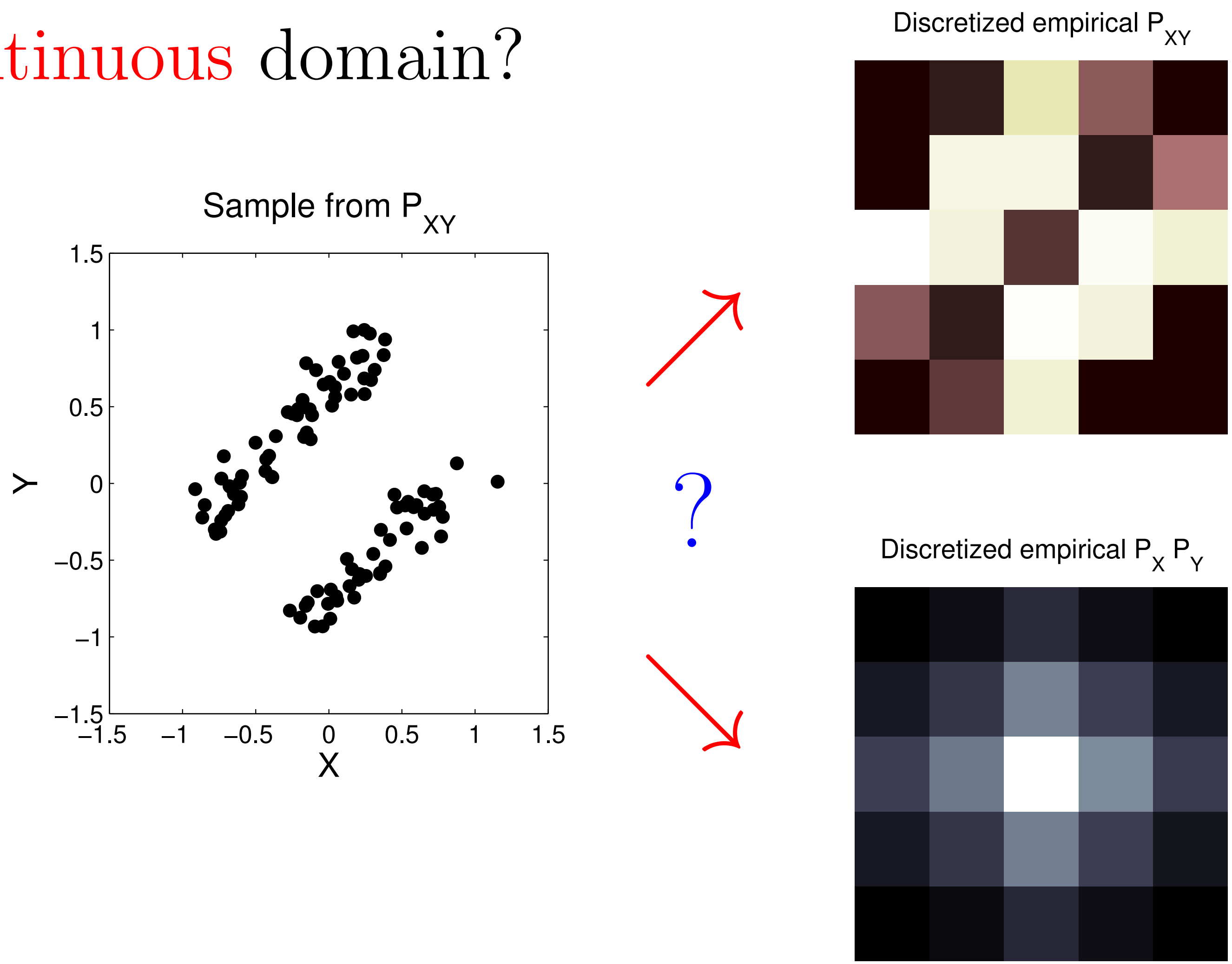
First motivating question

- How do you detect dependence. . .
- . . .in a continuous domain?



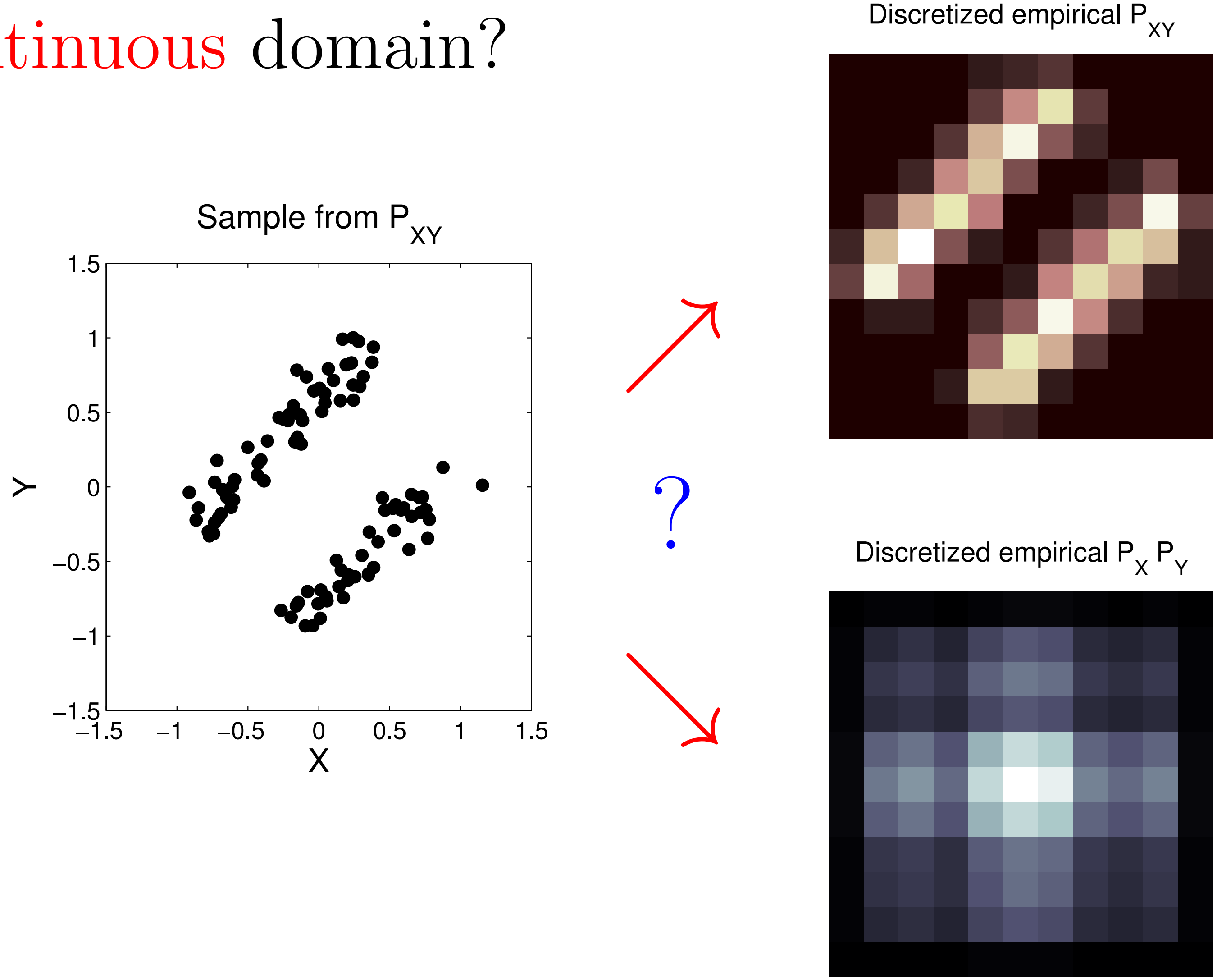
First motivating question

- How do you detect dependence. . .
- . . .in a continuous domain?



First motivating question

- How do you detect dependence. . .
- . . .in a continuous domain?



First motivating question

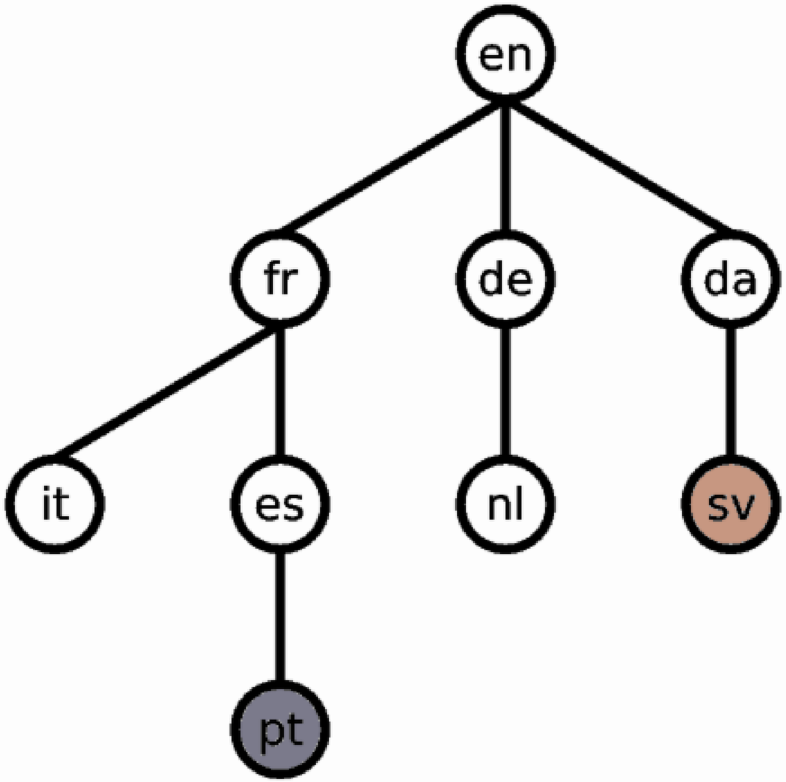
- How do you detect dependence…
- …in a continuous domain?
- Problem: fails even in "low" dimensions! [NIPS07a, ALT08]
  - X and Y in $\mathbb{R}^4$, statistic=Power divergence, samples= 1024, cases where dependence detected=0/500
- Too few points per bin

First motivating question

- How do you detect dependence. . .
- . . .in a continuous domain?
- Problem: fails even in "low" dimensions! [NIPS07a, ALT08]
  - X and Y in $\mathbb{R}^4$, statistic=Power divergence, samples= 1024, cases where dependence detected=0/500
- Too few points per bin

Can we represent and compare distributions in high dimensions?

Second question: 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

Talk Outline

- Kernel metric on the space of probability measures: Maximum Mean Discrepancy $MMD(\mathbf{P}, \mathbf{Q})$
  - Distance between means of (nonlinear) features
  - Function revealing differences in distributions
  - Dependence detection: $\mathbf{P}_{xy}$ vs $\mathbf{P}_x \mathbf{P}_y$ using $MMD(\mathbf{P}_{xy}, \mathbf{P}_x \mathbf{P}_y)$

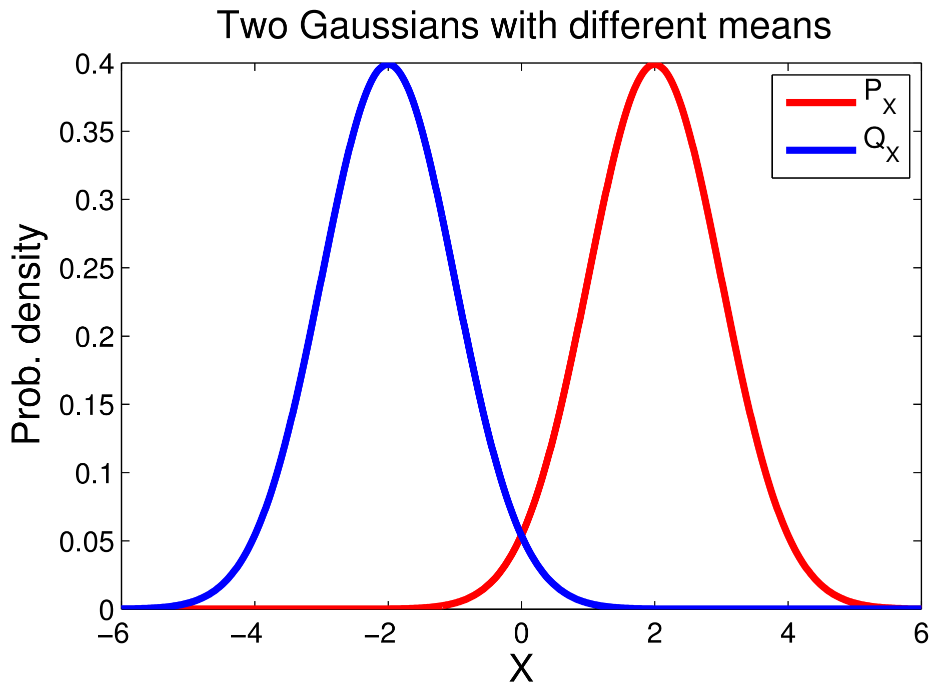
Talk Outline

- Kernel metric on the space of probability measures: Maximum Mean Discrepancy $MMD(\mathbf{P}, \mathbf{Q})$
  - Distance between means of (nonlinear) features
  - Function revealing differences in distributions
  - Dependence detection: $\mathbf{P}_{xy}$ vs $\mathbf{P}_x\mathbf{P}_y$ using $MMD(\mathbf{P}_{xy}, \mathbf{P}_x\mathbf{P}_y)$
- Kernel belief propagation:
  - Model learned from training data
  - No good parametric model
  - Other nonparametric methods fail in high dimensions, expensive

Kernel distance between distributions

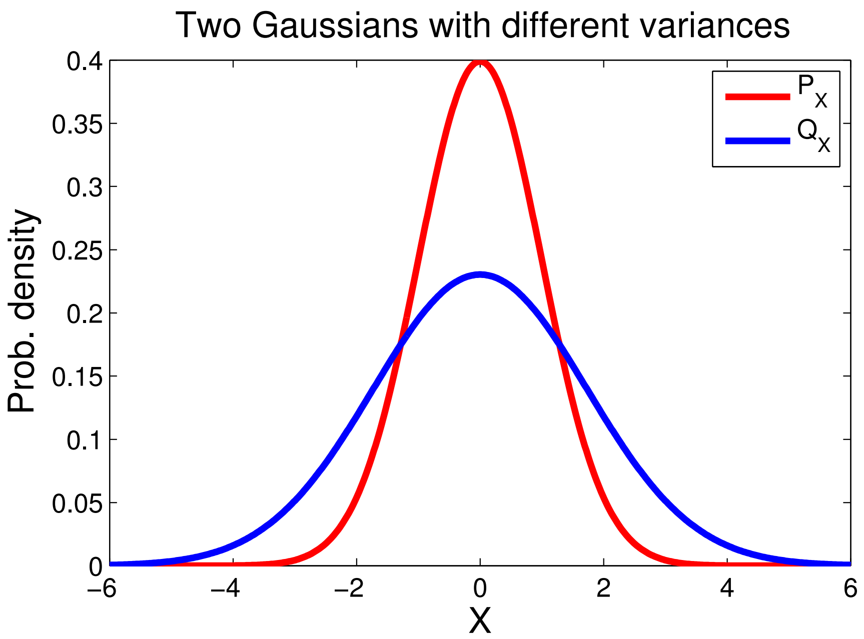
Feature mean difference

- Simple example: 2 Gaussians with different means
- Answer: t-test



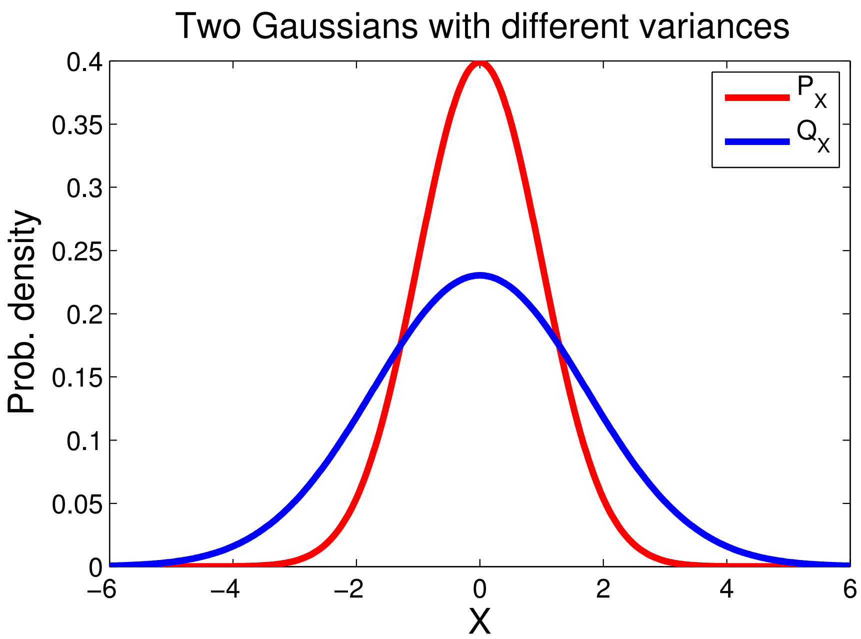
Feature mean difference

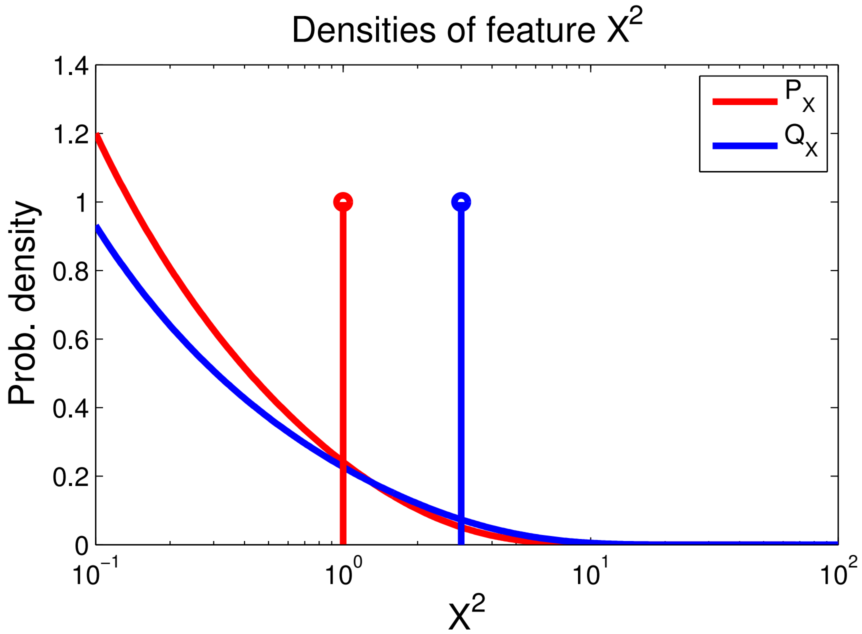
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi_x = x^2$



Feature mean difference

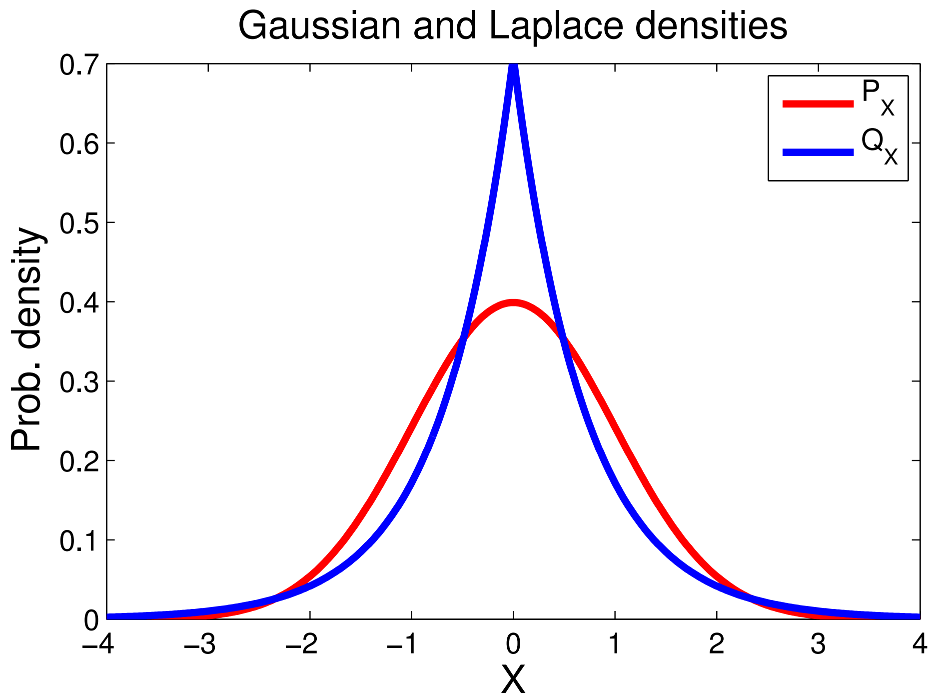
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi_x = x^2$





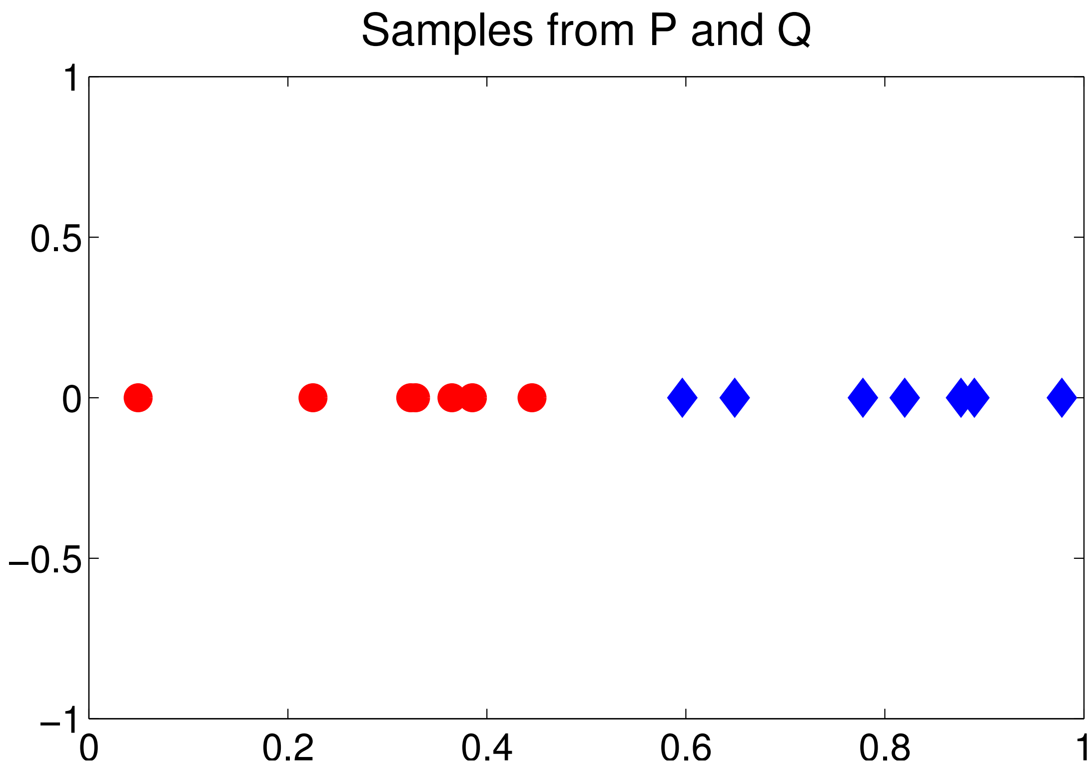
Feature mean difference

- Gaussian and Laplace distributions
- Same mean *and* same variance
- Difference in means using higher order features



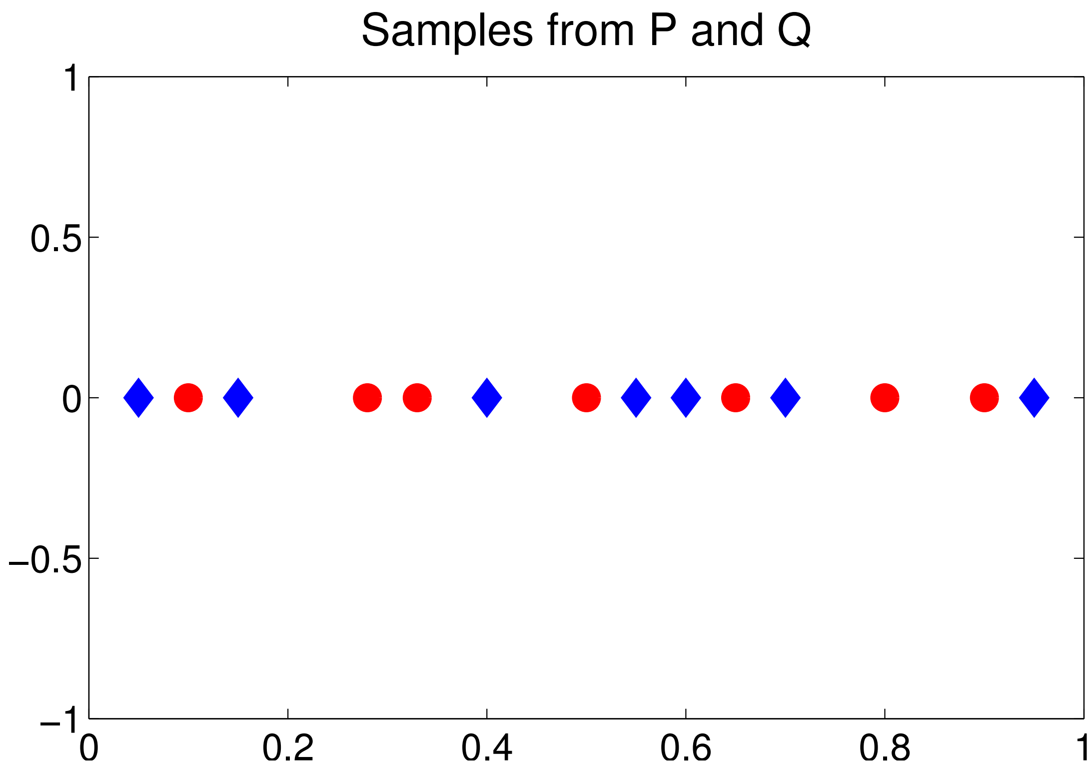
Function Showing Difference in Distributions

- Are **P** and **Q** different?



Function Showing Difference in Distributions

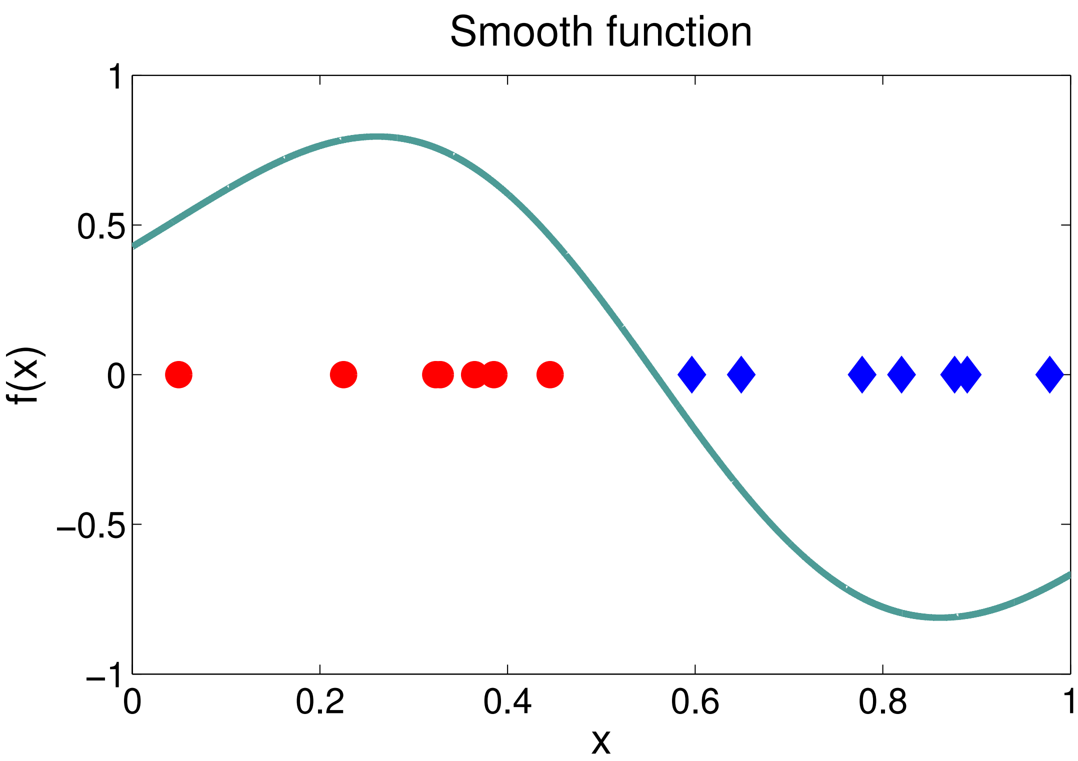
- Are **P** and **Q** different?



Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

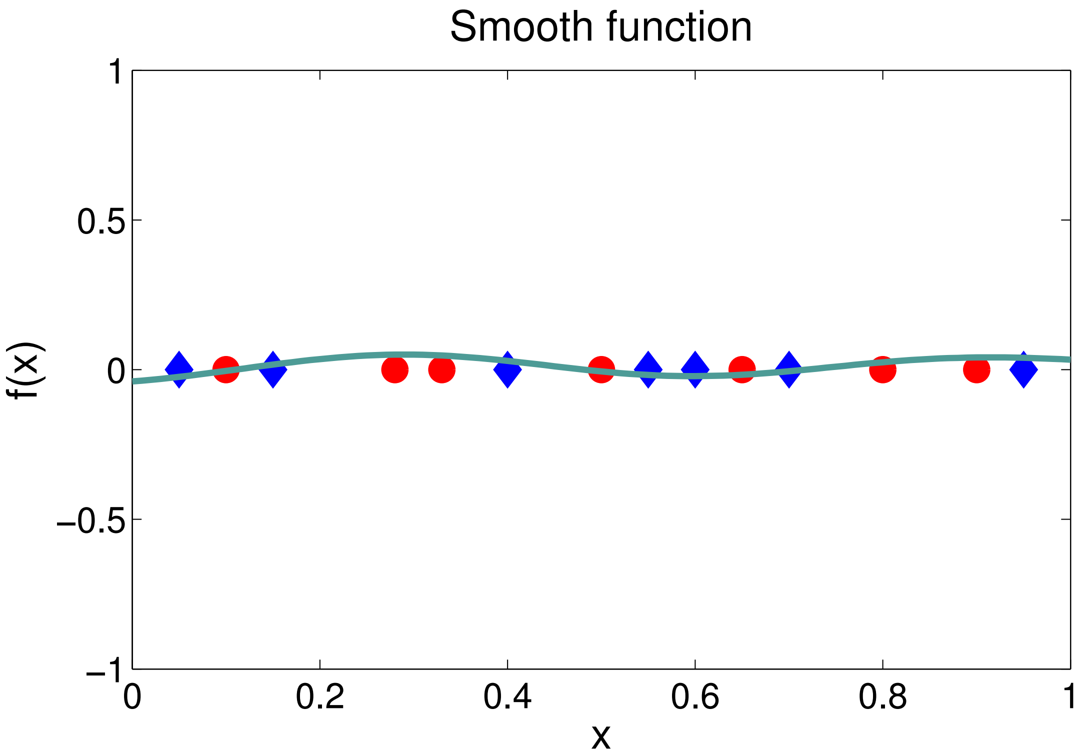
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathsf{y})\right].$$



Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

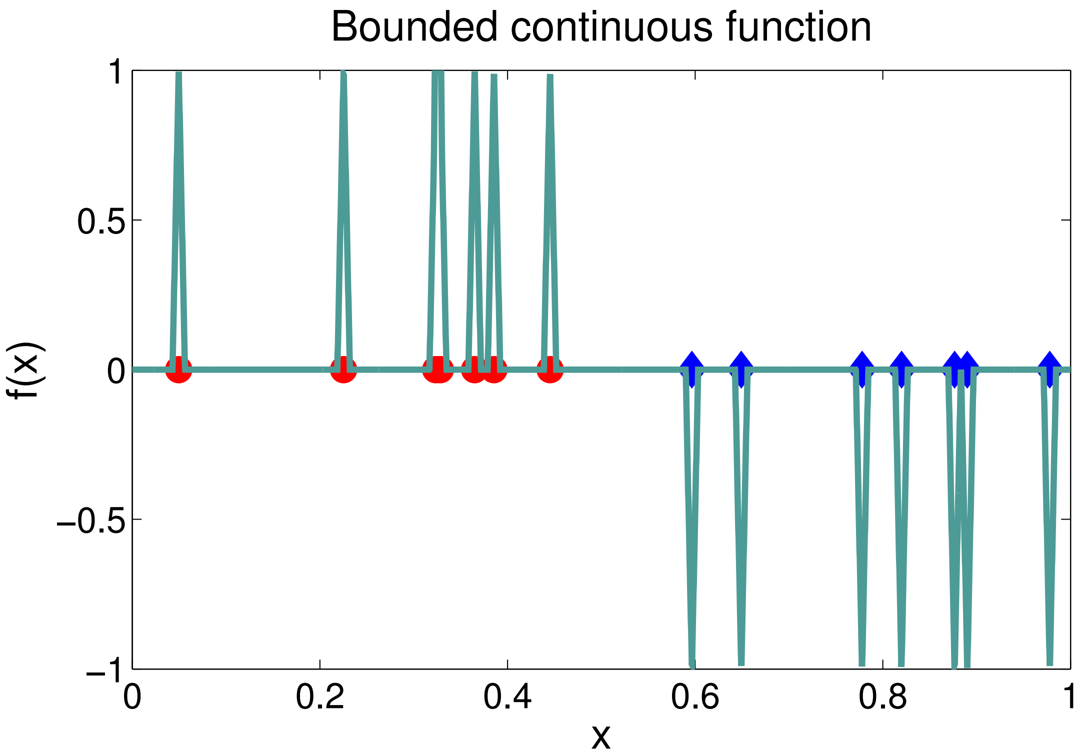
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathsf{y})\right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

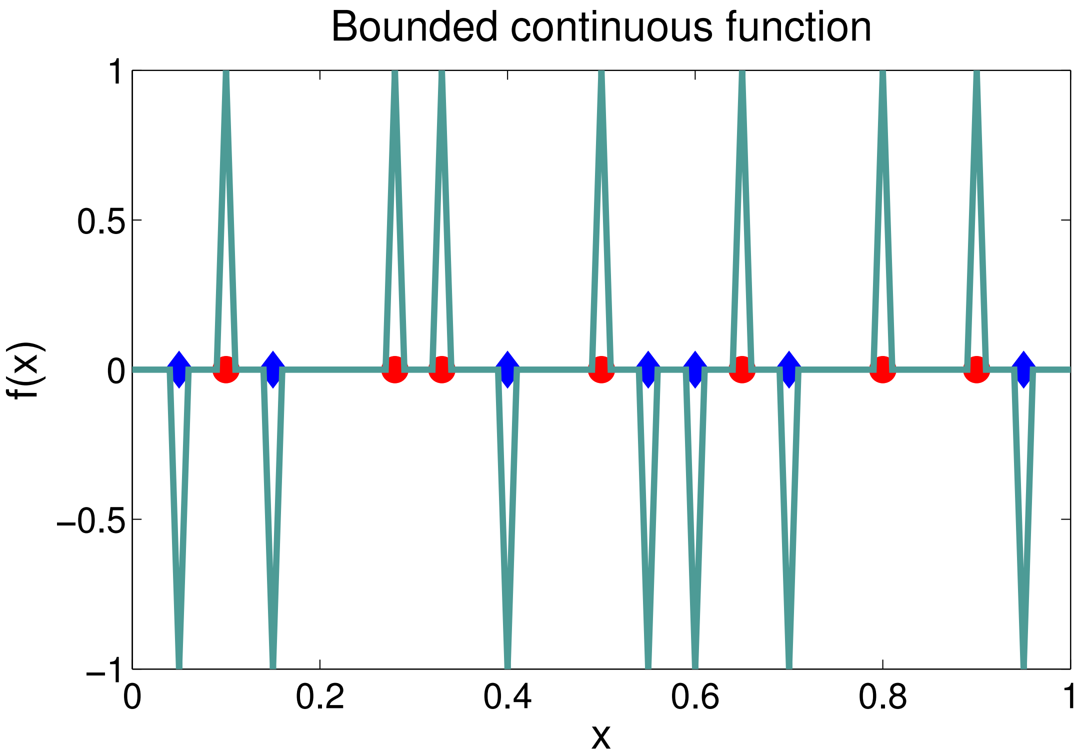
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathrm{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathrm{y}) \right].$$

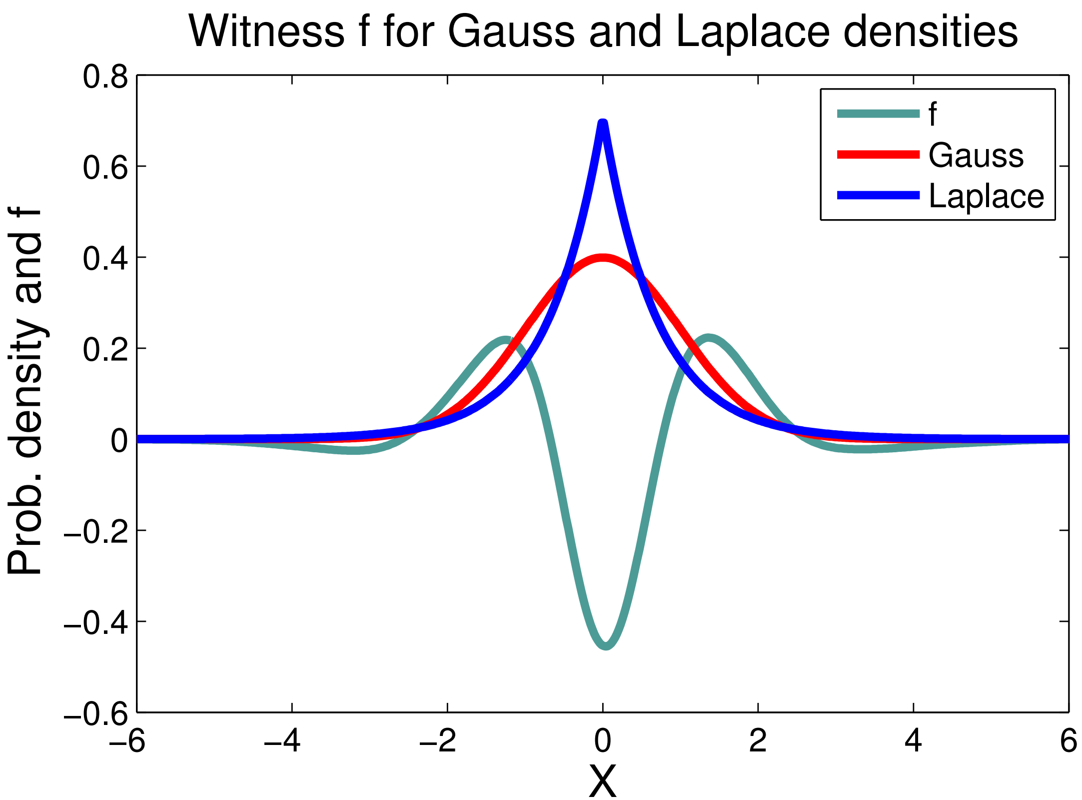


Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathsf{y})\right].$$

- Gauss **P** vs Laplace **Q**



Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$

- Classical results: $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when
  - $F =$bounded continuous [Dudley, 2002]
  - $F =$ bounded variation 1 (Kolmogorov metric) [Müller, 1997]
  - $F =$ bounded Lipschitz (Earth mover's distances) [Dudley, 2002]

Function Showing Difference in Distributions

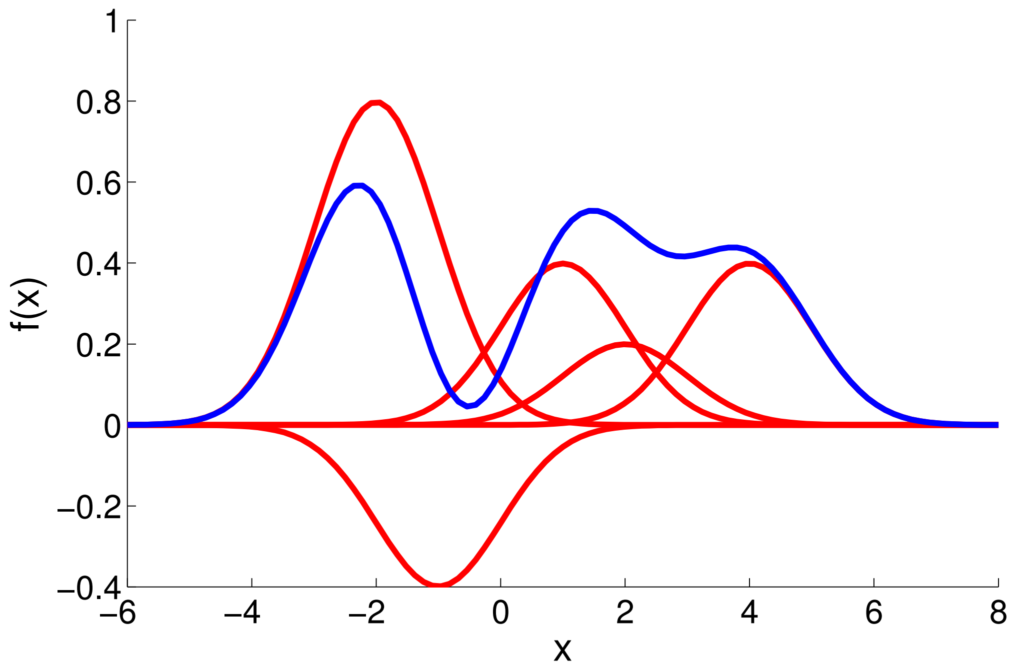
- Maximum mean discrepancy: smooth function for $\mathbf{P}$ vs $\mathbf{Q}$

$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$

- Classical results: $\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when
  - $F =$bounded continuous [Dudley, 2002]
  - $F =$ bounded variation 1 (Kolmogorov metric) [Müller, 1997]
  - $F =$ bounded Lipschitz (Earth mover's distances) [Dudley, 2002]
- $\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F =$the unit ball in a characteristic RKHS $\mathcal{F}$ [ISMB06, NIPS06a, NIPS07b, NIPS08a, JMLR10]

Functions in the RKHS

- $\mathcal{F}$ 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 $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, $x_i \in \mathcal{X}$.



The RKHS as feature map

- Feature map of $x \in \mathbb{R}^2$, written φ_x

$$\varphi_x^{(p)} = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad\qquad \varphi_x^{(g)} = \exp\left(-\lambda \left\| x - \cdot \right\|^2\right)$$

The RKHS as feature map

- Feature map of $x \in \mathbb{R}^2$, written φ_x

$$\varphi_x^{(p)} = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad\qquad \varphi_x^{(g)} = \exp\left(-\lambda \left\| x - \cdot \right\|^2\right)$$

- Inner product between feature maps:

$$\left\langle \varphi_x^{(p)}, \varphi_y^{(p)} \right\rangle_{\mathcal{F}} = \langle x, y \rangle^2 \qquad\qquad \left\langle \varphi_x^{(g)}, \varphi_y^{(g)} \right\rangle_{\mathcal{F}} = \exp\left(-\lambda \left\| x - y \right\|^2\right)$$

The RKHS as feature map

- Feature map of $x \in \mathbb{R}^2$, written φ_x

$$\varphi_x^{(p)} = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad \varphi_x^{(g)} = \exp\left(-\lambda \left\| x - \cdot \right\|^2\right)$$

- Inner product between feature maps:

$$\left\langle \varphi_x^{(p)}, \varphi_y^{(p)} \right\rangle_{\mathcal{F}} = \langle x, y \rangle^2 \qquad \left\langle \varphi_x^{(g)}, \varphi_y^{(g)} \right\rangle_{\mathcal{F}} = \exp\left(-\lambda \left\| x - y \right\|^2\right)$$

- In general,

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

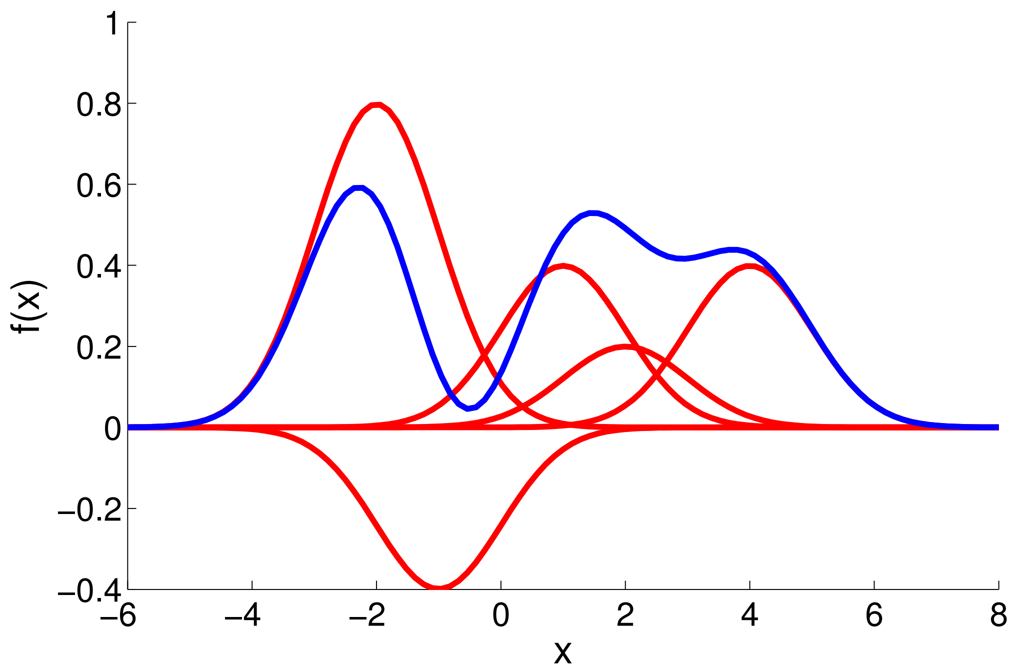
for positive definite $k(x, y)$

$\boxed{\text{Kernels are inner products of feature maps}}$

The RKHS as feature map

- Example:

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

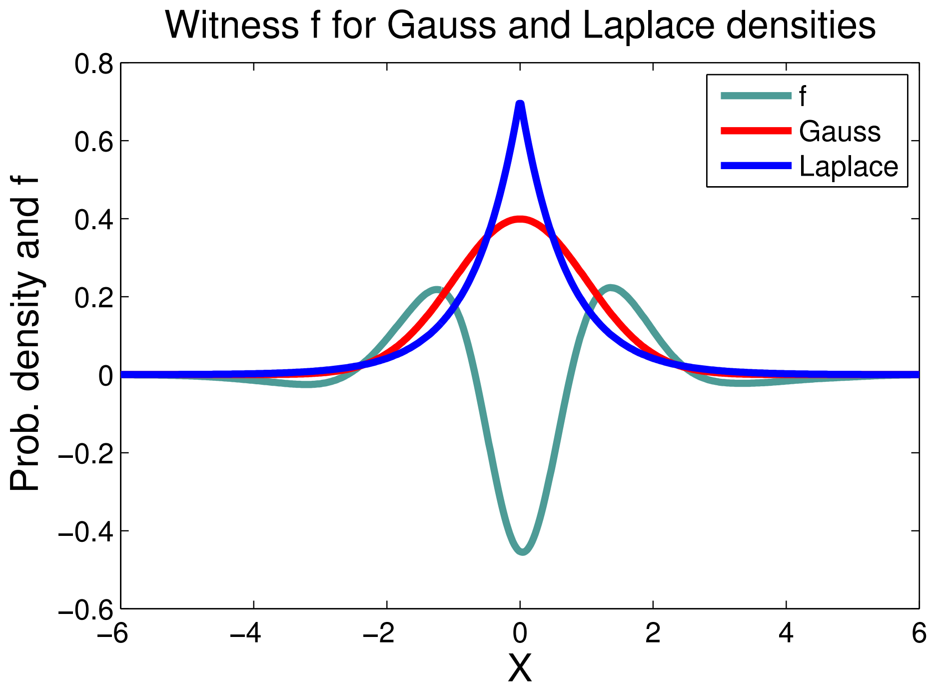


Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

$$\mathrm{MMD}^2(\mathbf{P},\mathbf{Q};F)$$

$$= \left(\sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y})\right]\right)^2$$



Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

$$\mathrm{MMD}^2(\mathbf{P},\mathbf{Q};F)$$

$$= \left(\sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y})\right]\right)^2$$

use

$$\begin{aligned}\mathbf{E}_{\mathbf{P}}(f(\mathsf{x})) &= \mathbf{E}_{\mathbf{P}}\left[\langle \varphi_x, f\rangle_{\mathcal{F}}\right] \\ &=: \langle \mu_{\mathbf{P}}, f\rangle_{\mathcal{F}}\end{aligned}$$

Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

$$\mathrm{MMD}^2(\mathsf{P}, \mathsf{Q}; F)$$

$$= \left(\sup_{f \in F} \left[\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y}) \right] \right)^2$$

$$= \left(\sup_{f \in F} \langle f, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} \right)^2$$

use

$$\begin{aligned} \mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) &= \mathbf{E}_{\mathsf{P}} \left[\langle \varphi_x, f \rangle_{\mathcal{F}} \right] \\ &=: \langle \mu_{\mathsf{P}}, f \rangle_{\mathcal{F}} \end{aligned}$$

Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

$$
\begin{aligned}
&\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}; F) \\
&= \left(\sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right] \right)^2 \\
&= \left(\sup_{f \in F} \langle f, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \right)^2 \\
&= \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2
\end{aligned}
$$

use

$$
\|\theta\|_{\mathcal{F}} = \sup_{f \in F} \langle f, \theta \rangle_{\mathcal{F}}
$$

$\boxed{\text{Function view and feature view equivalent}}$

Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

$$\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}; F)$$

$$= \left(\sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right] \right)^2$$

$$= \left(\sup_{f \in F} \langle f, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \right)^2$$

$$= \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2$$

use

$$\|\theta\|_{\mathcal{F}} = \sup_{f \in F} \langle f, \theta \rangle_{\mathcal{F}}$$

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

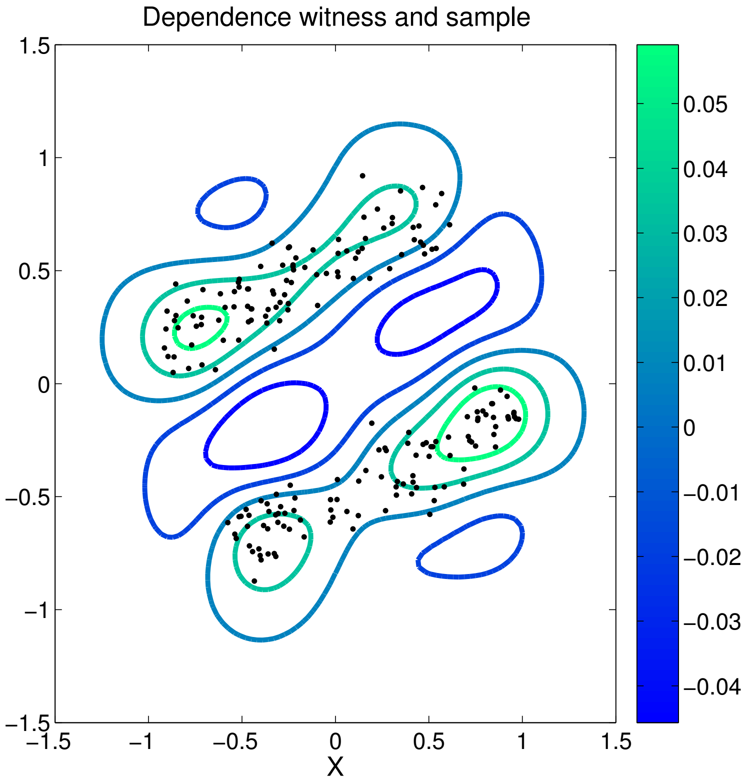
$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j) \right]$$

MMD for independence

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

$$\left(\sup_f \left[\mathbf{E}_{\mathsf{P}_{XY}} f - \mathbf{E}_{\mathsf{P}_X \mathsf{P}_Y} f\right]\right)^2 = \sup_{\|f\| \leq 1} \langle f, \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \rangle^2_{\mathcal{F} \times \mathcal{G}}$$

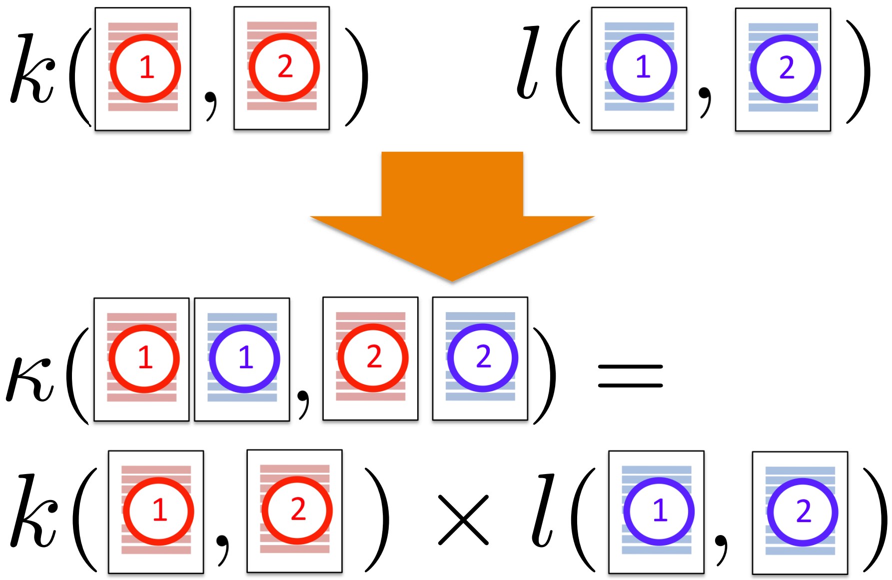
$$= \|\mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y}\|^2_{\mathcal{F} \times \mathcal{G}} := MMD(\mathsf{P}_{XY}, \mathsf{P}_X \mathsf{P}_Y)$$



MMD for independence

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

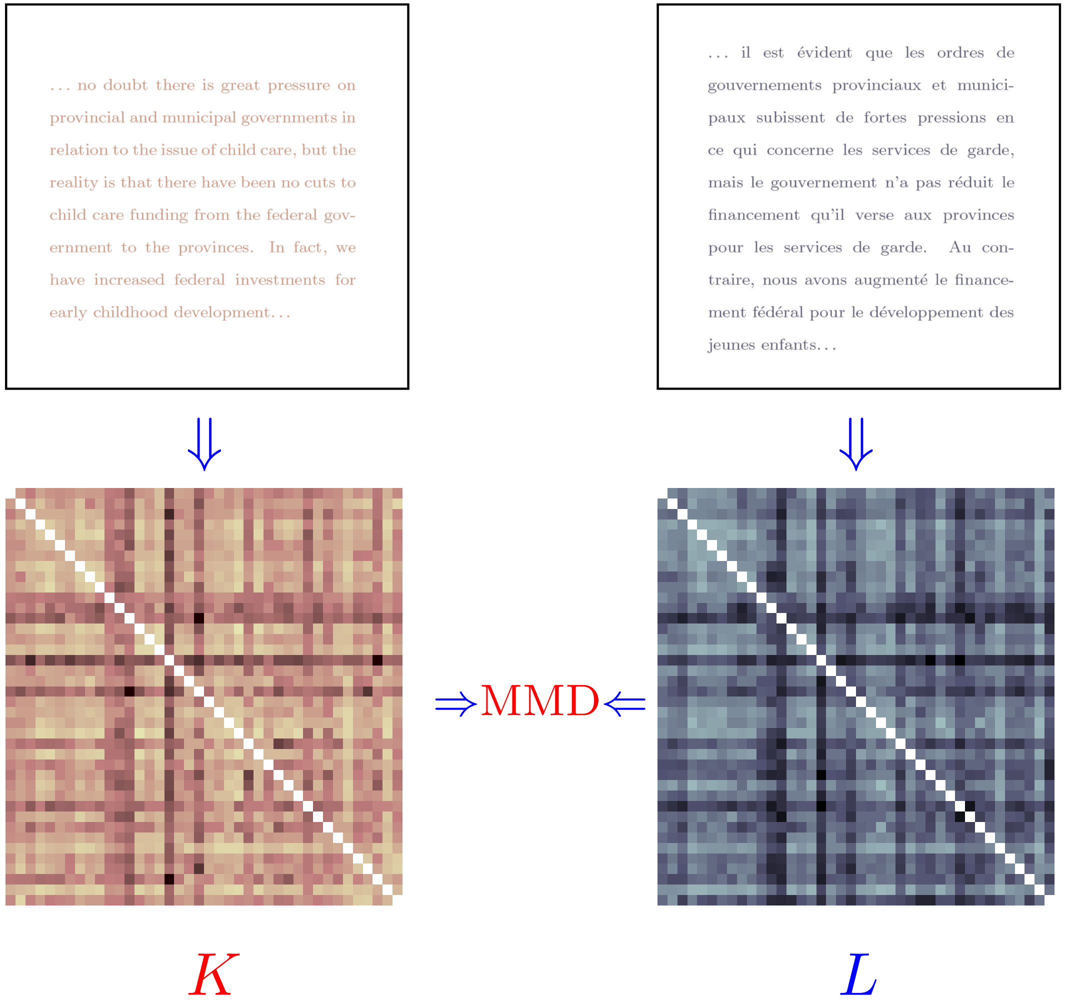
$$
\begin{aligned}
\left(\sup_f \left[\mathbf{E}_{\mathsf{P}_{XY}} f - \mathbf{E}_{\mathsf{P}_X \mathsf{P}_Y} f\right]\right)^2 &= \sup_{\|f\| \leq 1} \left\langle f, \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\rangle^2_{\mathcal{F} \times \mathcal{G}} \\
= \left\| \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\|^2_{\mathcal{F} \times \mathcal{G}} &:= MMD(\mathsf{P}_{XY}, \mathsf{P}_X \mathsf{P}_Y)
\end{aligned}
$$



Experiment: dependence testing for translation

- Translation example: [NIPS07b] Canadian Hansard (agriculture)
- 5-line extracts, k-spectrum kernel, $k = 10$, repetitions=300, sample size 10
- Empirical $MMD(\mathbf{P}_{XY}, \mathbf{P}_X \mathbf{P}_Y)$:

$$\frac{1}{m^2} \mathrm{trace}(\mathbf{KHLH})$$



- k-spectrum kernel: average Type II error 0 ($\alpha = 0.05$)
- Bag of words kernel: average Type II error 0.18

Kernel Belief Propagation

Nonparametric belief propagation

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

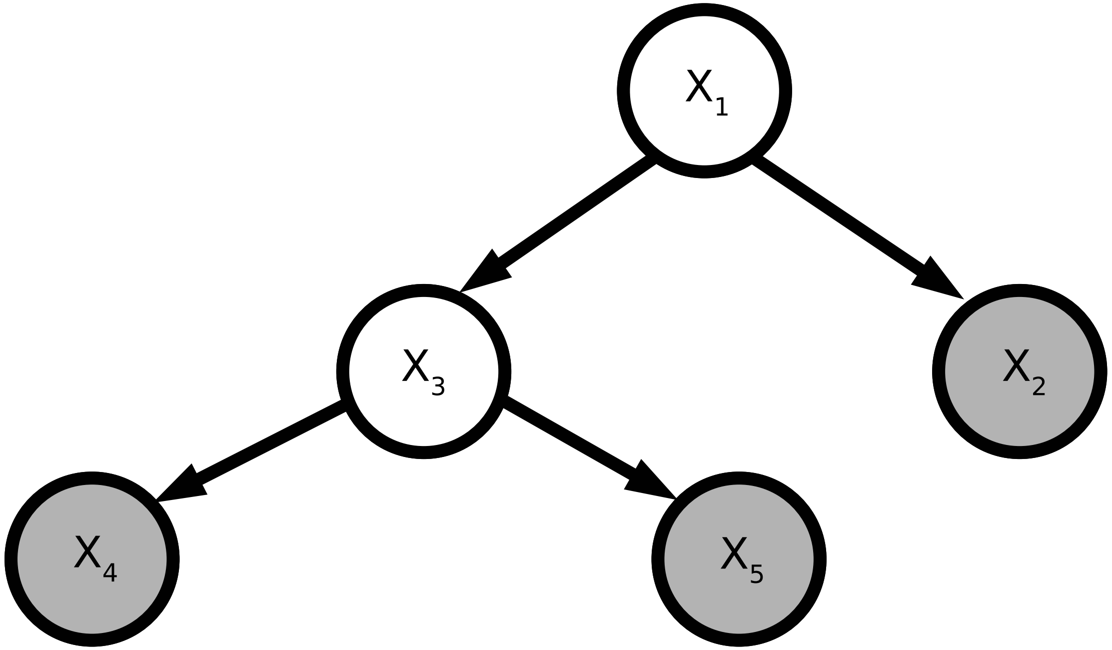
Nonparametric belief propagation

- Why use a non-parametric (kernel) algorithm?
  - Model learned from training data
  - Complex high-dimensional/structured data (discretization fails)
  - Non-Gaussian/multimodal (Gaussian BP fails)
  - Numerical integration too expensive (Parzen window approximations fail)
- Exact inference on trees [Song, Gretton, and Guestrin, 2010]
  - Cross-language document retrieval
  - Camera orientation recovery from images
- Loopy BP on pairwise MRFs [Song, Gretton, Bickson, Low, and Guestrin, 2011]
  - Depth recovery from 2D images
  - Predicting paper categories from citation networks
  - Protein structure prediction

Nonparametric belief propagation

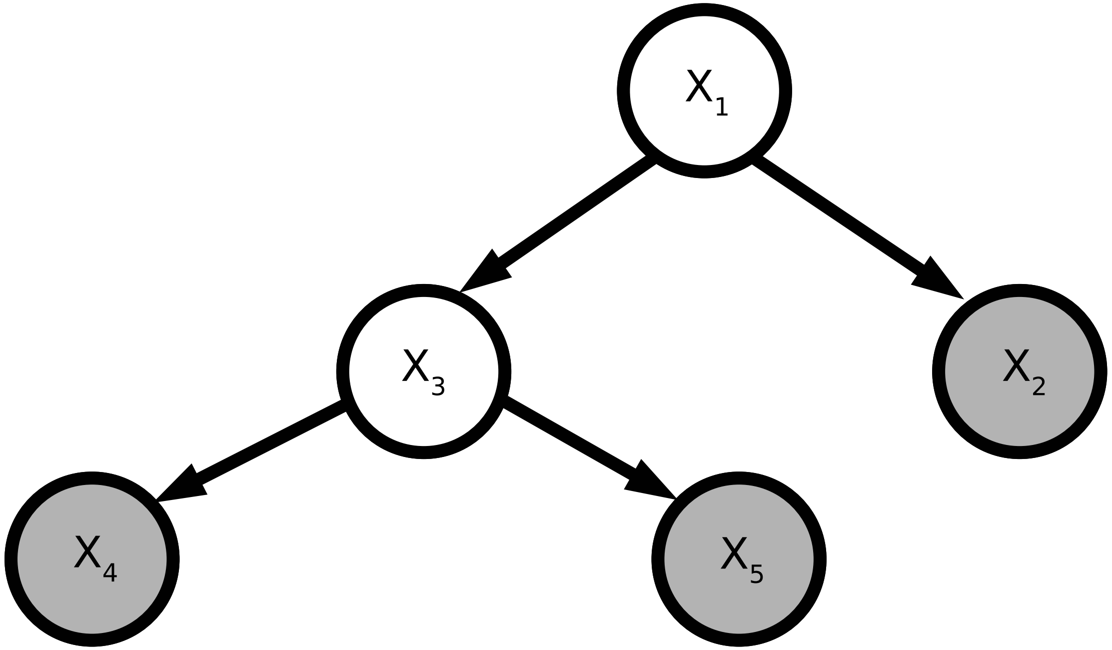
- Why use a non-parametric (kernel) algorithm?
  - Model learned from training data
  - Complex high-dimensional/structured data (discretization fails)
  - Non-Gaussian/multimodal (Gaussian BP fails)
  - Numerical integration too expensive (Parzen window approximations fail)
- Exact inference on trees [Song, Gretton, and Guestrin, 2010]
  - Cross-language document retrieval
  - Camera orientation recovery from images
- Loopy BP on pairwise MRFs [Song, Gretton, Bickson, Low, and Guestrin, 2011]
  - Depth recovery from 2D images
  - Predicting paper categories from citation networks
  - Protein structure prediction

Message passing on directed graphical models



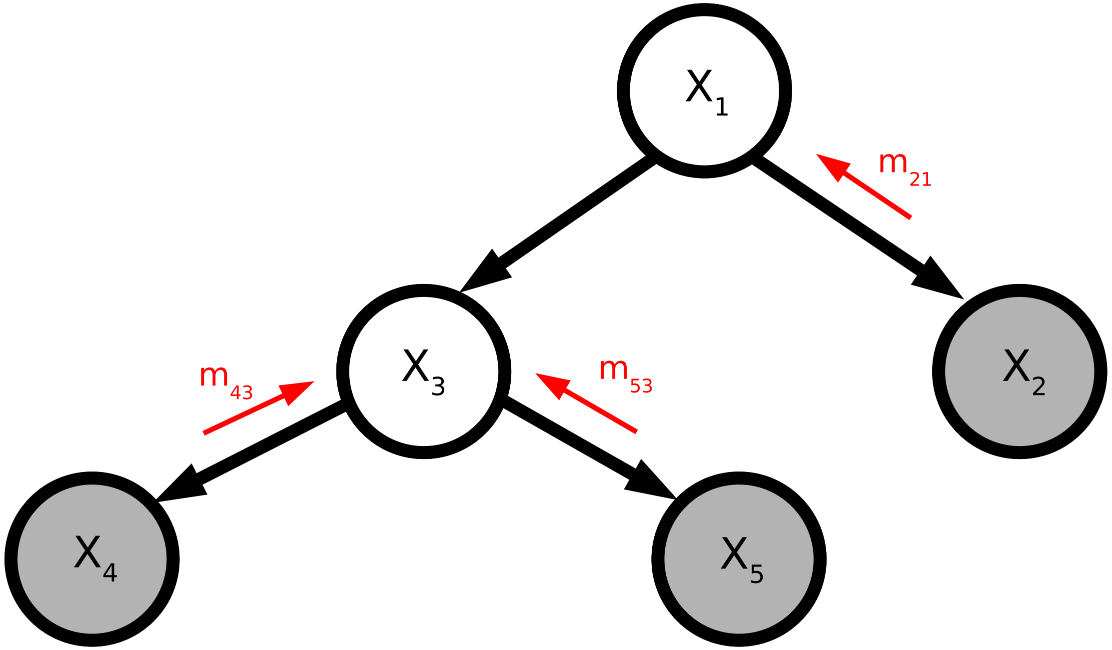
$$\mathbf{P}(X_1, x_2, x_4, x_5) = \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3)$$

Message passing on directed graphical models



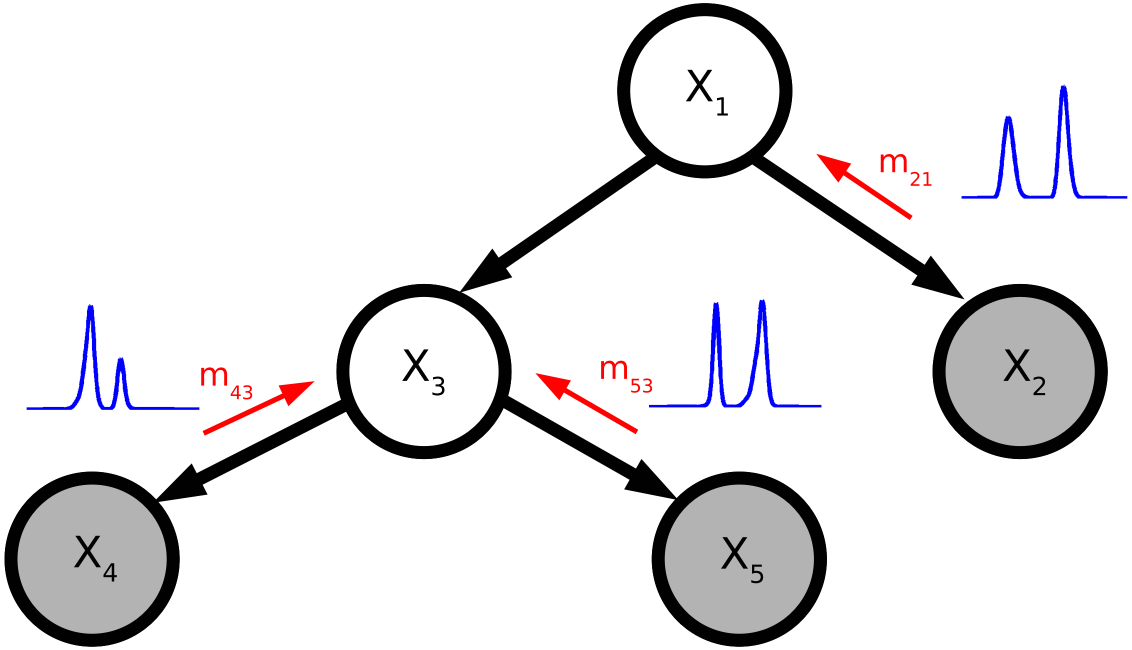
$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\int_{x_3} \mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3)
\end{aligned}
$$

Message passing on directed graphical models



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

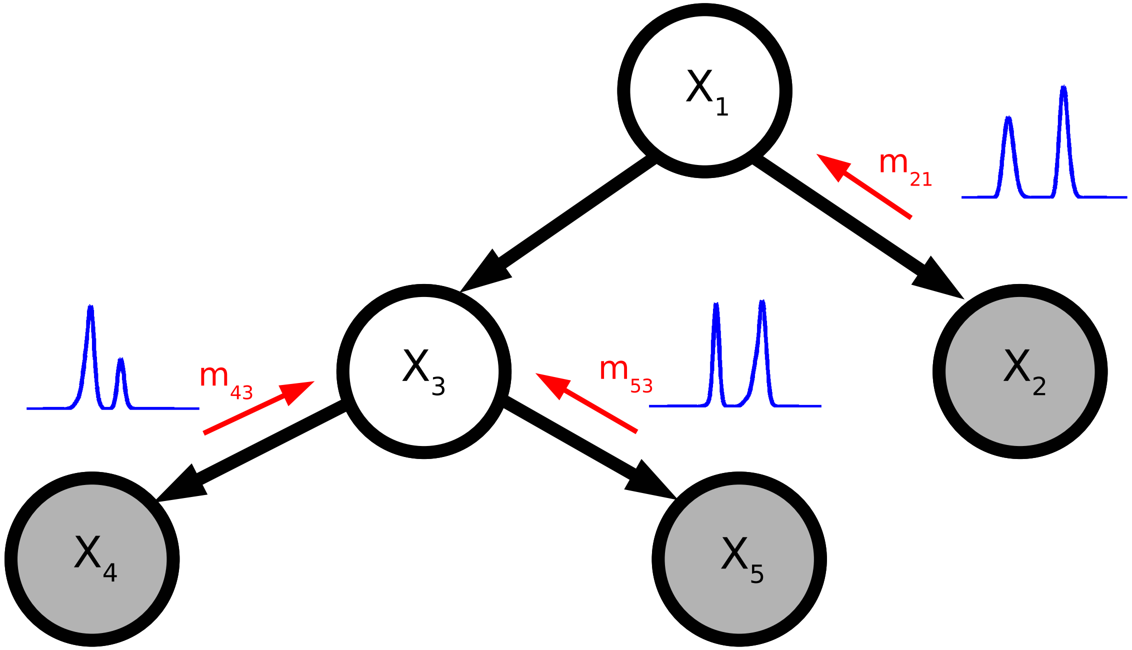
Message passing on directed graphical models



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

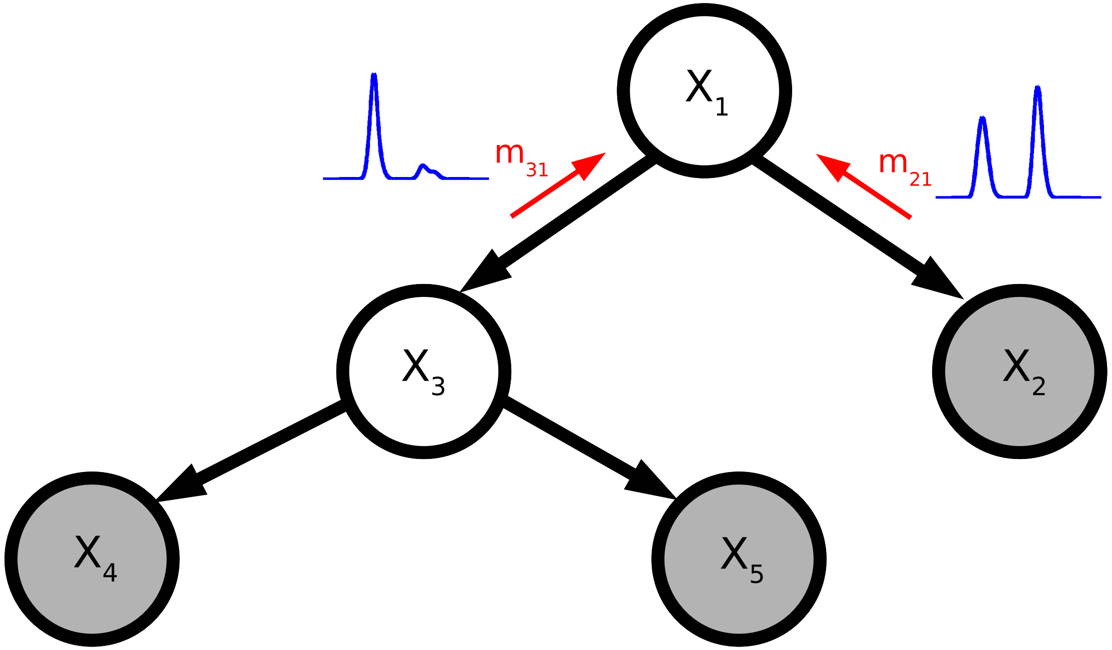
$$= \mathbf{P}(X_1)\underbrace{\mathbf{P}(x_2|X_1)}_{m_{21}(X_1)}\int_{x_3} \mathbf{P}(X_3|X_1)\underbrace{\mathbf{P}(x_4|X_3)}_{m_{43}(X_3)}\underbrace{\mathbf{P}(x_5|X_3)}_{m_{53}(X_3)}$$

Message passing on directed graphical models



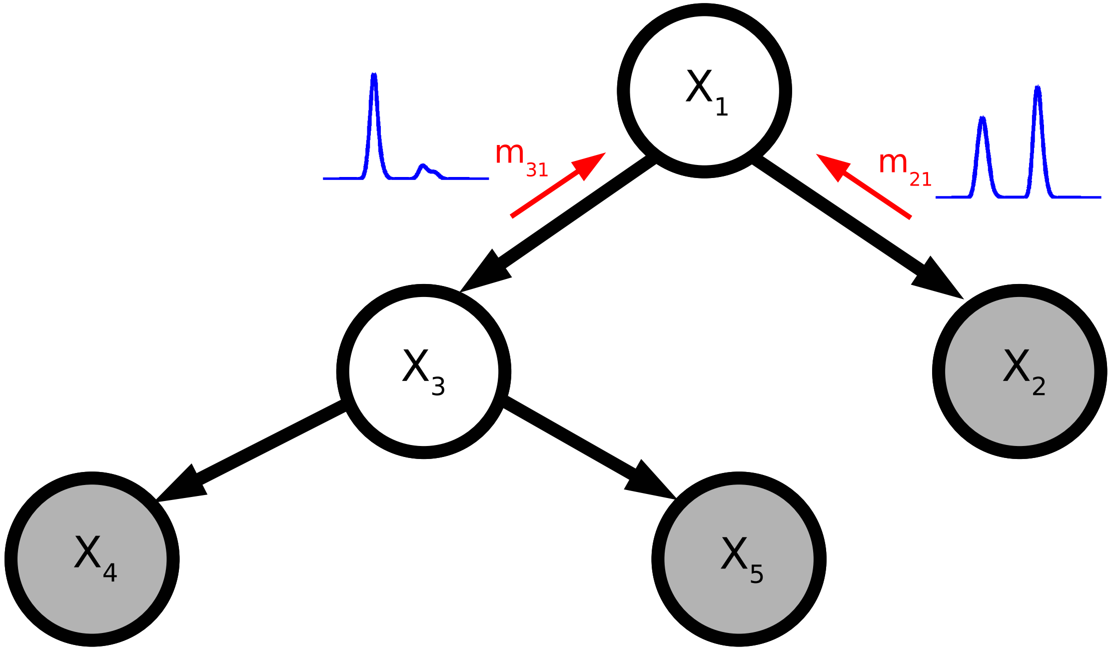
$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)\color{red}{m_{21}(X_1)}\color{black} \int_{x_3} \mathbf{P}(X_3|X_1)\color{red}{m_{43}(X_3)m_{53}(X_3)}
\end{aligned}
$$

Message passing on directed graphical models



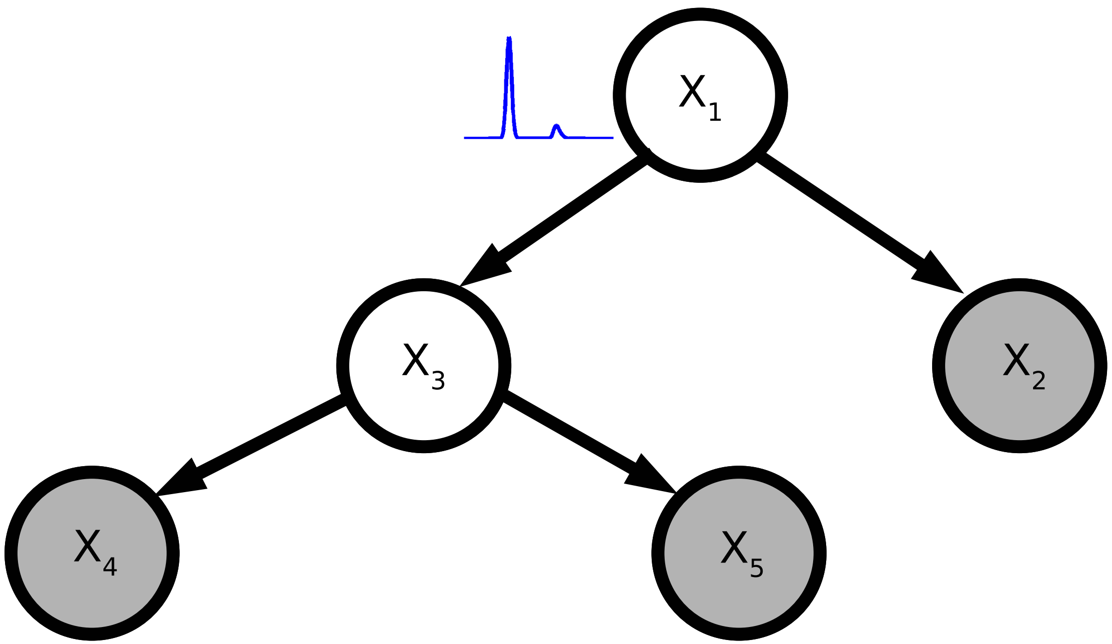
$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)\color{red}{m_{21}(X_1)}\color{black}\underbrace{\int_{x_3} \mathbf{P}(X_3|X_1)m_{43}(X_3)m_{53}(X_3)}_{\color{red}{m_{31}(X_1)}}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

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



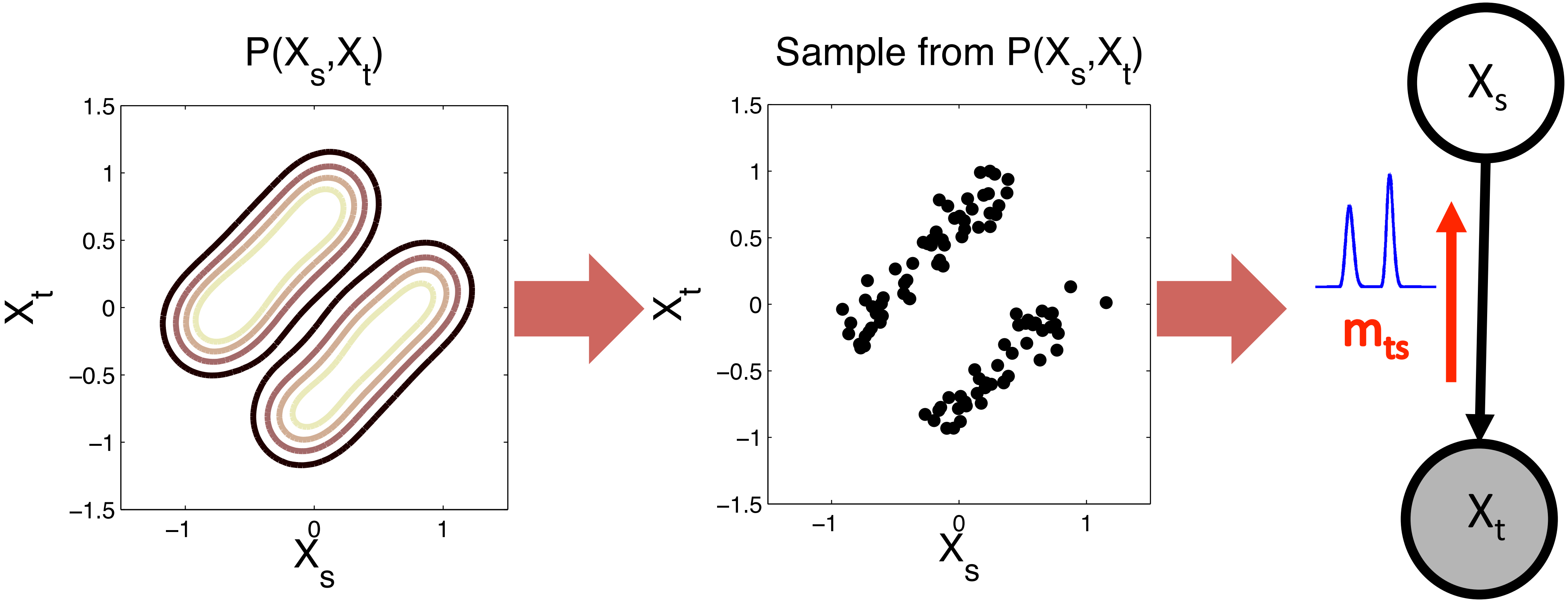
$$
\begin{aligned}
\mathbf{P}(X_1, x_2, x_4, x_5) &= \int_{x_3} \mathbf{P}(X_1)\mathbf{P}(x_2|X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(x_4|X_3)\mathbf{P}(x_5|X_3) \\
&= \mathbf{P}(X_1)m_{21}(X_1)m_{31}(X_1)
\end{aligned}
$$

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

Messages from observed leaves

- Pairwise interaction learned from training data
- 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 $m_{ts} := \mathbf{P}(x_t|X_s)$

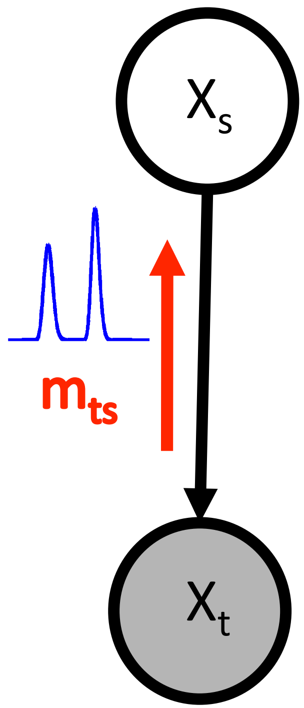
- Training data

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

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

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

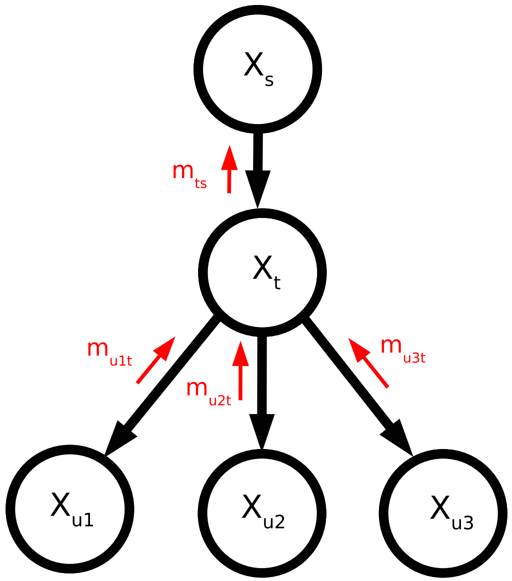
$$\beta_{ts} = ((K_t + \lambda I)(K_s + \lambda I))^{-1} k_t$$



Marginalize over internal nodes

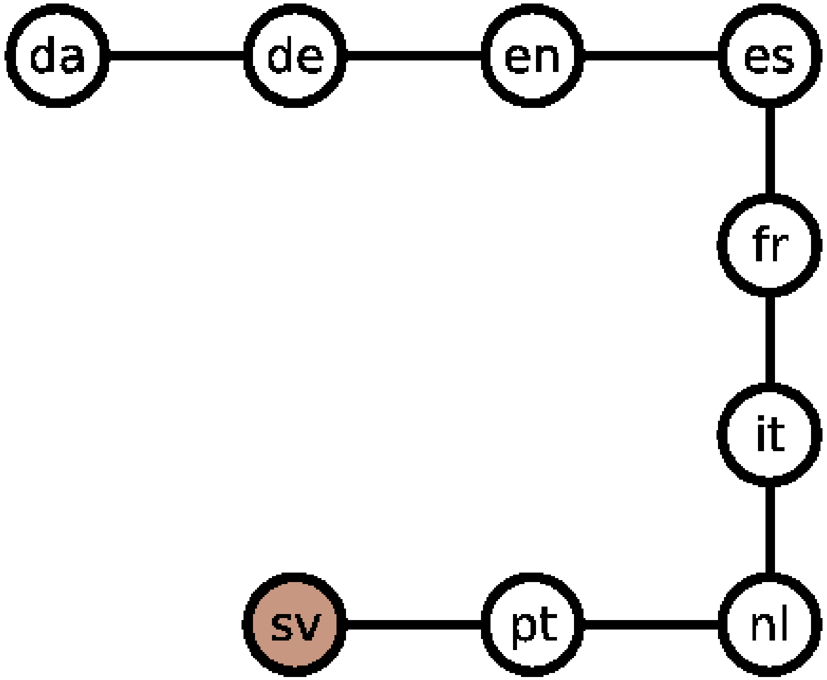
- Marginalize over X_t:

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



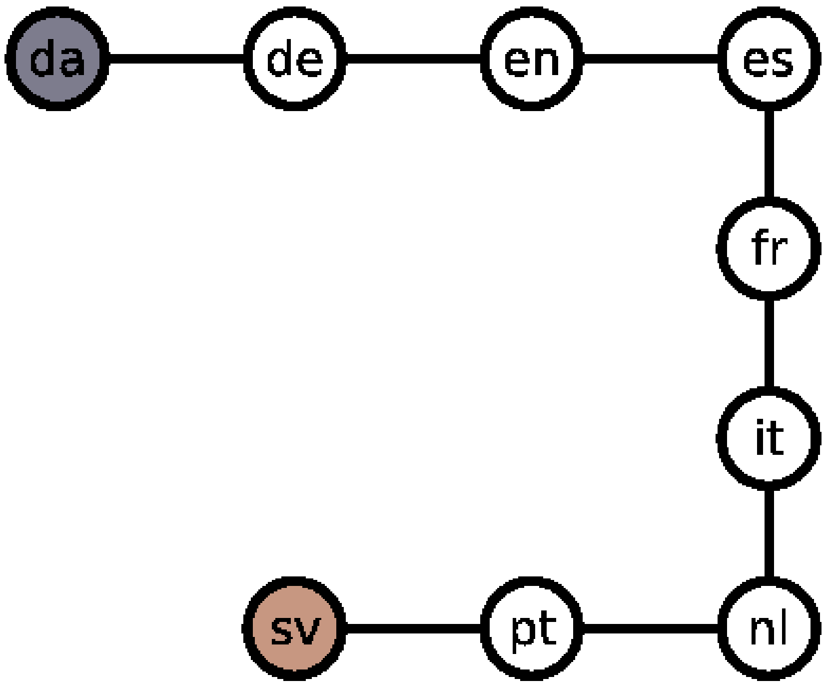
- Advantages:
  - Cost increase not exponential in depth unlike Gaussian Mixture Models (GMM) [Sudderth et al., 2003]
  - Nonparametric model learned from data unlike Gaussian BP, parametric approaches

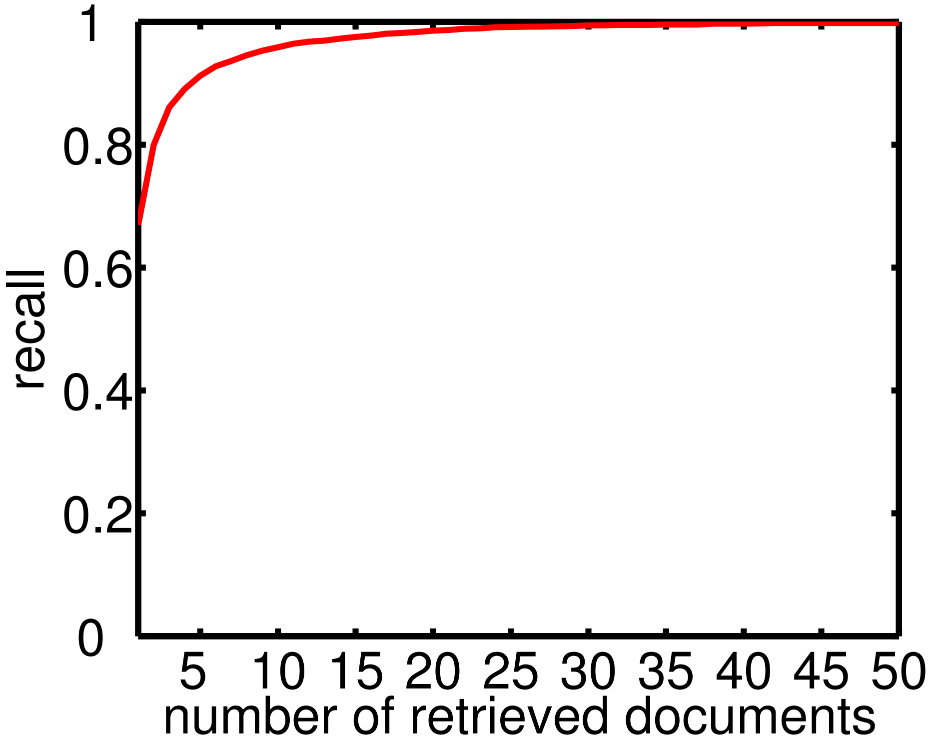
Cross-language document retrieval



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

Cross-language document retrieval

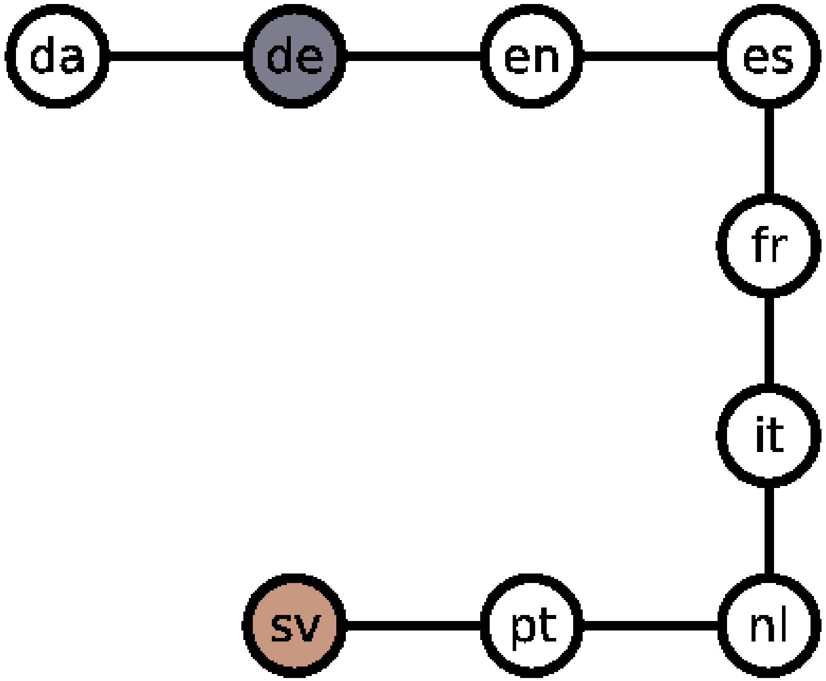


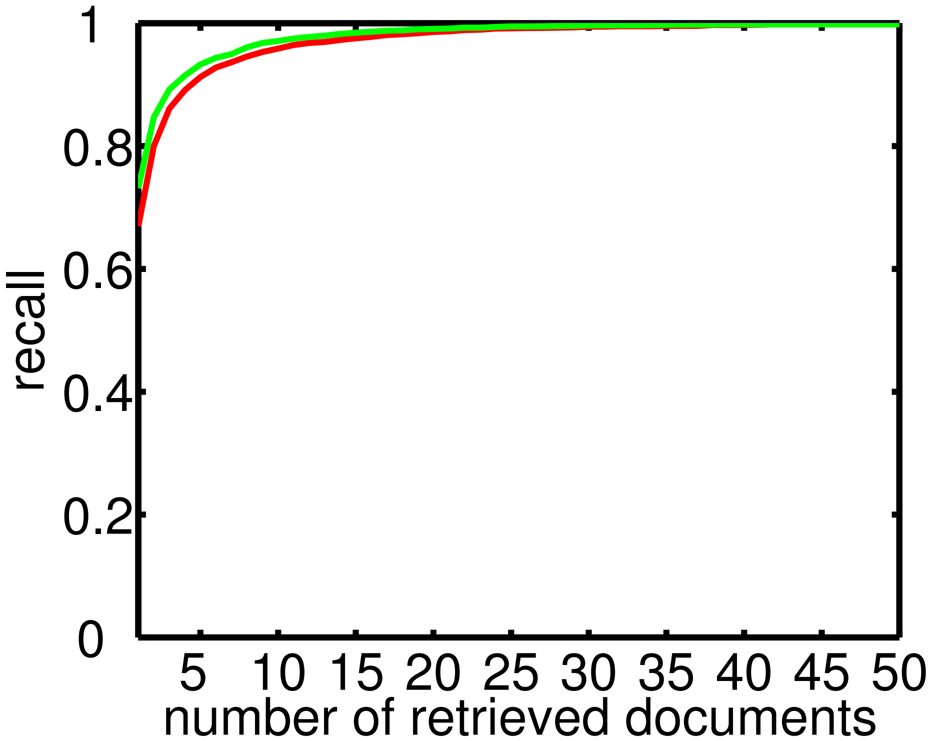


Recall score: whether target document is in set of retrieved documents

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

Cross-language document retrieval

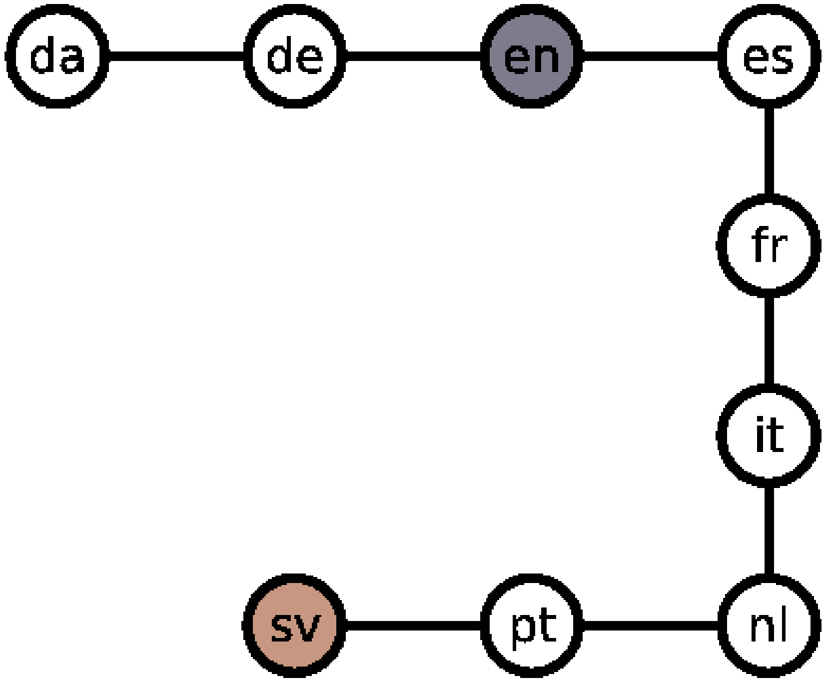


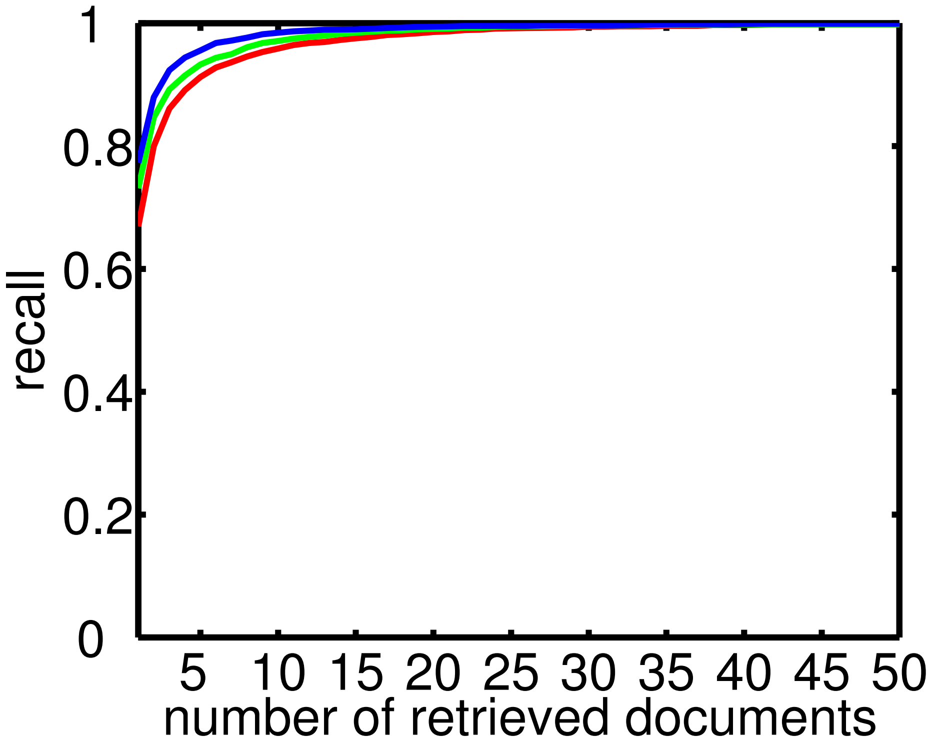


Recall score: whether target document is in set of retrieved documents

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

Cross-language document retrieval

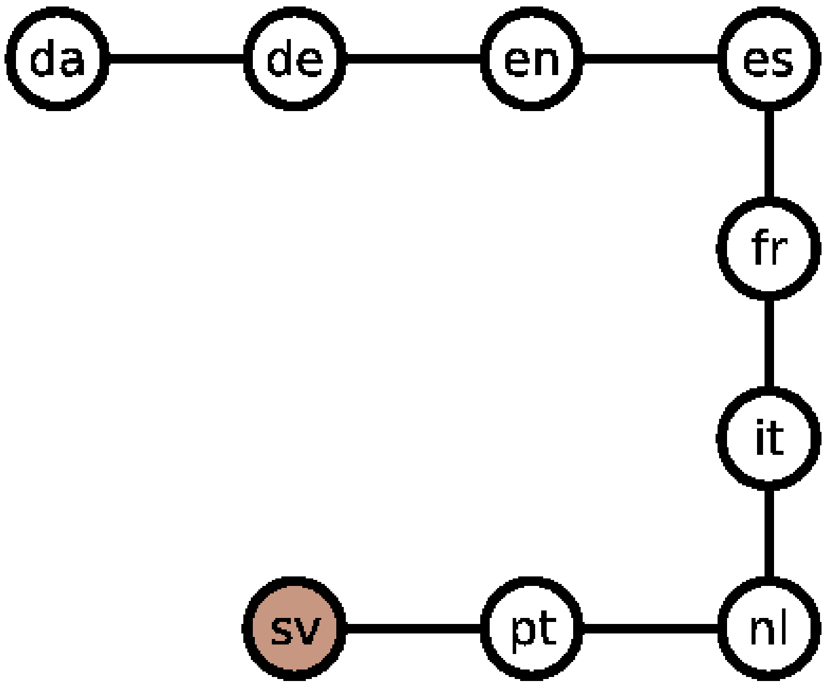


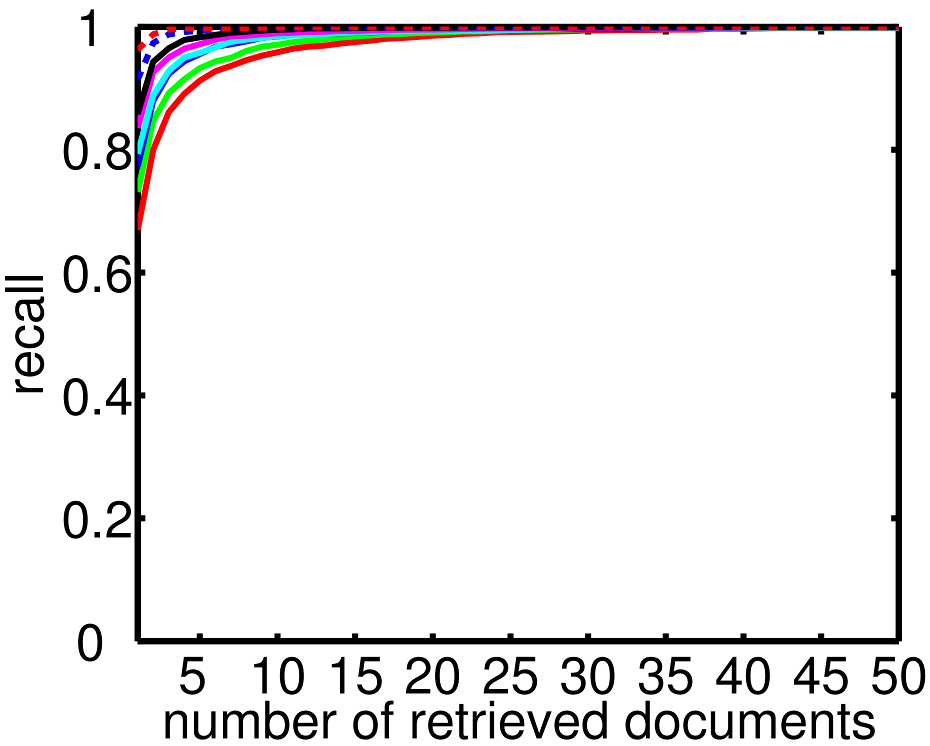


Recall score: whether target document is in set of retrieved documents

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

Cross-language document retrieval

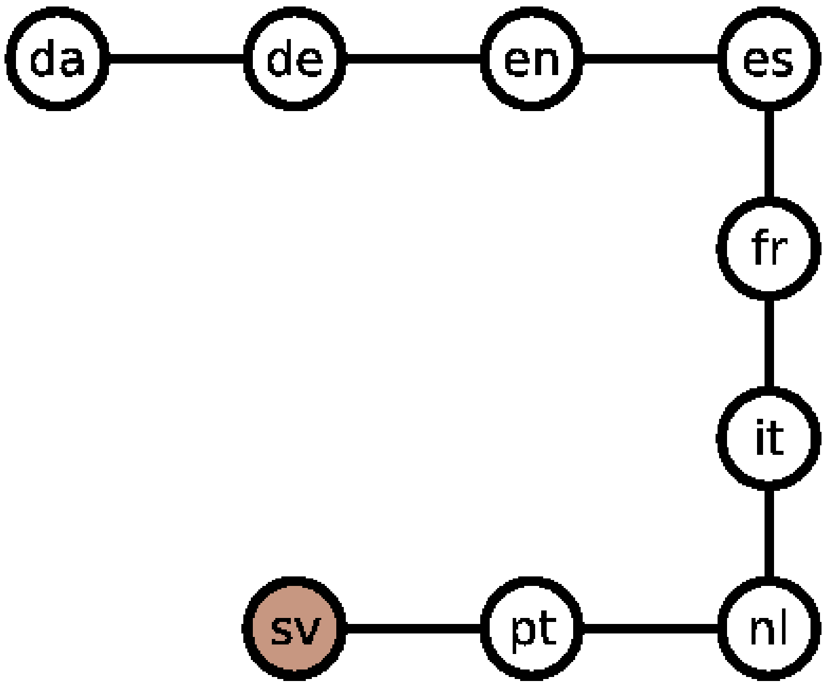


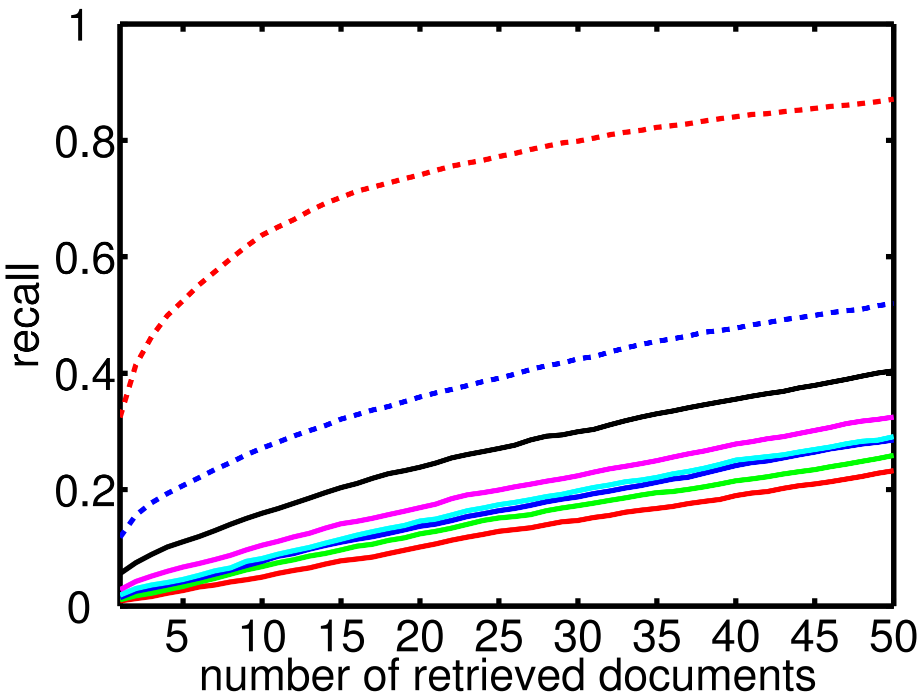


Recall score: whether target document is in set of retrieved documents

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

Cross-language document retrieval

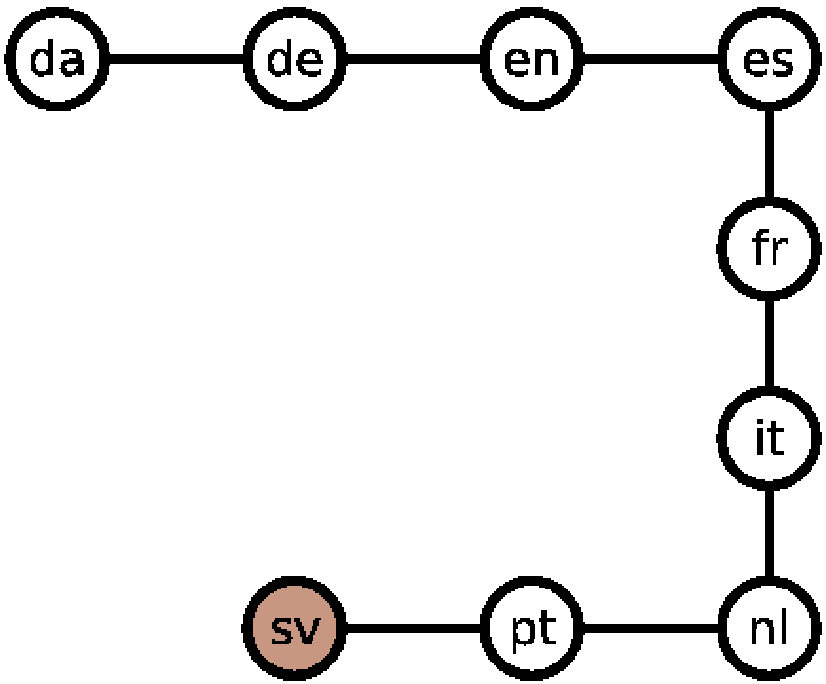


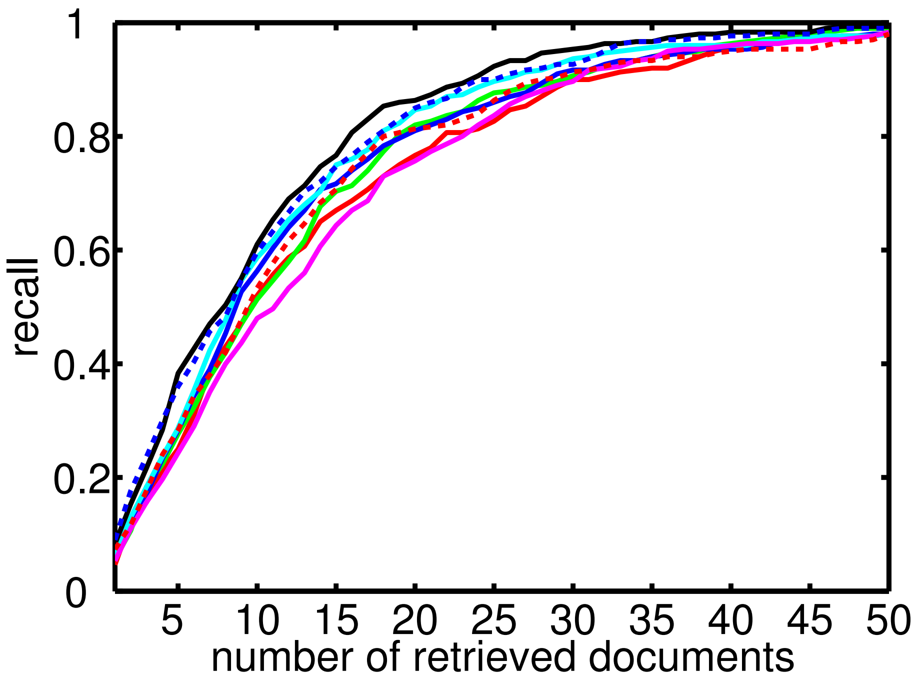


Recall score: whether target document is in set of retrieved documents

- 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



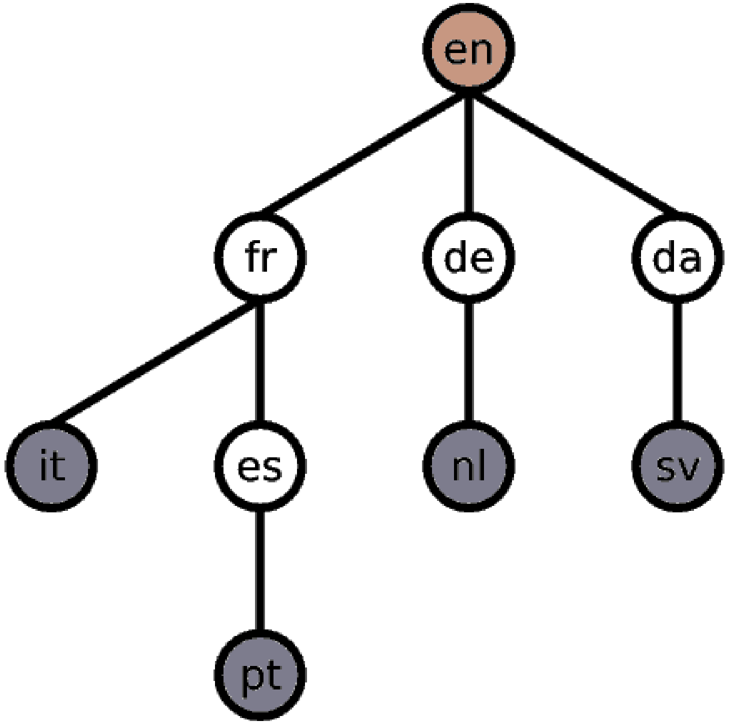


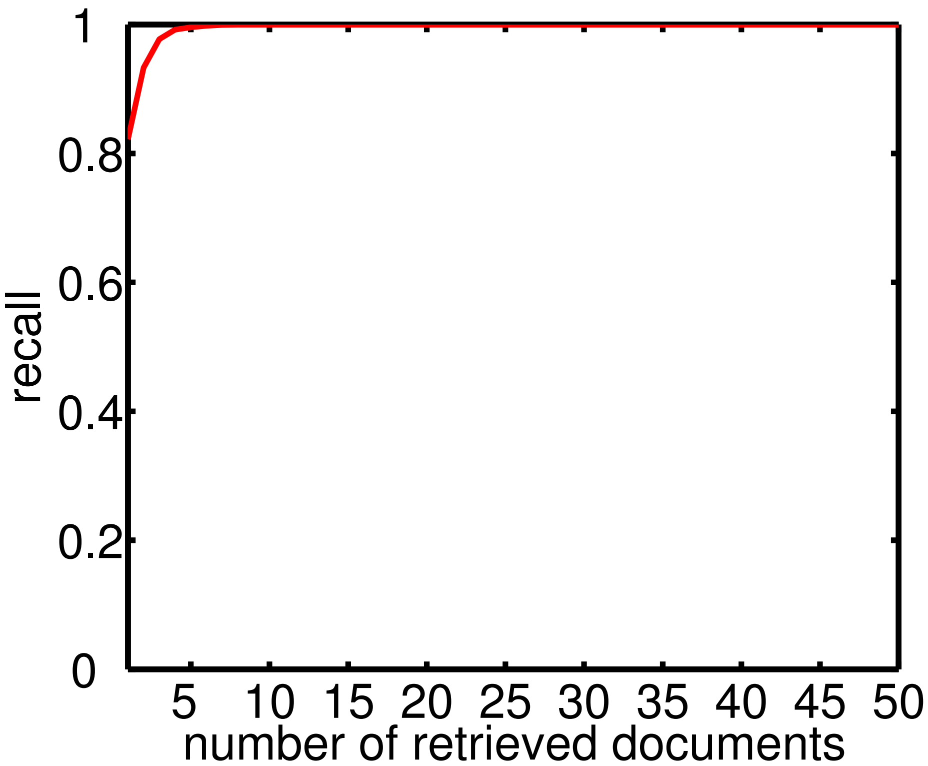
Recall score: whether target document is in set of retrieved documents

Normalized document length [Gale and Church, 1991]

- Chain length irrelevant

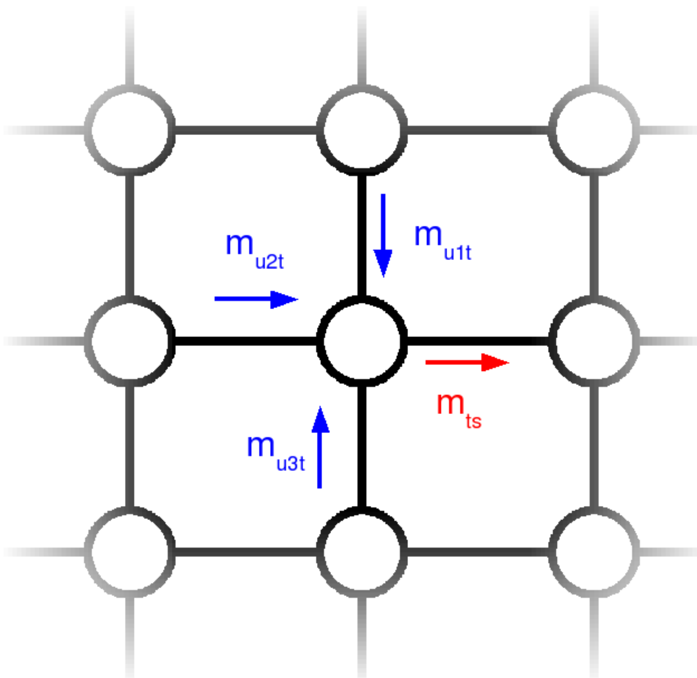
Cross-language document retrieval





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

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

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,

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

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

- BP update: can be kernelized [Song, Gretton, Bickson, Low, and Guestrin, 2011]

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

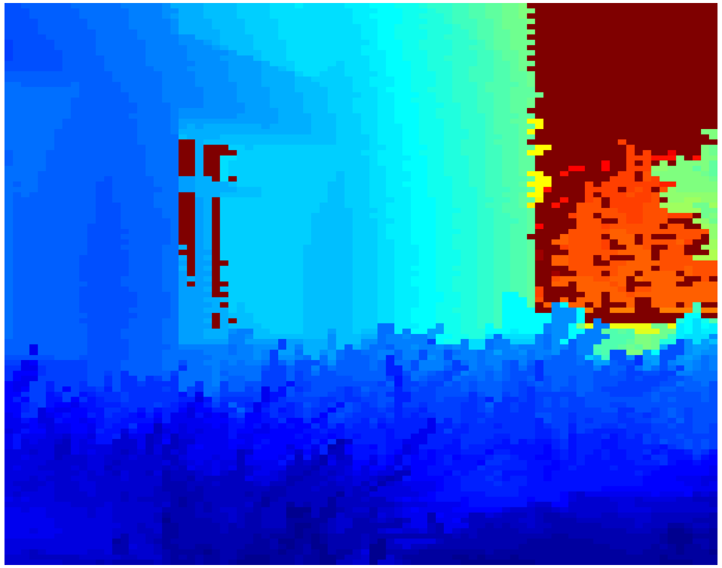
$$= \mathbf{E}_{X_t | X_s} \left[\prod_{u \in \Gamma_{t \backslash s}} m_{ut}(X_t) \; dX_t \right].$$

Application: depth from 2D images

- 3D depth reconstruction from 2D image features.

  [Song, Gretton, Bickson, Low, and Guestrin, 2011]

- 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

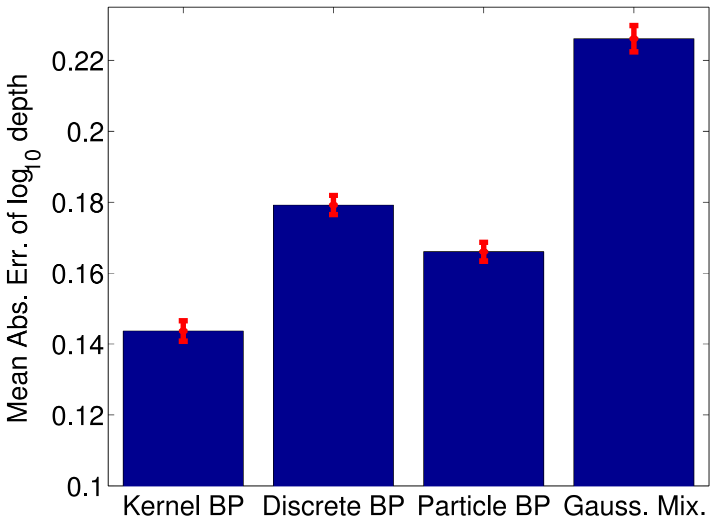
Application: depth from 2D images

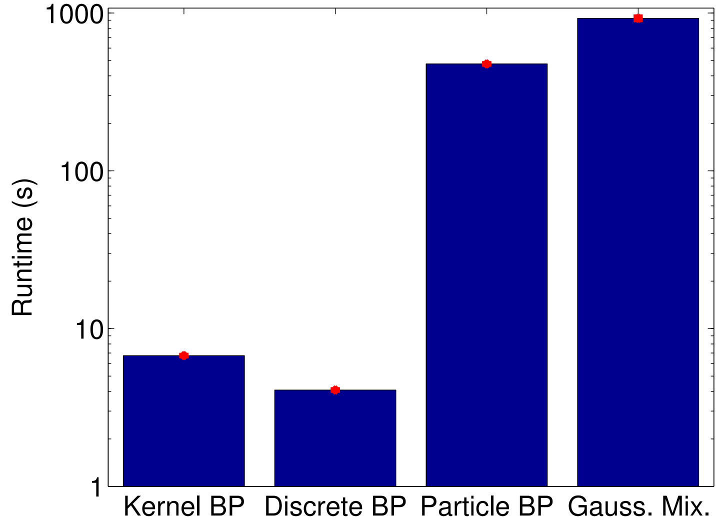
- Templatized model
  - Depth $y_i \in \mathbb{R}$ hidden var. for each image patch, in 2D grid
  - Depth linked to image features $x_i \in \mathbb{R}^{273}$
  - Potentials $\Psi(y_i, x_i)$ between features and depth unknown, as are $\Psi(y_i, y_k)$
- Kernels: Gaussian RBF on depth, linear on features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

- With RKHS distribution embeddings, compare distributions in high dimensions and on structured objects
  - Easier than density estimation
  - Works on complex high-dimensional/structured data
  - Special case: independence testing
- Kernel nonparametric message passing:
  - Exact inference on trees
  - Loopy BP on pairwise MRFs
  - Numerical integration of mixture models too expensive
  - Don't need models, just need observations!

Questions?


THE KERNEL
BREWERY LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)\right]$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j) \right]$$

- Proof:

$$\begin{aligned} \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\ &= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j) \right]$$

- Proof:

$$\begin{aligned} \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\ &= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i}^m [k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)]$$

- Proof:

$$\begin{aligned}
\|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\
&= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \\
&= \langle \mathbf{E}_{\mathbf{P}} \varphi_x, \mathbf{E}_{\mathbf{P}} \varphi_x \rangle + \ldots
\end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)\right]$$

- Proof:

$$\begin{aligned} \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\ &= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \\ &= \langle \mathbf{E}_{\mathbf{P}} \varphi_x, \mathbf{E}_{\mathbf{P}} \varphi_x \rangle + \ldots \\ &= \mathbf{E}_{\mathbf{P}} \langle \varphi_x, \varphi_{x'} \rangle + \ldots \end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathsf{P}$ and $\{y_i\}_{i=1}^m \sim \mathsf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m [k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)]$$

- Proof:

$$\begin{aligned}
\|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathsf{P}} - \mu_{\mathsf{Q}}, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} \\
&= \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle \\
&= \langle \mathbf{E}_{\mathsf{P}} \varphi_x, \mathbf{E}_{\mathsf{P}} \varphi_x \rangle + \ldots \\
&= \mathbf{E}_{\mathsf{P}} \langle \varphi_x, \varphi_{x'} \rangle + \ldots \\
&= \mathbf{E}_{\mathsf{P}} k(x, x') + \mathbf{E}_{\mathsf{Q}} k(y, y') - 2\mathbf{E}_{\mathsf{P},\mathsf{Q}} k(x, y)
\end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathsf{P}$ and $\{y_i\}_{i=1}^m \sim \mathsf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)\right]$$

- Proof:

$$\begin{aligned}
\|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathsf{P}} - \mu_{\mathsf{Q}}, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} \\
&= \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle \\
&= \langle \mathbf{E}_{\mathsf{P}} \varphi_x, \mathbf{E}_{\mathsf{P}} \varphi_x \rangle + \ldots \\
&= \mathbf{E}_{\mathsf{P}} \langle \varphi_x, \varphi_{x'} \rangle + \ldots \\
&= \mathbf{E}_{\mathsf{P}} k(x, x') + \mathbf{E}_{\mathsf{Q}} k(y, y') - 2\mathbf{E}_{\mathsf{P},\mathsf{Q}} k(x, y)
\end{aligned}$$

Then $\widehat{\mathbf{E}} k(x, x') = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(x_i, x_j)$

μ_{P} is feature map of probability

Embedding of P to feature space

- $\mu_{\mathsf{P}} := \mathbf{E}_{\mathsf{P}} \varphi_x \in \mathcal{F}$

  $$\langle \mu_{\mathsf{P}}, f \rangle = \langle \mathbf{E}_{\mathsf{P}} \varphi_x, f \rangle = E_X f(X).$$

- What does prob. feature map look like?

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

  Expectation of kernel!

- Empirical estimate:

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

$\mu_{\mathbf{P}}$ is feature map of probability

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

- $\mu_{\mathbf{P}} := \mathbf{E}_{\mathbf{P}} \varphi_x \in \mathcal{F}$

  $$\langle \mu_{\mathbf{P}}, f \rangle = \langle \mathbf{E}_{\mathbf{P}} \varphi_x, f \rangle = E_X f(X).$$

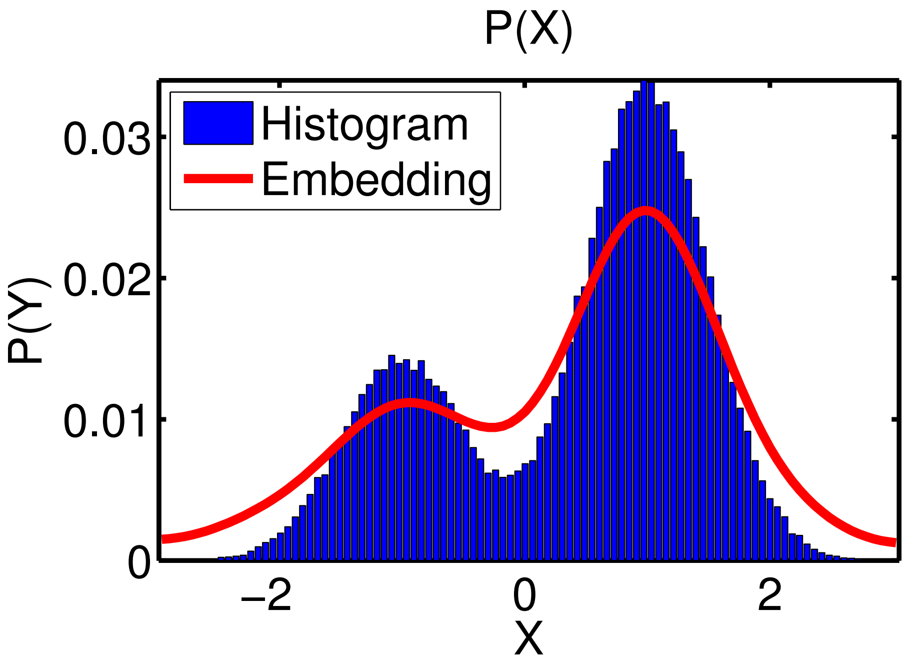
- What does prob. feature map look like?

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

  Expectation of kernel!

- Empirical estimate:

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



Bibliography

References

R. M. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, UK, 2002.

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*, pages 256–263, 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.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

T. Read and N. Cressie. *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis*. Springer-Verlag, New York, 1988.

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, 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, 2010.