Kernel approaches to covariate shift

Arthur Gretton

Carnegie Mellon University
Max Planck Institute for Biological Cybernetics

December 2009

Transfer learning and covariate shift

- Patterns $\mathcal{X}$, labels $\mathcal{Y}$
- Training: get Z_{tr} are n_{tr} pairs $(x^{\mathrm{tr}}, y^{\mathrm{tr}})$ from $\mathbf{P}_{\mathrm{tr}}$
- Test: get Z_{te} are n_{te} pairs $(x^{\mathrm{te}}, y^{\mathrm{te}})$ from $\mathbf{P}_{\mathrm{te}}$
- Predict on $\mathbf{P}_{\mathrm{te}}$ given data from $\mathbf{P}_{\mathrm{tr}}$
- Examples:
  - Medical diagnosis
  - Brain computer interfaces
  - Gene expression profiles

Transfer learning and covariate shift

- Patterns $\mathcal{X}$, labels $\mathcal{Y}$
- Training: get Z_{tr} are n_{tr} pairs $(x^{\text{tr}}, y^{\text{tr}})$ from $\mathbf{P}_{\text{tr}}$
- Test: get Z_{te} are n_{te} pairs $(x^{\text{te}}, y^{\text{te}})$ from $\mathbf{P}_{\text{te}}$
- Predict on $\mathbf{P}_{\text{te}}$ given data from $\mathbf{P}_{\text{tr}}$
- Examples:
  - Medical diagnosis
  - Brain computer interfaces
  - Gene expression profiles

Does this make sense?

Transfer learning and covariate shift

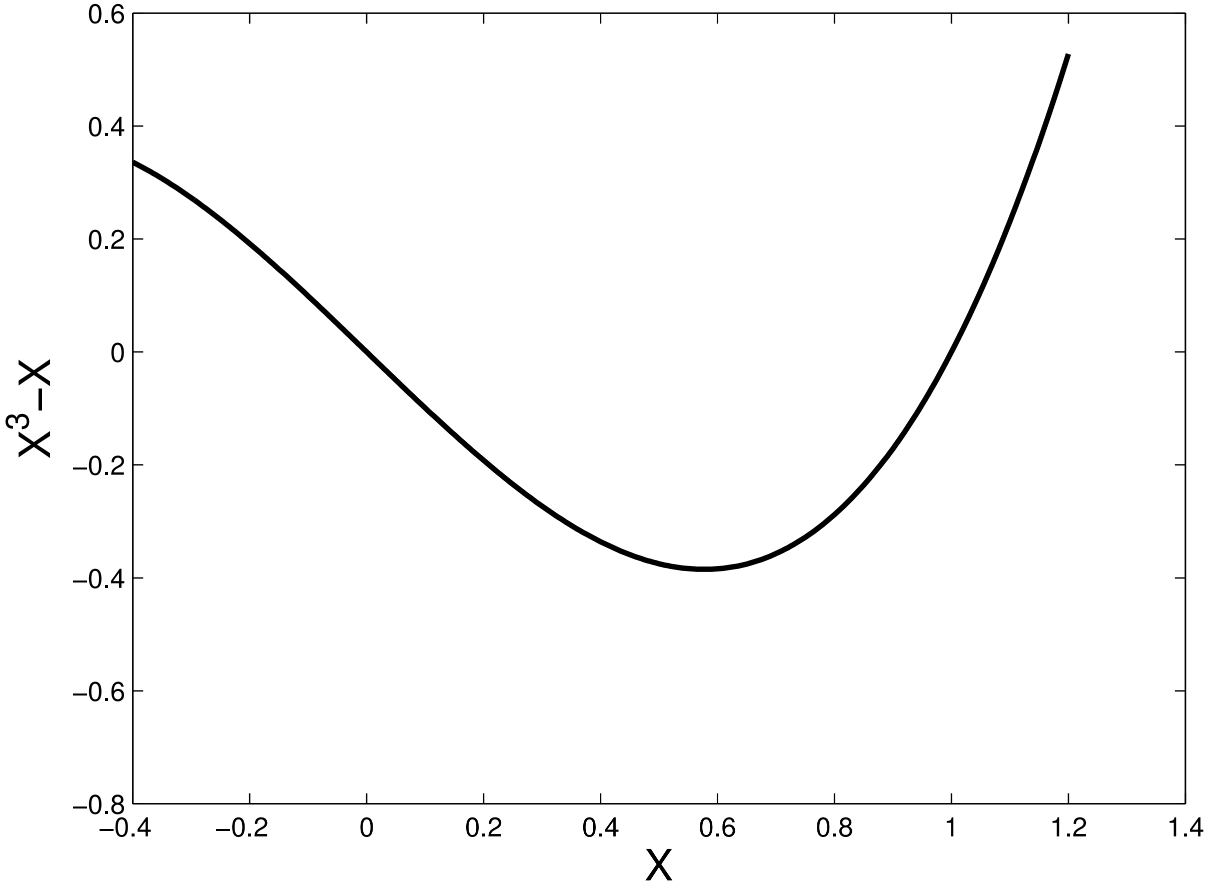
- Patterns $\mathcal{X}$, labels $\mathcal{Y}$
- Training: get Z_{tr} are n_{tr} pairs $(x^{\mathrm{tr}}, y^{\mathrm{tr}})$ from $\mathbf{P}_{\mathrm{tr}}$
- Test: get Z_{te} are n_{te} pairs $(x^{\mathrm{te}}, y^{\mathrm{te}})$ from $\mathbf{P}_{\mathrm{te}}$
- Predict on $\mathbf{P}_{\mathrm{te}}$ given data from $\mathbf{P}_{\mathrm{tr}}$
- Examples:
  - Medical diagnosis
  - Brain computer interfaces
  - Gene expression profiles
- Assumption: $\mathbf{P}_{\mathrm{tr}}(x, y) = \mathbf{P}(y|x)\mathbf{P}_{\mathrm{tr}}(x)$ and $\mathbf{P}_{\mathrm{te}}(x, y) = \mathbf{P}(y|x)\mathbf{P}_{\mathrm{te}}(x)$

Conditional probs unchanged: covariate shift

A toy example

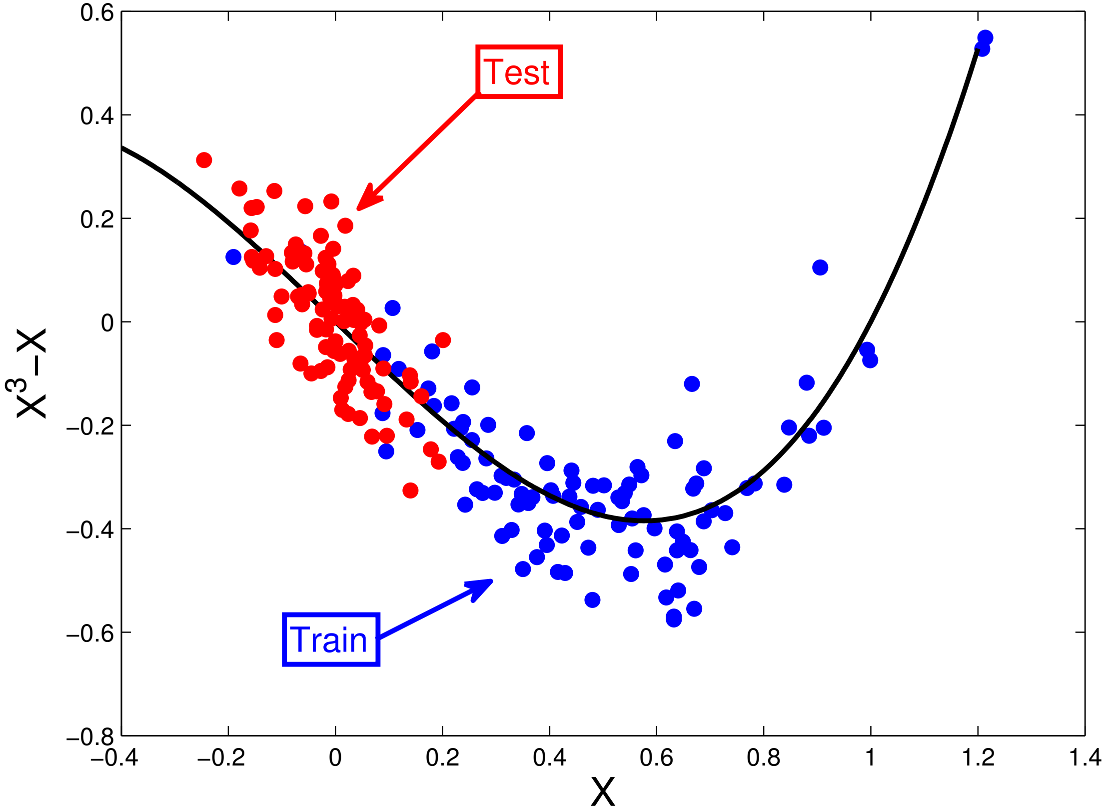
- Toy data [Shimodaira, 2000]
  - $\mathbf{P}_{\mathrm{tr}}(x) \sim \mathcal{N}(0.5, 0.5^2)$,
  - $\mathbf{P}_{\mathrm{te}}(x) \sim \mathcal{N}(0, 0.3^2)$

- $y = -x + x^3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.3^2)$
- Linear regression



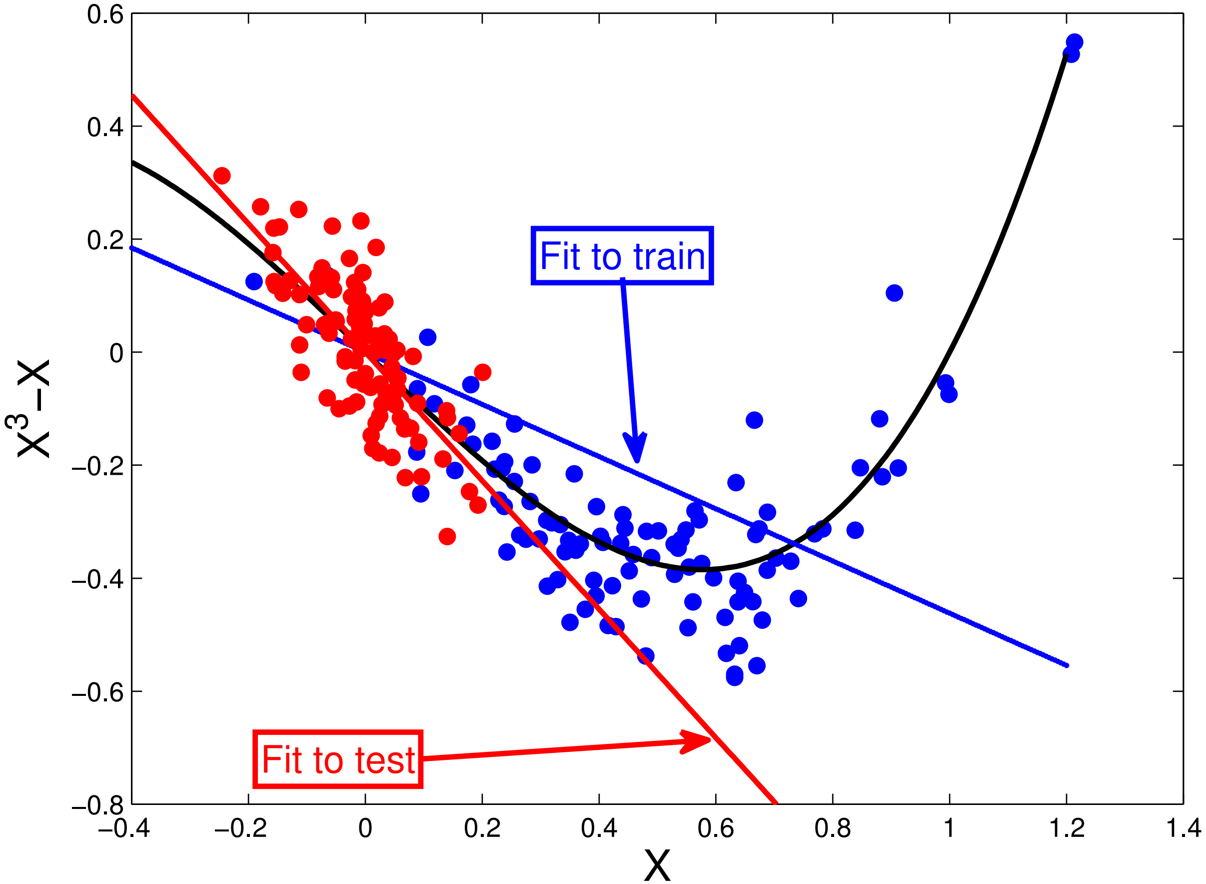
A toy example

- Toy data [Shimodaira, 2000]
  - $\mathbf{P}_{\mathrm{tr}}(x) \sim \mathcal{N}(0.5, 0.5^2)$,
  - $\mathbf{P}_{\mathrm{te}}(x) \sim \mathcal{N}(0, 0.3^2)$
- $y = -x + x^3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.3^2)$
- Linear regression



A toy example

- Toy data [Shimodaira, 2000]
  - $\mathbf{P}_{\mathrm{tr}}(x) \sim \mathcal{N}(0.5, 0.5^2)$,
  - $\mathbf{P}_{\mathrm{te}}(x) \sim \mathcal{N}(0, 0.3^2)$
- $y = -x + x^3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.3^2)$
- Linear regression



The solution procedure

- Classical setting: (regularized) expected risk

$$R[\mathbf{P}, l(x, y, \theta)] = \mathbf{E}\left[l(x, y, \theta)\right] + \lambda\Omega[\theta]$$

  - Loss $l(x, y, \theta)$, eg $-\log \mathbf{P}(y|x, \theta)$
  - Minimize over θ

The solution procedure

- Classical setting: (regularized) expected risk

$$R[\mathbf{P}, l(x,y,\theta)] = \mathbf{E}\left[l(x,y,\theta)\right] + \lambda\Omega[\theta]$$

  - Loss $l(x,y,\theta)$, eg $-\log \mathbf{P}(y|x,\theta)$
  - Minimize over θ

- Covariate shift setting:

$$\begin{aligned} R[\mathbf{P}_{\mathrm{te}}, l(x,y,\theta)] &= \mathbf{E}_{\mathbf{P}_{\mathrm{te}}}\left[l(x,y,\theta)\right] + \lambda\Omega[\theta] \\ &= \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}\left[\beta(x,y) l(x,y,\theta)\right] + \lambda\Omega[\theta] \end{aligned}$$

The solution procedure

- Classical setting: (regularized) expected risk

$$R[\mathbf{P}, l(x,y,\theta)] = \mathbf{E}\left[l(x,y,\theta)\right] + \lambda\Omega[\theta]$$

  - Loss $l(x,y,\theta)$, eg $-\log \mathbf{P}(y|x,\theta)$
  - Minimize over θ

- Covariate shift setting:

$$\begin{aligned} R[\mathbf{P}_{\mathrm{te}}, l(x,y,\theta)] &= \mathbf{E}_{\mathbf{P}_{\mathrm{te}}}\left[l(x,y,\theta)\right] + \lambda\Omega[\theta] \\ &= \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}\left[\beta(x,y) l(x,y,\theta)\right] + \lambda\Omega[\theta] \end{aligned}$$

- Importance weighting:

$$\mathbf{E}_{\mathbf{P}_{\mathrm{te}}}\left[l(x,y,\theta)\right] = \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}\Bigg[\underbrace{\frac{\mathbf{P}_{\mathrm{te}}(x,y)}{\mathbf{P}_{\mathrm{tr}}(x,y)}}_{:=\beta_{\mathrm{imp}}(x,y)} l(x,y,\theta)\Bigg] \quad \text{provided} \quad \mathbf{P}_{\mathrm{te}} \ll \mathbf{P}_{\mathrm{tr}}$$

Importance weighting

- Variance of importance weighted risk [Robert and Casella, 2004]

$$\mathrm{var}\left(l(x,y,\theta)\frac{\mathbf{P}_{\mathrm{te}}(x,y)}{\mathbf{P}_{\mathrm{tr}}(x,y)} \right)$$
$$= \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}\left[l^2(x,y,\theta)\frac{{\mathbf{P}_{\mathrm{te}}}^2(x,y)}{{\mathbf{P}_{\mathrm{tr}}}^2(x,y)} \right] - \left(\mathbf{E}_{\mathbf{P}_{\mathrm{te}}}\left[l(x,y,\theta)\right]\right)^2$$

Importance weighting

- Variance of importance weighted risk [Robert and Casella, 2004]

$$\begin{aligned}
&\mathrm{var}\left(l(x,y,\theta)\frac{\mathsf{P}_{\mathrm{te}}(x,y)}{\mathsf{P}_{\mathrm{tr}}(x,y)}\right)\\
&= \mathbf{E}_{\mathsf{P}_{\mathrm{tr}}}\left[l^2(x,y,\theta)\frac{{\mathsf{P}_{\mathrm{te}}}^2(x,y)}{{\mathsf{P}_{\mathrm{tr}}}^2(x,y)}\right] - R^2[\mathsf{P}_{\mathrm{te}},\theta,l(x,y,\theta)]\\
&= \mathbf{E}_{\mathsf{P}_{\mathrm{te}}}\left[l^2(x,y,\theta)\frac{\mathsf{P}_{\mathrm{te}}(x,y)}{\mathsf{P}_{\mathrm{tr}}(x,y)}\right] - R^2[\mathsf{P}_{\mathrm{te}},\theta,l(x,y,\theta)]
\end{aligned}$$

Importance weighting

- Variance of importance weighted risk [Robert and Casella, 2004]

$$
\begin{aligned}
&\mathrm{var}\left(l(x,y,\theta)\frac{\mathbf{P}_{\mathrm{te}}(x,y)}{\mathbf{P}_{\mathrm{tr}}(x,y)}\right) \\
&= \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}\left[l^2(x,y,\theta)\frac{\mathbf{P}_{\mathrm{te}}{}^2(x,y)}{\mathbf{P}_{\mathrm{tr}}{}^2(x,y)}\right] - R^2[\mathbf{P}_{\mathrm{te}},\theta,l(x,y,\theta)] \\
&= \mathbf{E}_{\mathbf{P}_{\mathrm{te}}}\left[l^2(x,y,\theta)\underbrace{\frac{\mathbf{P}_{\mathrm{te}}(x,y)}{\mathbf{P}_{\mathrm{tr}}(x,y)}}_{\leq B}\right] - R^2[\mathbf{P}_{\mathrm{te}},\theta,l(x,y,\theta)]
\end{aligned}
$$

Importance weighting

- Variance of importance weighted risk [Robert and Casella, 2004]

$$\begin{aligned}
&\mathrm{var}\left(l(x,y,\theta)\frac{\mathbf{P}_{\mathrm{te}}(x,y)}{\mathbf{P}_{\mathrm{tr}}(x,y)}\right)\\
&= \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}\left[l^2(x,y,\theta)\frac{\mathbf{P}_{\mathrm{te}}{}^2(x,y)}{\mathbf{P}_{\mathrm{tr}}{}^2(x,y)}\right] - R^2[\mathbf{P}_{\mathrm{te}},\theta,l(x,y,\theta)]\\
&= \mathbf{E}_{\mathbf{P}_{\mathrm{te}}}\left[l^2(x,y,\theta)\underbrace{\frac{\mathbf{P}_{\mathrm{te}}(x,y)}{\mathbf{P}_{\mathrm{tr}}(x,y)}}_{\leq B}\right] - R^2[\mathbf{P}_{\mathrm{te}},\theta,l(x,y,\theta)]
\end{aligned}$$

- $\mathbf{P}_{\mathrm{tr}}$ should have heavier tails than $\mathbf{P}_{\mathrm{te}}$

Importance weighting

- Example: kernel ridge regression
- Loss $l(x, y, \theta) = (y - \langle \Phi(x), \theta \rangle)^2$

Importance weighting

- Example: kernel ridge regression
- Loss $l(x, y, \theta) = (y - \langle \Phi(x), \theta \rangle)^2$
- Solve

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{n_{\text{tr}}} \beta_i (y_i^{\text{tr}} - \left\langle \Phi(x_i^{\text{tr}}), \theta \right\rangle)^2 + \lambda \left\| \theta \right\|^2. \tag{2}$$

Importance weighting

- Example: kernel ridge regression
- Loss $l(x, y, \theta) = (y - \langle \Phi(x), \theta \rangle)^2$
- Solve

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{n_{\text{tr}}} \beta_i (y_i^{\text{tr}} - \left\langle \Phi(x_i^{\text{tr}}), \theta \right\rangle)^2 + \lambda \|\theta\|^2. \quad (3)$$

- Equivalently:

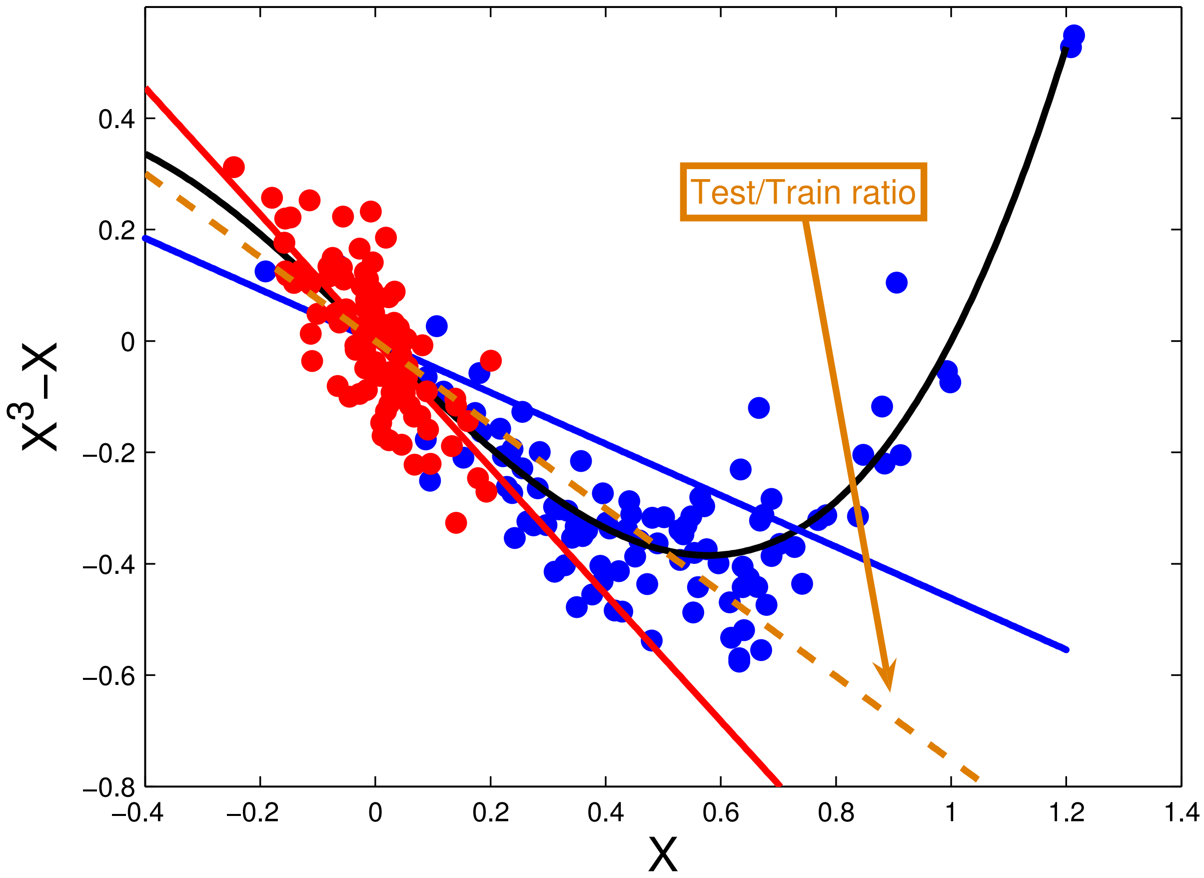
$$\underset{\alpha}{\text{minimize}} \quad (y - K\alpha)^\top \bar{\beta} (y - K\alpha) + \lambda \alpha^\top K \alpha$$

  - $\bar{\beta} = \text{diag}(\beta_1, \ldots, \beta_{n_{\text{tr}}})$
  - $K_{ij} = k(x_i^{\text{tr}}, x_j^{\text{tr}}) = \left\langle \Phi(x_i^{\text{tr}}), \Phi(x_j^{\text{tr}}) \right\rangle$
- Solution

$$\alpha = (\lambda \bar{\beta}^{-1} + K)^{-1} y$$

Importance weighting

- Ridge regression, linear kernel
- Importance weighting improves performance



Alternatives to density estimation

- Difficulties with direct density estimation
  - Empirical $\mathbf{P}_{\mathrm{tr}}$ and $\mathbf{P}_{\mathrm{te}}$ difficult for structured/high dimensional data
  - Variance can be large if empirical $\mathbf{P}_{\mathrm{te}}/\mathbf{P}_{\mathrm{tr}}$ large

Alternatives to density estimation

- Difficulties with direct density estimation
  - Empirical $\mathbf{P}_{\text{tr}}$ and $\mathbf{P}_{\text{te}}$ difficult for structured/high dimensional data
  - Variance can be large if empirical $\mathbf{P}_{\text{te}}/\mathbf{P}_{\text{tr}}$ large
- Some other reweighting approaches:
  - Minimize classification error of $\mathbf{P}_{\text{tr}}$ vs $\mathbf{P}_{\text{te}}$ [Qin, 1998, Cheng and Chu, 2004, Bickel et al., 2009]
  - Minimize KL divergence between $\mathbf{P}_{\text{tr}}$ and $\mathbf{P}_{\text{te}}$ (KLIEP) [Sugiyama et al., 2008]
  - Ratio $\mathbf{P}_{\text{te}}/\mathbf{P}_{\text{tr}}$ via least-squares function fitting [Kanamori et al., 2009]
  - Minimize Maximum Mean Discrepancy (MMD) between $\mathbf{P}_{\text{tr}}$ and $\mathbf{P}_{\text{te}}$ [Huang et al., 2007, Gretton et al., 2008]

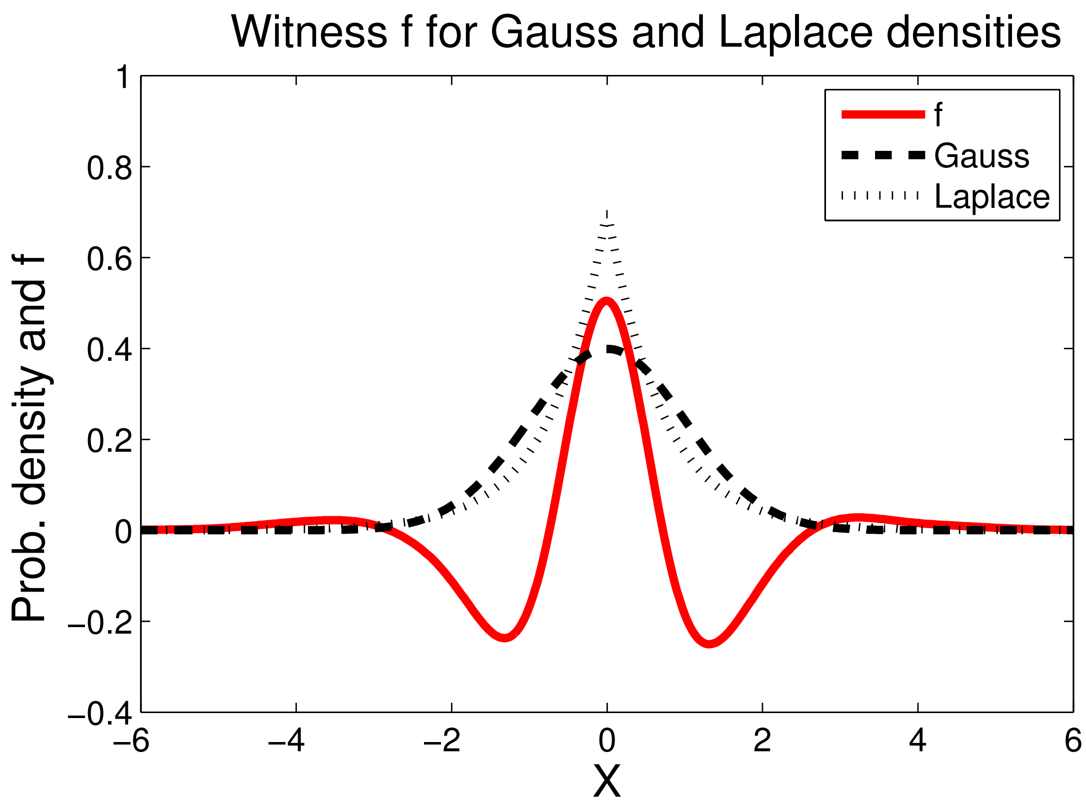
Maximum mean discrepancy

Function Showing Difference in Distributions

- Idea: avoid density estimation when comparing distributions **P** and **Q**

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

- Example: Gauss **P** vs Laplace **Q**



Function Showing Difference in Distributions

- Idea: avoid density estimation when comparing distributions $\mathbf{P}$ and $\mathbf{Q}$

  [Fortet and Mourier, 1953]

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

- 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

- Idea: avoid density estimation when comparing distributions $\mathbf{P}$ and $\mathbf{Q}$

  [Fortet and Mourier, 1953]

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

- 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 $\mathcal{F}$ [Fukumizu et al., 2008, Sriperumbudur et al., 2008]

Function Showing Difference in Distributions (2)

- $\mathcal{F}$ 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: $f(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$ for arbitrary $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, $x_i \in \mathcal{X}$.

Function Showing Difference in Distributions (2)

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

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

Function Showing Difference in Distributions (2)

- $\mathcal{F}$ 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: $f(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$ for arbitrary $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, $x_i \in \mathcal{X}$.
- Kernel is inner product between two feature maps:

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

- Evaluating functions at x

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

  - $\Phi(x)$ feature map

Function Showing Difference in Distributions

- The (kernel) MMD: [Gretton et al., 2007]

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

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

Function Showing Difference in Distributions

- The (kernel) MMD: [Gretton et al., 2007]

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

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

using

$$\mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) = \mathbf{E}_{\mathsf{P}} \left[\langle \Phi(\mathsf{x}), f \rangle_{\mathcal{F}} \right]$$

$$=: \langle \mu_x, f \rangle_{\mathcal{F}}$$

Function Showing Difference in Distributions

- The (kernel) MMD: [Gretton et al., 2007]

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

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

$$= \left(\sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2$$

using

$$\mathbf{E}_{\mathbf{P}}(f(\mathsf{x})) = \mathbf{E}_{\mathbf{P}} \left[\langle \Phi(\mathsf{x}), f \rangle_{\mathcal{F}} \right] =: \langle \mu_x, f \rangle_{\mathcal{F}}$$

Function Showing Difference in Distributions

- The (kernel) MMD: [Gretton et al., 2007]

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

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

$$= \left(\sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2$$

$$= \|\mu_x - \mu_y\|_{\mathcal{F}}^2$$

using

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

Function Showing Difference in Distributions

- The (kernel) MMD: [Gretton et al., 2007]

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

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

$$= \left(\sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2$$

$$= \|\mu_x - \mu_y\|_{\mathcal{F}}^2$$

$$= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}}$$

$$= \mathbf{E}_{\mathbf{P},\mathbf{P}} k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_{\mathbf{Q},\mathbf{Q}} k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})$$

- x' is a R.V. independent of x with distribution $\mathbf{P}$
- y' is a R.V. independent of y with distribution $\mathbf{Q}$.

Function Showing Difference in Distributions

- The (kernel) MMD: [Gretton et al., 2007]

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

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

$$= \left(\sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2$$

$$= \|\mu_x - \mu_y\|_{\mathcal{F}}^2$$

$$= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}}$$

$$= \mathbf{E}_{\mathbf{P},\mathbf{P}} k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_{\mathbf{Q},\mathbf{Q}} k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})$$

- x' is a R.V. independent of x with distribution $\mathbf{P}$
- y' is a R.V. independent of y with distribution $\mathbf{Q}$.

- Mean map:

$$\mu_x := \mathbf{E}_{\mathbf{P}} \Phi(\mathsf{x}) = \int k(\cdot, x)\, d\mathbf{P}(x)$$

Transfer learning using maximum mean discrepancy

Transfer learning by KMM

Kernel mean matching (KMM)

Transfer learning by KMM

Kernel mean matching (KMM)

- Reweight training points so feature means match

$$\underset{\beta}{\text{minimize}} \quad \left\| \mu(\mathsf{P}_{\text{te}}) - \mathbf{E}_{\mathsf{P}_{\text{tr}}}\left[\beta(x)\Phi(x)\right] \right\|$$
$$\text{subject to } \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathsf{P}_{\text{tr}}}\left[\beta(x)\right] = 1.$$

- If $\mathsf{P}_{\text{te}} \ll \mathsf{P}_{\text{tr}}$, characteristic kernel, solution is $\mathsf{P}_{\text{te}}(x) = \beta_{\text{imp}}(x)\mathsf{P}_{\text{tr}}(x)$

Transfer learning by KMM

Kernel mean matching (KMM)

- Reweight training points so feature means match

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \left\| \mu(\mathbf{P}_{\text{te}}) - \mathbf{E}_{\mathbf{P}_{\text{tr}}} \left[\beta(x) \Phi(x) \right] \right\| \\ \text{subject to } & \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathbf{P}_{\text{tr}}} \left[\beta(x) \right] = 1. \end{aligned}$$

- If $\mathbf{P}_{\text{te}} \ll \mathbf{P}_{\text{tr}}$, characteristic kernel, solution is $\mathbf{P}_{\text{te}}(x) = \beta_{\text{imp}}(x) \mathbf{P}_{\text{tr}}(x)$
- What about non-characteristic?

Transfer learning by KMM

Kernel mean matching (KMM)

- Reweight training points so feature means match

$$\underset{\beta}{\text{minimize}} \quad \|\mu(\mathbf{P}_{\text{te}}) - \mathbf{E}_{\mathbf{P}_{\text{tr}}}[\beta(x)\Phi(x)]\|$$
$$\text{subject to } \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathbf{P}_{\text{tr}}}[\beta(x)] = 1.$$

- If $\mathbf{P}_{\text{te}} \ll \mathbf{P}_{\text{tr}}$, characteristic kernel, solution is $\mathbf{P}_{\text{te}}(x) = \beta_{\text{imp}}(x)\mathbf{P}_{\text{tr}}(x)$
- Empirical:

$$\min_{\beta} \left\| \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \beta_i \Phi(x_i^{\text{tr}}) - \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \Phi(x_i^{\text{te}}) \right\|^2 = \frac{1}{{n_{\text{tr}}}^2} \beta^\top K \beta - \frac{2}{{n_{\text{tr}}}^2} \kappa^\top \beta + \text{const.}$$

Transfer learning by KMM

Kernel mean matching (KMM)

- Reweight training points so feature means match

$$\underset{\beta}{\text{minimize}} \quad \| \mu(\mathbf{P}_{\text{te}}) - \mathbf{E}_{\mathbf{P}_{\text{tr}}}[\beta(x)\Phi(x)] \|$$
$$\text{subject to } \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathbf{P}_{\text{tr}}}[\beta(x)] = 1.$$

- If $\mathbf{P}_{\text{te}} \ll \mathbf{P}_{\text{tr}}$, characteristic kernel, solution is $\mathbf{P}_{\text{te}}(x) = \beta_{\text{imp}}(x)\mathbf{P}_{\text{tr}}(x)$
- Empirical:

$$\min_{\beta} \left\| \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \beta_i \Phi(x_i^{\text{tr}}) - \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \Phi(x_i^{\text{te}}) \right\|^2 = \frac{1}{{n_{\text{tr}}}^2} \beta^\top K \beta - \frac{2}{{n_{\text{tr}}}^2} \kappa^\top \beta + \text{const}.$$

$$\text{subject to } \beta_i \in [0, B] \quad \text{and} \quad \left| \sum_{i=1}^{n_{\text{tr}}} \beta_i - n_{\text{tr}} \right| \leq \sqrt{n_{\text{tr}}}\epsilon.$$

Transfer learning by KMM

Kernel mean matching (KMM)

- Reweight training points so feature means match

$$\underset{\beta}{\text{minimize}} \quad \|\mu(\mathbf{P}_{\mathrm{te}}) - \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}[\beta(x)\Phi(x)]\|$$
$$\text{subject to } \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathbf{P}_{\mathrm{tr}}}[\beta(x)] = 1.$$

- If $\mathbf{P}_{\mathrm{te}} \ll \mathbf{P}_{\mathrm{tr}}$, characteristic kernel, solution is $\mathbf{P}_{\mathrm{te}}(x) = \beta_{\mathrm{imp}}(x)\mathbf{P}_{\mathrm{tr}}(x)$
- Empirical:

$$\min_{\beta} \left\| \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \beta_i \Phi(x_i^{\mathrm{tr}}) - \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \Phi(x_i^{\mathrm{te}}) \right\|^2 = \frac{1}{n_{\mathrm{tr}}{}^2} \beta^\top K \beta - \frac{2}{n_{\mathrm{tr}}{}^2} \kappa^\top \beta + \text{const.}$$

$$\text{subject to } \beta_i \in [0, B] \quad \text{and} \quad \underbrace{\left| \sum_{i=1}^{n_{\mathrm{tr}}} \beta_i - n_{\mathrm{tr}} \right| \leq \sqrt{n_{\mathrm{tr}}}\epsilon.}_{\left[\frac{1}{\sqrt{n_{\mathrm{tr}}}} \sum_i \beta_{\mathrm{imp}}(x_i^{\mathrm{tr}}) - \sqrt{n_{\mathrm{tr}}}\right] \ \underset{D}{\to} \ \mathcal{N}(0,\sigma^2)}$$

Transfer learning by KMM

- What if given β_{imp}: finite sample effects?
- Assume $k(x, x) \leq R^2$ for all $x \in \mathcal{X}$.

Transfer learning by KMM

- What if given β_{imp}: finite sample effects?
- Assume $k(x,x) \leq R^2$ for all $x \in \mathcal{X}$.
- With probability at least $1-\delta$,

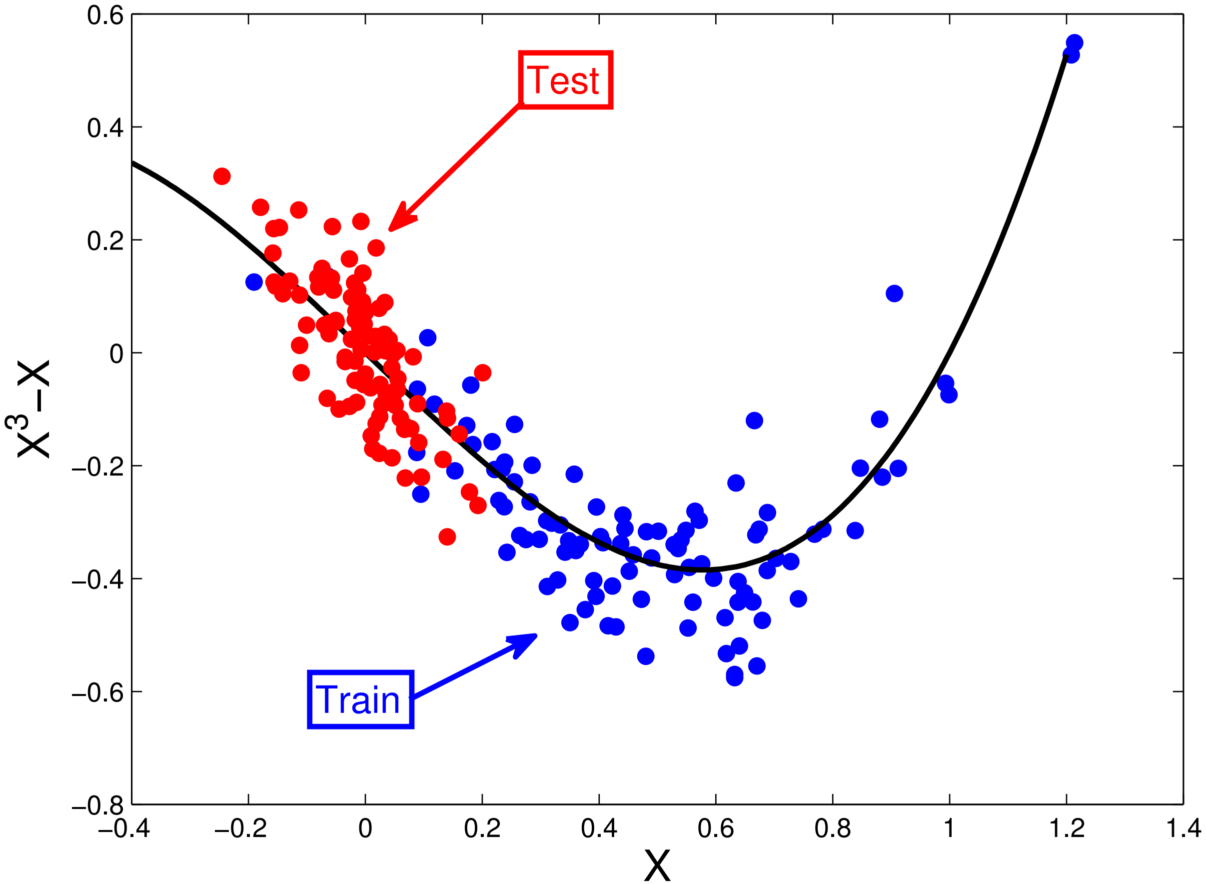
$$\left\| \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \beta_{\mathrm{imp}}(x_i^{\mathrm{tr}}) \Phi(x_i^{\mathrm{tr}}) - \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \Phi(x_i^{\mathrm{te}}) \right\| \leq \left(1 + \sqrt{2 \log 2/\delta}\right) R \sqrt{B^2/n_{\mathrm{tr}} + 1/n_{\mathrm{te}}}.$$

- Still (potentially) high variance for large B.

- Convergence of KMM procedure: [Cortes et al., 2008]

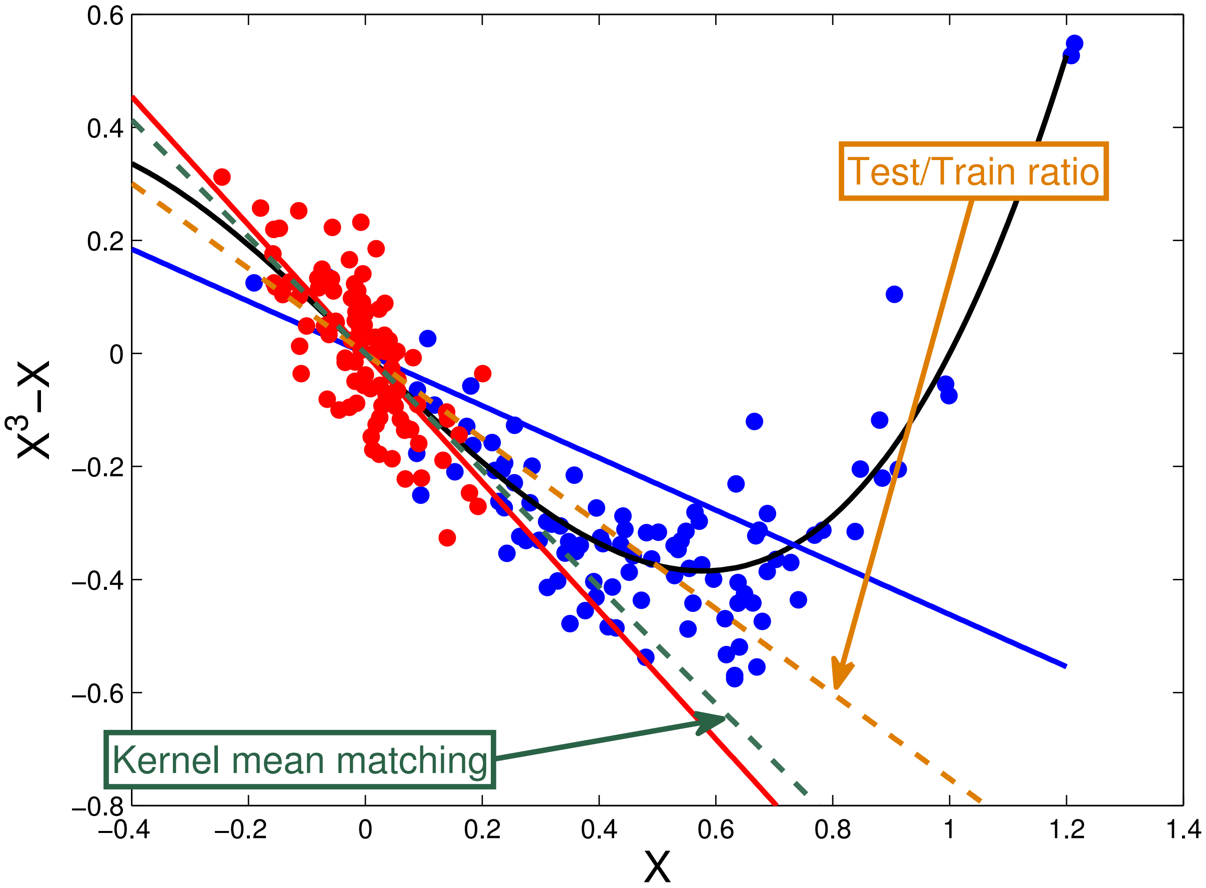
Transfer learning by KMM

- Compare KMM and importance sampling



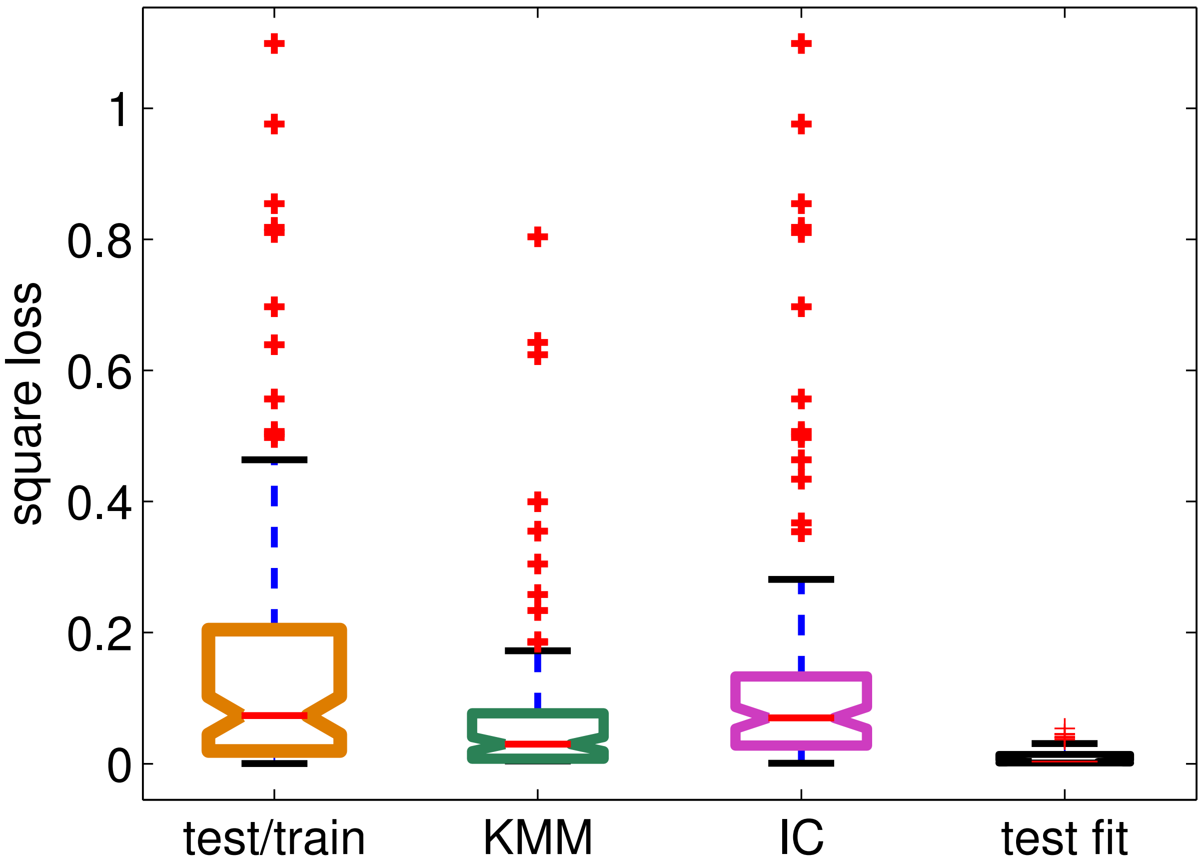
Transfer learning by KMM

- Compare KMM and importance sampling



Transfer learning by KMM

- Compare KMM and importance sampling



IC method due to [Shimodaira, 2000]

Reweighting by classification

- Use train/test classification error to reweight [Qin, 1998, Cheng and Chu, 2004, Bickel et al., 2009]
- $\mathbf{P}\left(S|x^{\mathrm{tr}}, x^{\mathrm{te}}, \theta_{\mathrm{shift}}\right)$ classifies training $(s = 1)$ vs test $(s = 0)$

Reweighting by classification

- Use train/test classification error to reweight [Qin, 1998, Cheng and Chu, 2004, Bickel et al., 2009]
- $\mathbf{P}\left(S|x^{\mathrm{tr}}, x^{\mathrm{te}}, \theta_{\mathrm{shift}}\right)$ classifies training $(s=1)$ vs test $(s=0)$
- Estimate importance ratio:

$$\frac{\mathbf{P}_{\mathrm{te}}(x_i^{\mathrm{tr}})}{\mathbf{P}_{\mathrm{tr}}(x_i^{\mathrm{tr}})} = \frac{\mathbf{P}(s=1)}{\mathbf{P}(s=0)}\left(\mathbf{P}^{-1}\left(s=1|x_i^{\mathrm{tr}}, \theta_{\mathrm{shift}}\right) - 1\right)$$

- Learn two classifiers: train vs test and covariate to label

Reweighting by classification

- Use train/test classification error to reweight [Qin, 1998, Cheng and Chu, 2004, Bickel et al., 2009]
- $\mathbf{P}\left(S|x^{\text{tr}}, x^{\text{te}}, \theta_{\text{shift}}\right)$ classifies training $(s = 1)$ vs test $(s = 0)$
- Estimate importance ratio:

$$\frac{\mathbf{P}_{\text{te}}(x_i^{\text{tr}})}{\mathbf{P}_{\text{tr}}(x_i^{\text{tr}})} = \frac{\mathbf{P}(s=1)}{\mathbf{P}(s=0)}\left(\mathbf{P}^{-1}\left(s=1|x_i^{\text{tr}}, \theta_{\text{shift}}\right) - 1\right)$$

- Learn two classifiers: train vs test and covariate to label
- Single joint optimization? [Bickel et al., 2009]

$$\max_{\theta_{\text{shift}}, \theta_{\text{learn}}} \mathbf{P}\left(y^{\text{tr}}|S, x^{\text{tr}}, \theta_{\text{shift}}, \theta_{\text{learn}}\right) \mathbf{P}\left(S|x^{\text{tr}}, x^{\text{te}}, \theta_{\text{shift}}\right) \mathbf{P}(\theta_{\text{shift}})\mathbf{P}(\theta_{\text{learn}})$$

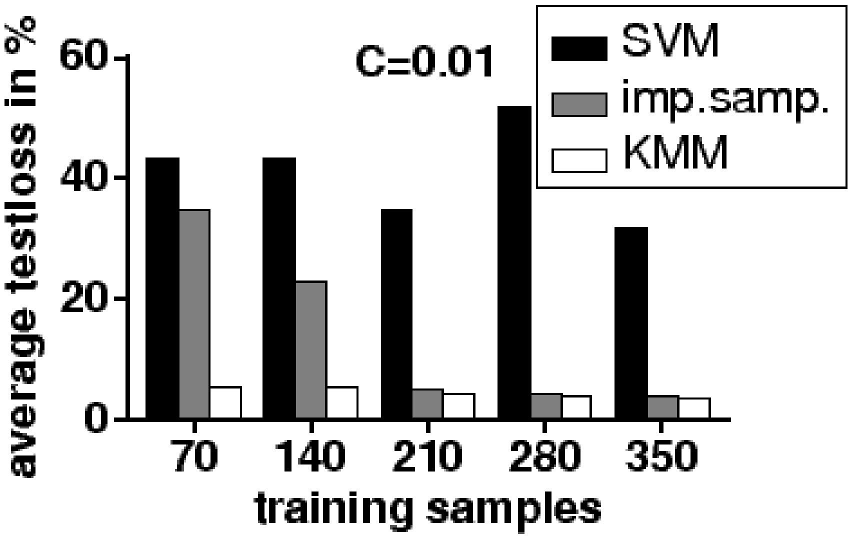
Experiments

Breast Cancer data

- Gaussian kernel $\exp(-|x_i - x_j|^2/(2\sigma))$ for KMM and SVN, $\sigma = 5$
- Performance vs C
  - Small C $\rightarrow$ prioritize smoothness
- Selection procedure:
  - Random training/test split
  - Training set from 10% - 50% of test
  - $P(s_i = 1|x_i) \propto \exp(-0.05\|x_i - \overline{x}\|^2)$

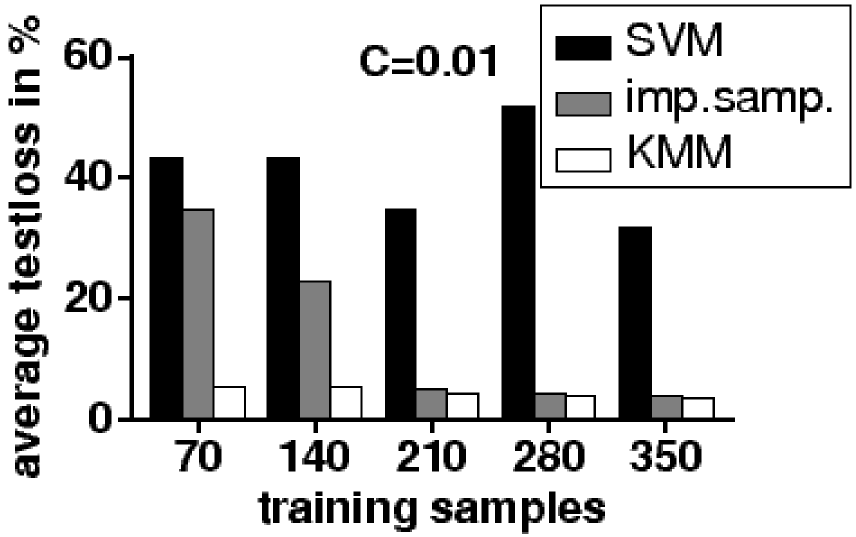
Breast Cancer data

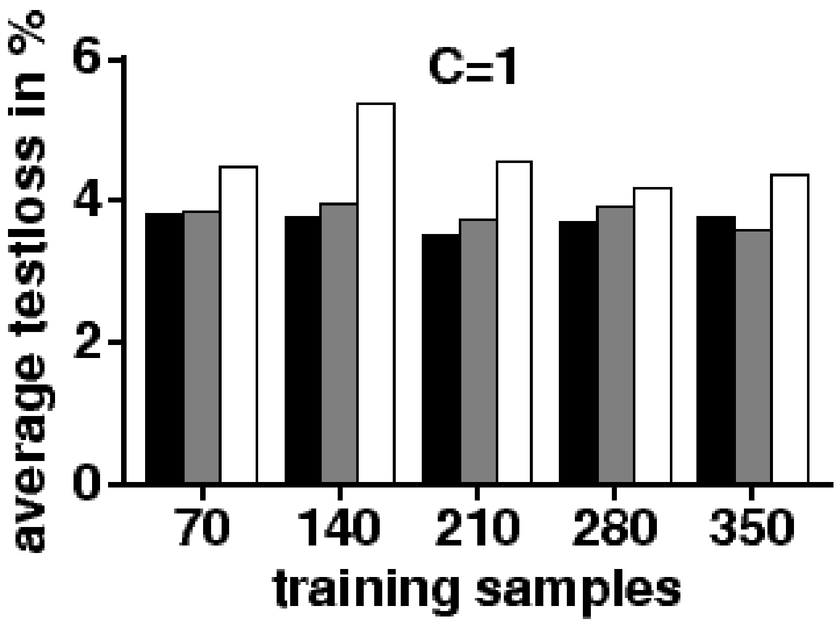
- Reweighting greatly improves performance
- KMM outperforms IS at small sample sizes



Breast Cancer data

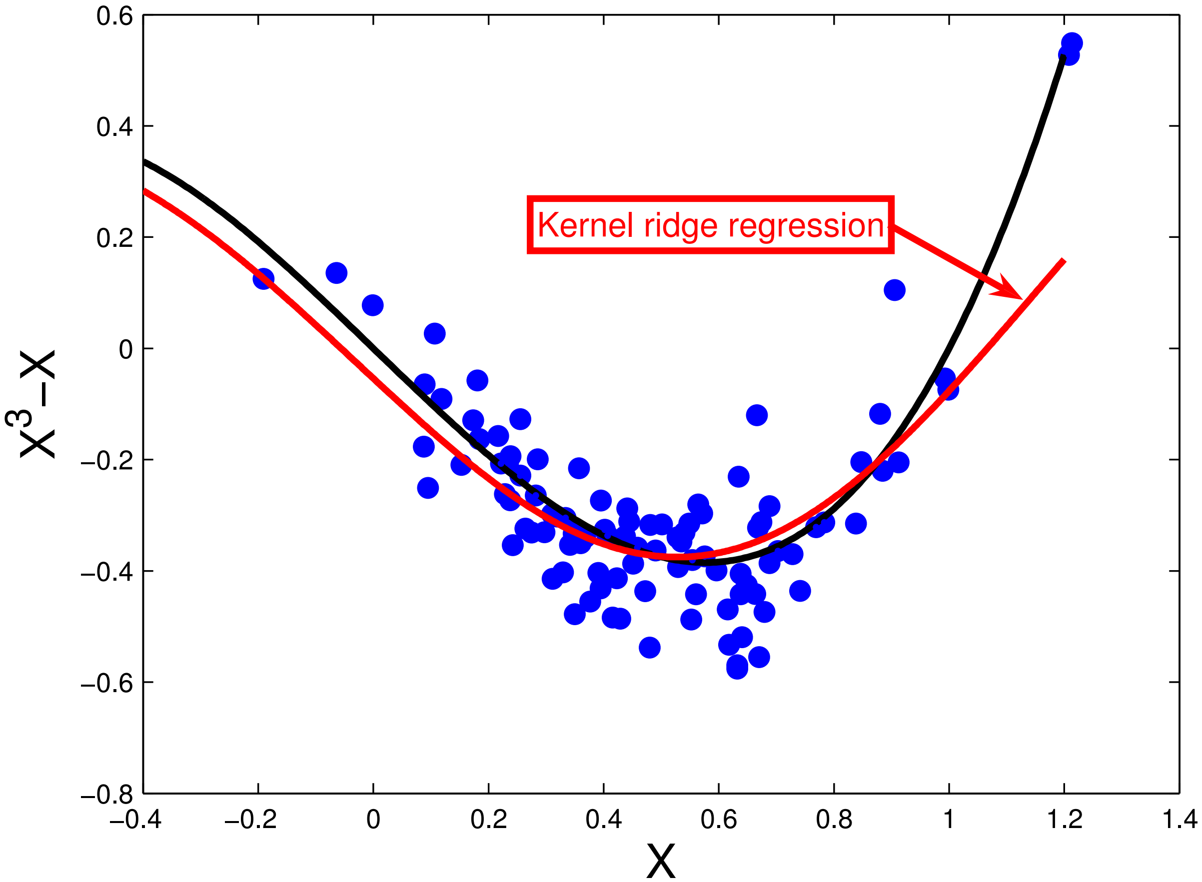
- KMM slightly decreases performance
- IS does not help





Toy example revisited

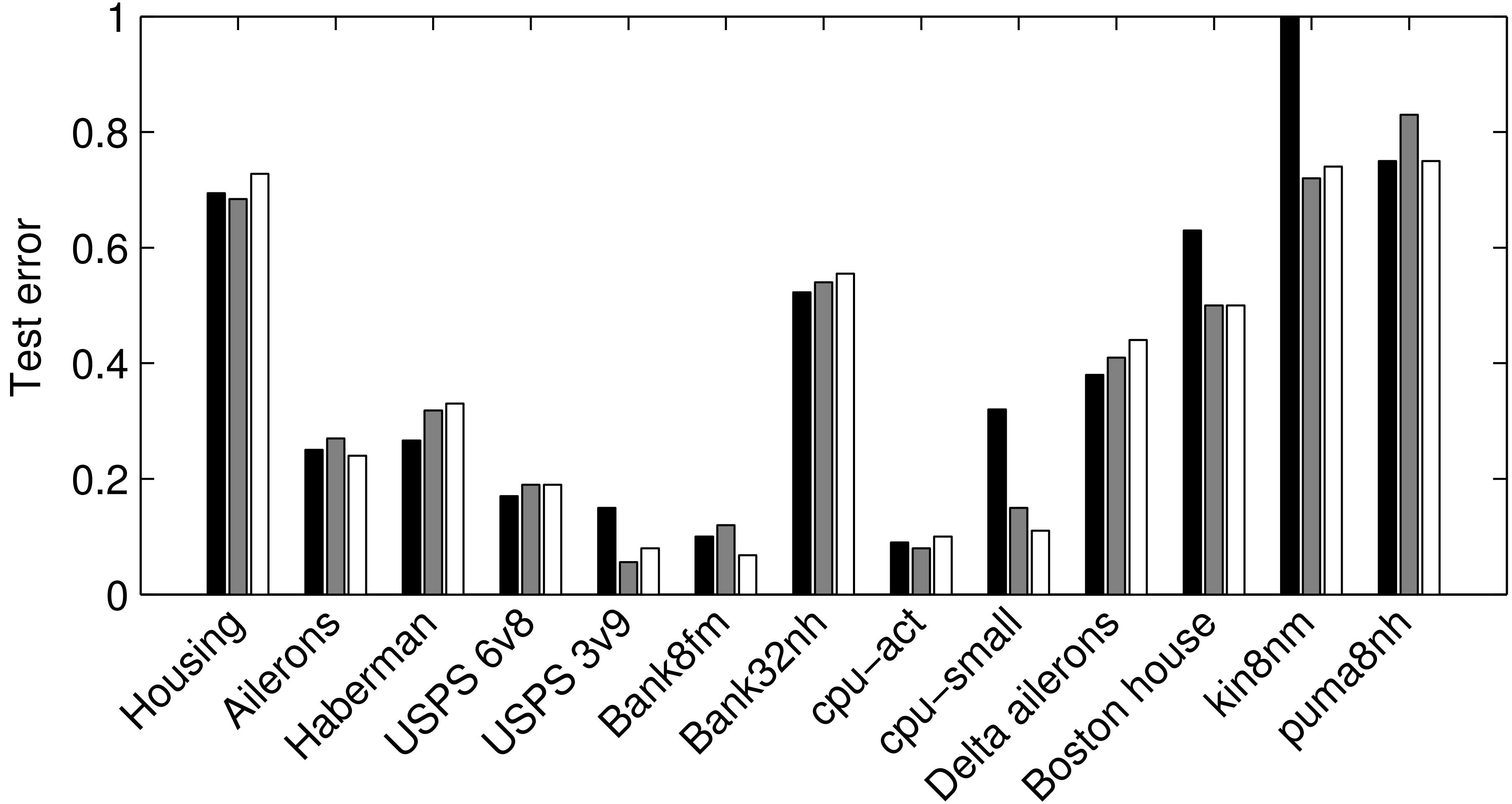
- Kernel ridge regression result



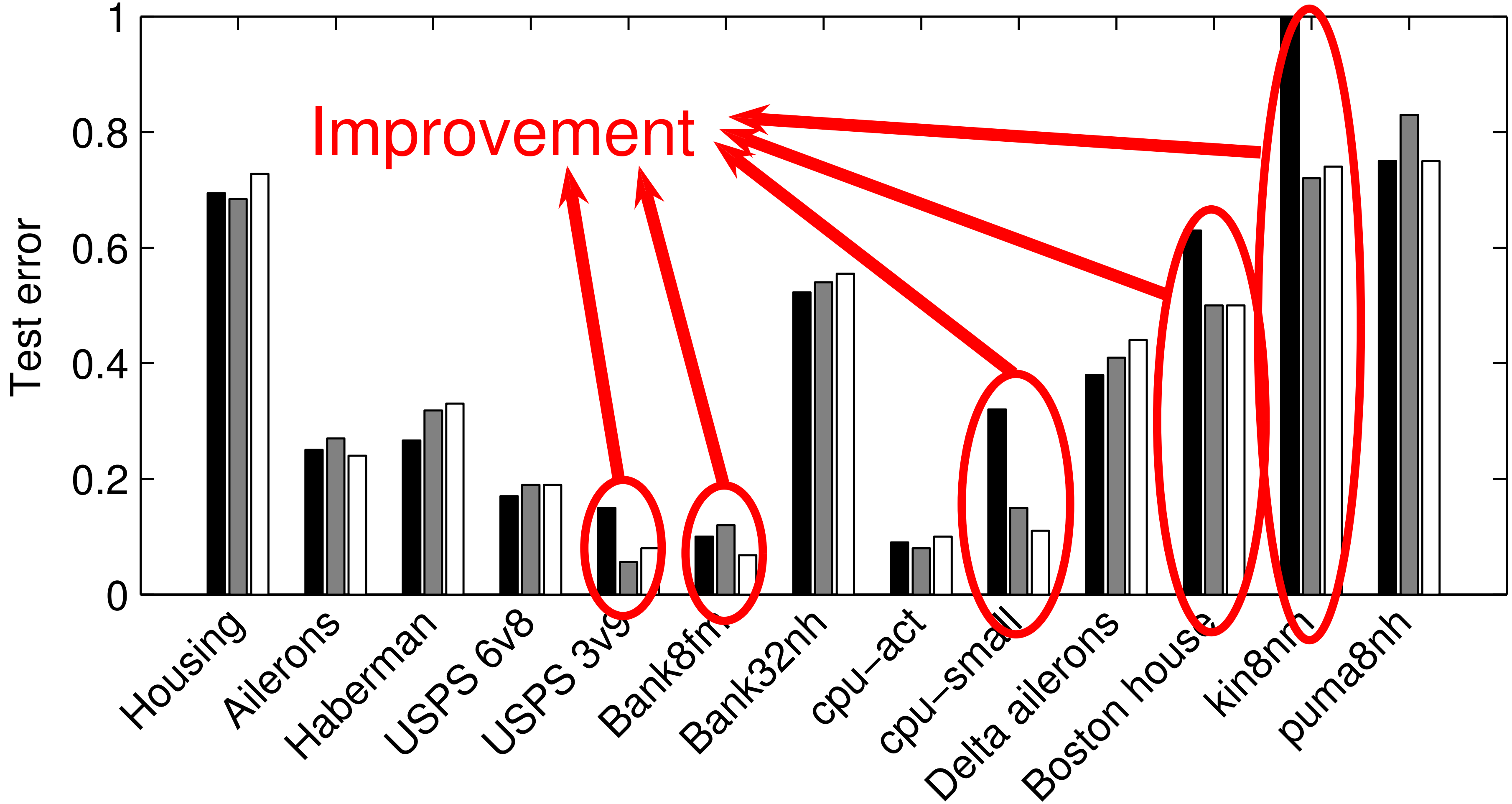
Large scale experiments

- Regression and classification
- Sampling scheme: training data missing at random
  - Sampling by Gaussian distribution on first principal component
- Cross validate on unweighted training set for C and σ
- Same σ for classifier/regressor and KMM

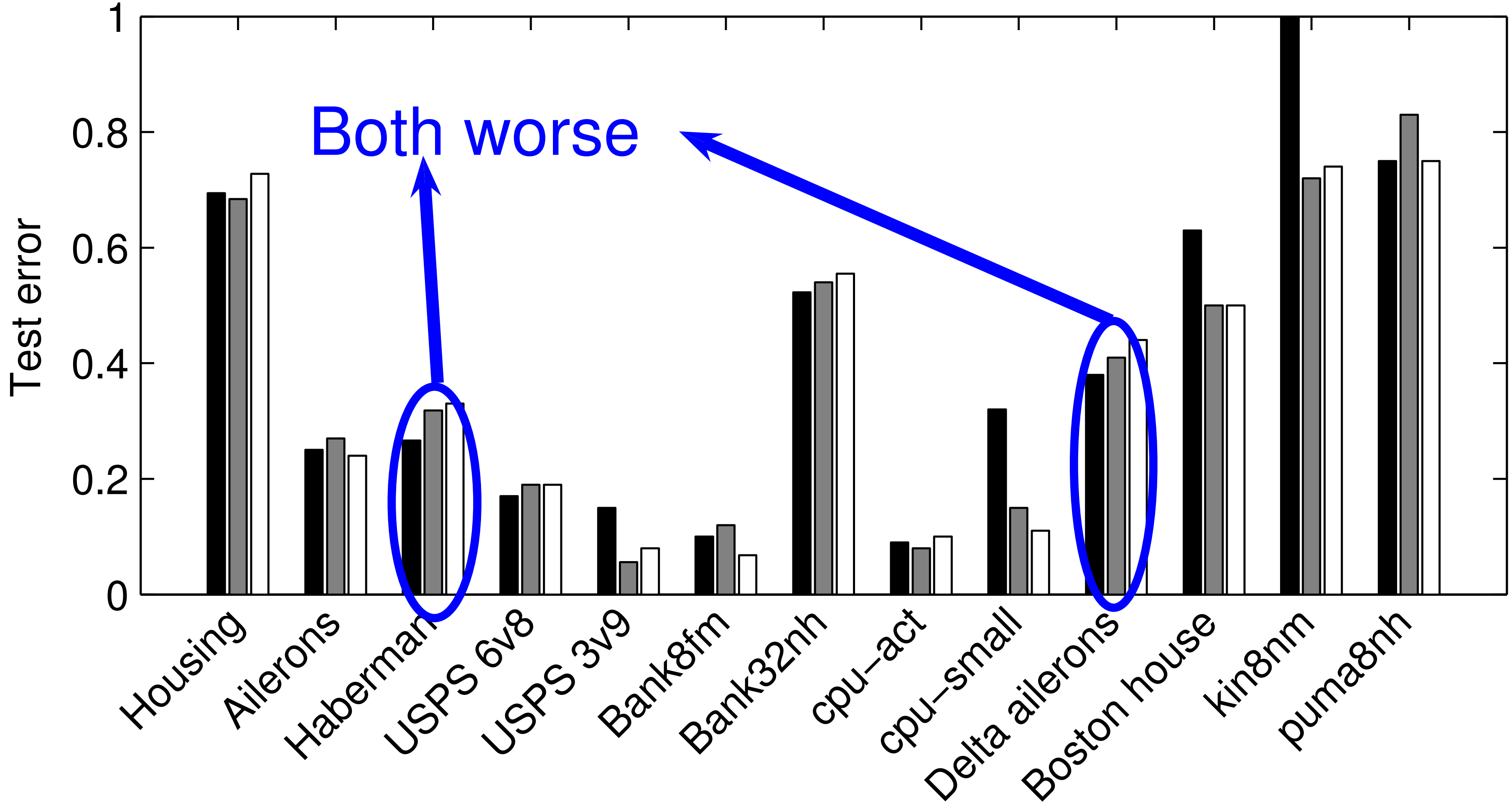
Large scale experiments



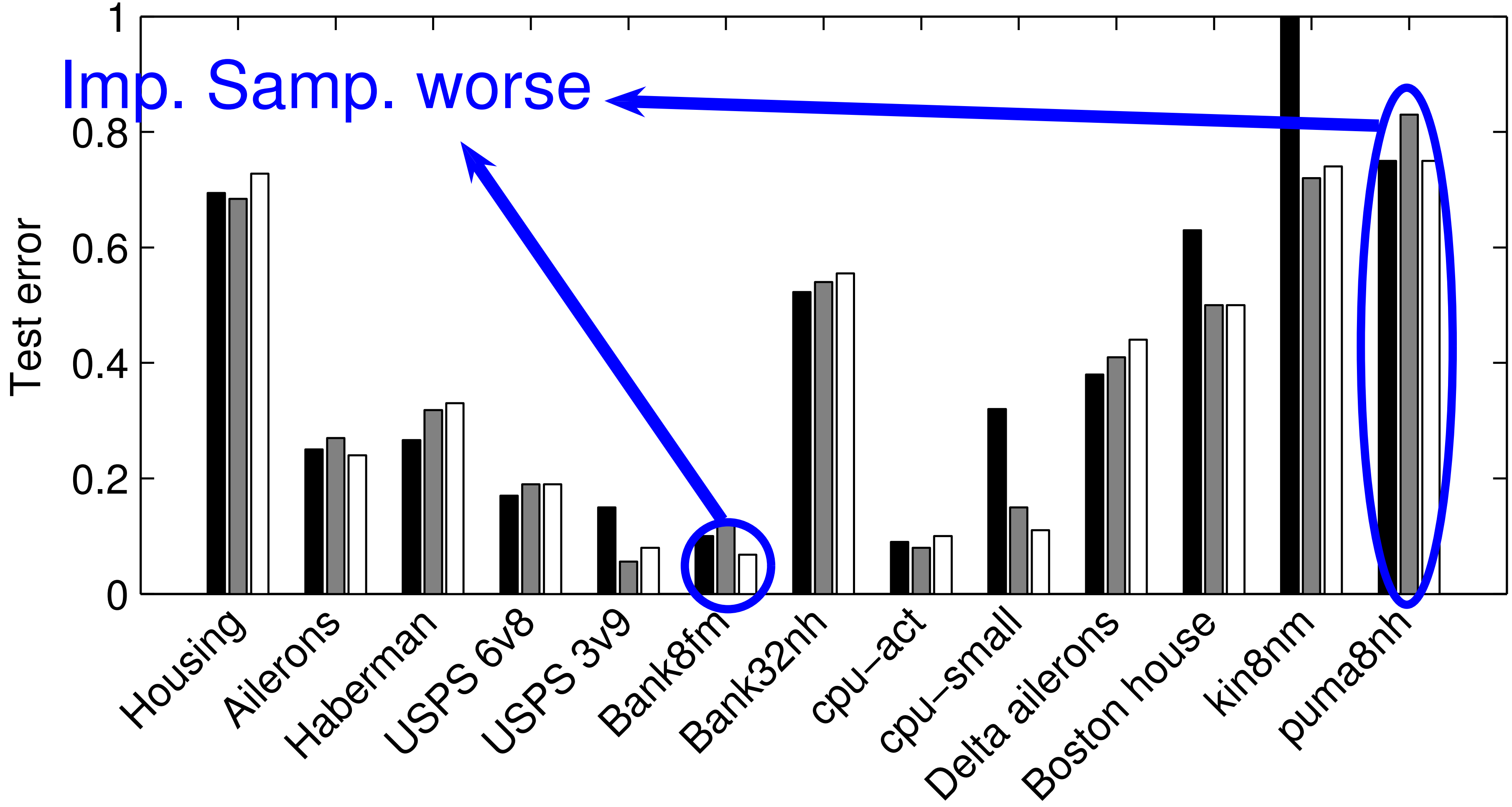
Large scale experiments



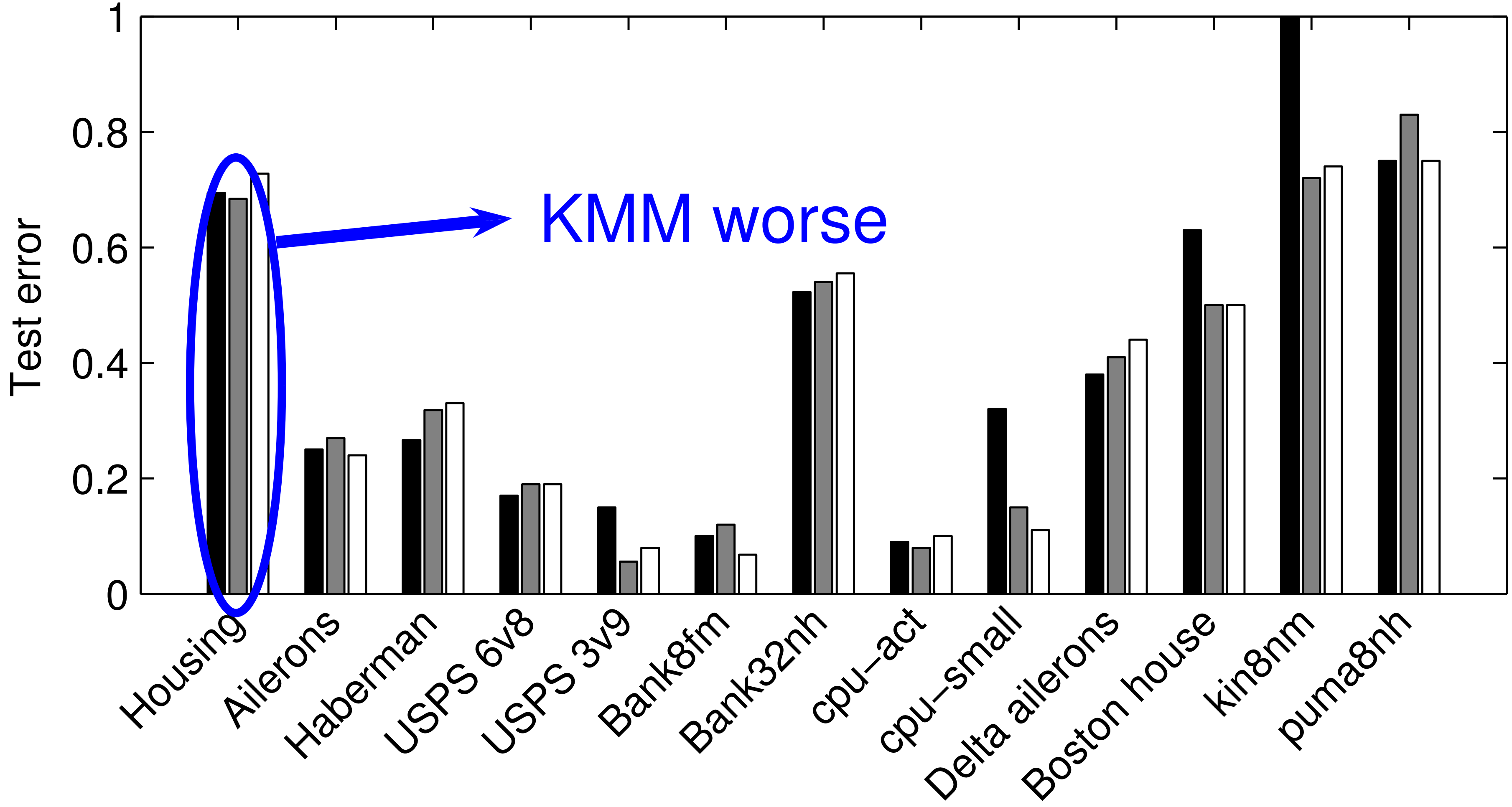
Large scale experiments



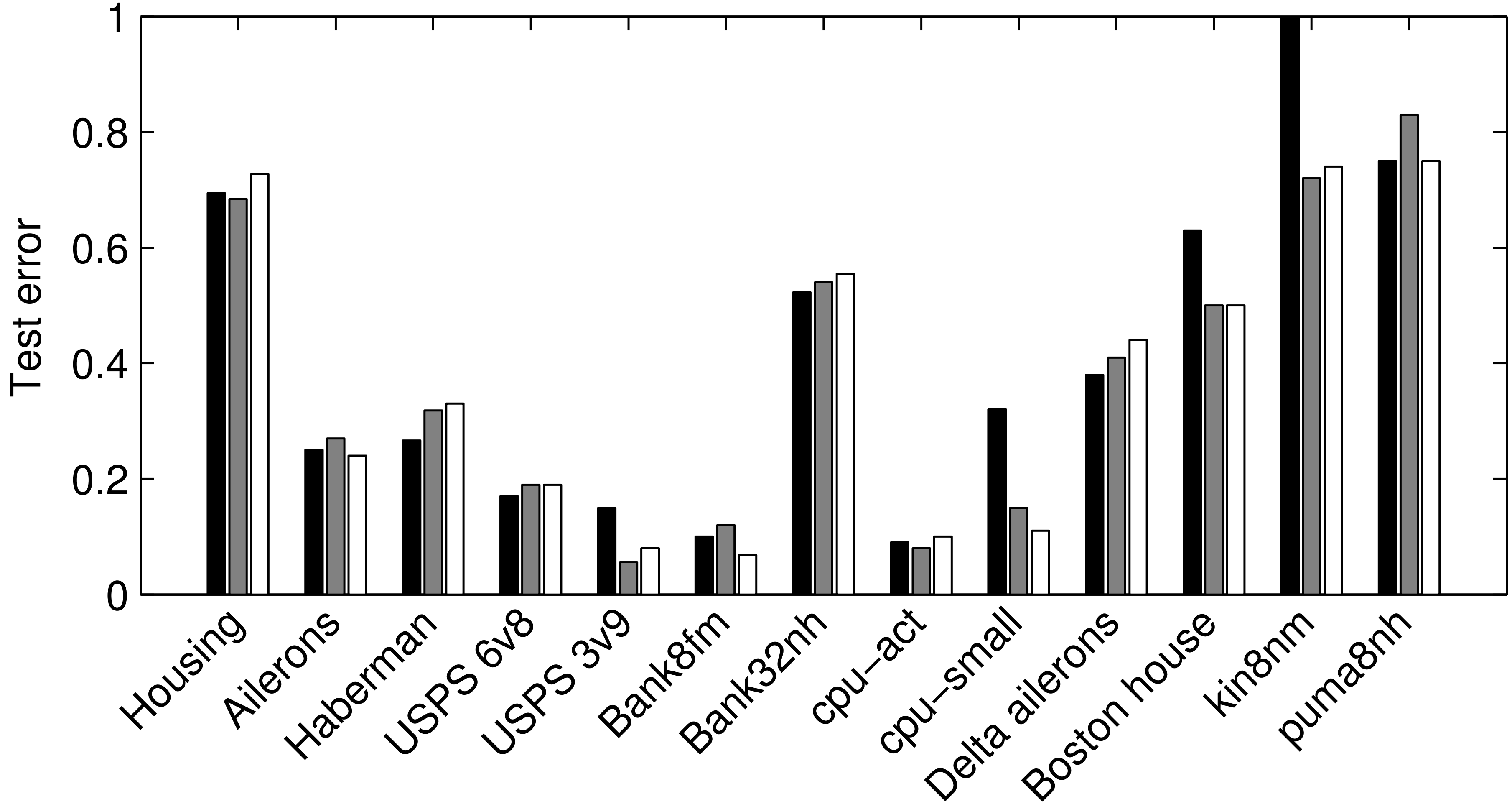
Large scale experiments



Large scale experiments

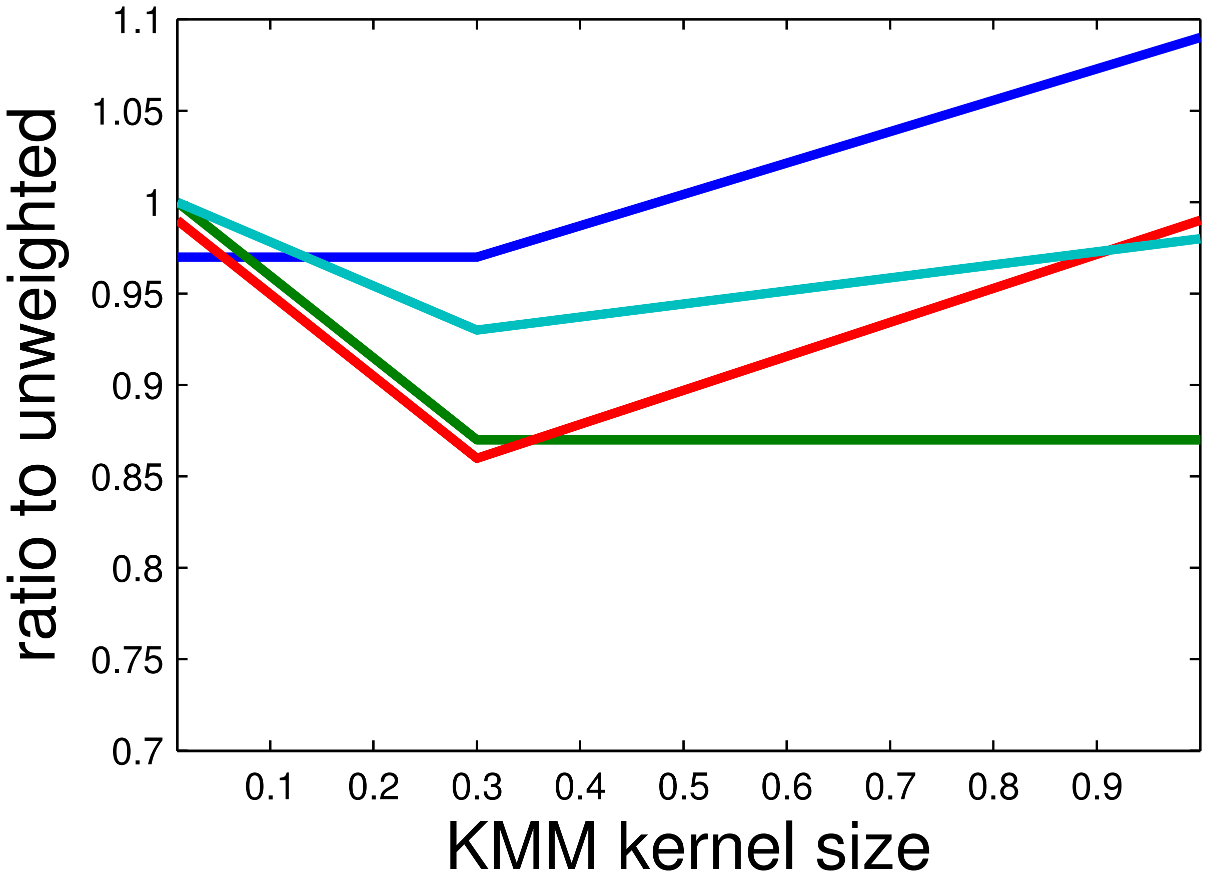


Large scale experiments



Further work: model selection

- Model selection for covariate shift
- Results from [Sugiyama et al., 2008]
- Data have 18-21 dimensions



Further work: model selection

- Model selection for covariate shift

- Some strategies [Bickel et al., 2009]
  - Systematic drift: can be learned [Bickel et al., 2009]
  - Cross validation to obtain error for current β estimate [Sugiyama et al., 2008, Kanamori et al., 2009]
  - Classifier of training vs test: again, cross-validate [Bickel et al., 2009]
  - Supremum of MMD over set of kernels? (this NIPS) [Sriperumbudur et al., 2010]
- Does knowing something about the learning problem help?

Further work: model selection

- Model selection for covariate shift

- Some strategies [Bickel et al., 2009]
  - Systematic drift: can be learned [Bickel et al., 2009]
  - Cross validation to obtain error for current β estimate [Sugiyama et al., 2008, Kanamori et al., 2009]
  - Classifier of training vs test: again, cross-validate [Bickel et al., 2009]
  - Supremum of MMD over set of kernels? (this NIPS) [Sriperumbudur et al., 2010]

- Does knowing something about the learning problem help?

- Model selection for weighted learning: bias for unweighted? [Kanamori et al., 2009]

Summary

- Kernel mean matching: perform covariate shift. . .
  - . . .without density estimation
  - . . .using only particular covariate features
  - . . .on structured domains
- Large performance advantage for “simple” learning algorithms
- Mixed results for powerful learning algorithms
- Model selection remains an issue

Acknowledgements

- Co-authors on KMM papers:
  - Karsten Borgwardt
  - Jiayuan Huang
  - Marcel Schmittful
  - Bernhard Schölkopf
  - Alex Smola
- Discussions
  - Paul von Bünau
  - Corinna Cortes
  - Klaus-Robert Müller
  - Masashi Sugiyama

Questions?

Bibliography

References

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *JMLR*, 10:2137–2155, 2009.

K. F. Cheng and C. K. Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *ALT*, 2008.

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

R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, 70:266–285, 1953.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520, Cambridge, MA, 2007. MIT Press.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Dataset shift in machine learning. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, editors, *Covariate Shift and Local Learning by Distribution Matching*, pages 131–160, Cambridge, MA, 2008. MIT Press.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.

T. Kanamori, S. Hido, , and M Sugiyama. A least-squares approach to direct importance estimation. *Journal of*

Characteristic kernels

Characteristic Kernels (1)

- Characteristic: MMD a metric (MMD = 0 iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

Characteristic Kernels (1)

- Characteristic: MMD a metric (MMD = 0 iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Translation invariant kernels: $k(x, y) = k(x - y)$

Characteristic Kernels (1)

- Characteristic: MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Translation invariant kernels: $k(x, y) = k(x - y)$
- Bochner's theorem:

$$k(x) = \int_{\mathbb{R}^d} e^{-ix^\top \omega} d\Lambda(\omega)$$

  - Λ finite non-negative Borel measure

Characteristic Kernels (1)

- Characteristic: MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Translation invariant kernels: $k(x, y) = k(x - y)$
- Bochner's theorem:

$$k(x) = \int_{\mathbb{R}^d} e^{-ix^\top \omega} d\Lambda(\omega)$$

  - Λ finite non-negative Borel measure

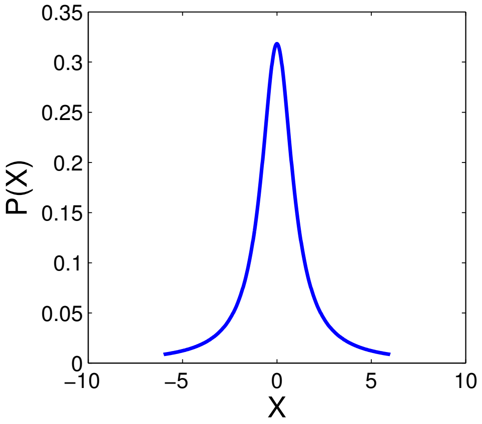
- Fourier representation of MMD:

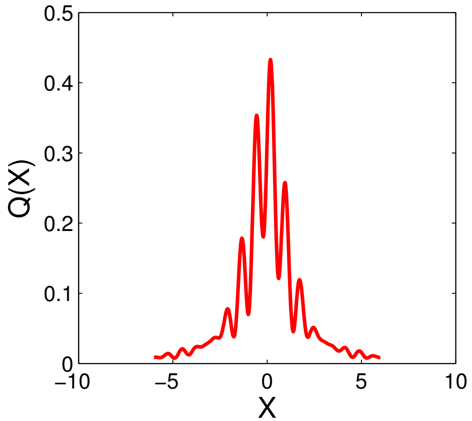
$$\text{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \left[\left(\bar{\phi}_{\mathbf{P}} - \bar{\phi}_{\mathbf{Q}} \right) \Lambda \right]^{\vee} \right\|_{\mathcal{F}}$$

  - $\phi_{\mathbf{P}}$ characteristic function of $\mathbf{P}$
  - $f^{\wedge}$ is Fourier transform, $f^{\vee}$ is inverse Fourier transform
  - $\mu_x := \int k(\cdot, x)\, d\mathbf{P}(x)$

Characteristic Kernels (2)

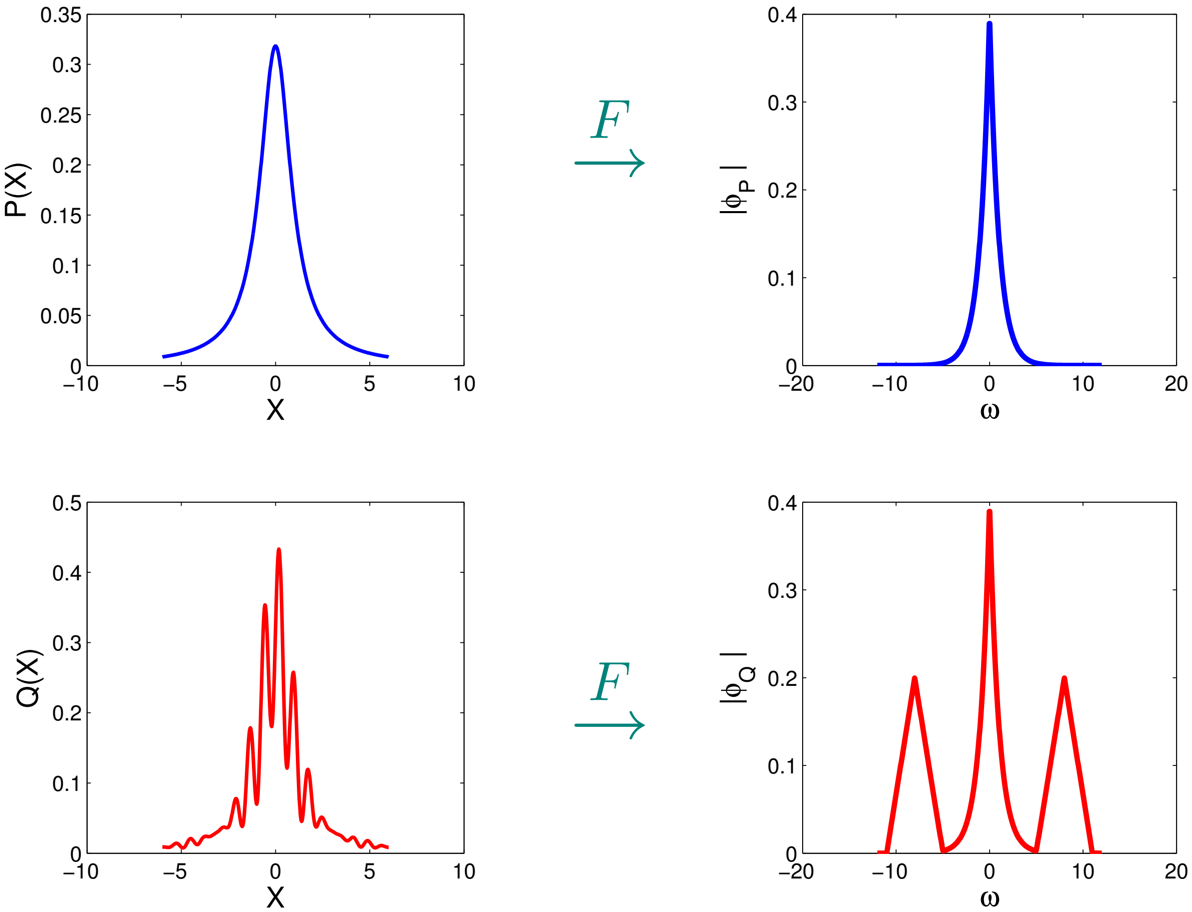
- Example: **P** differs from **Q** at (roughly) one frequency





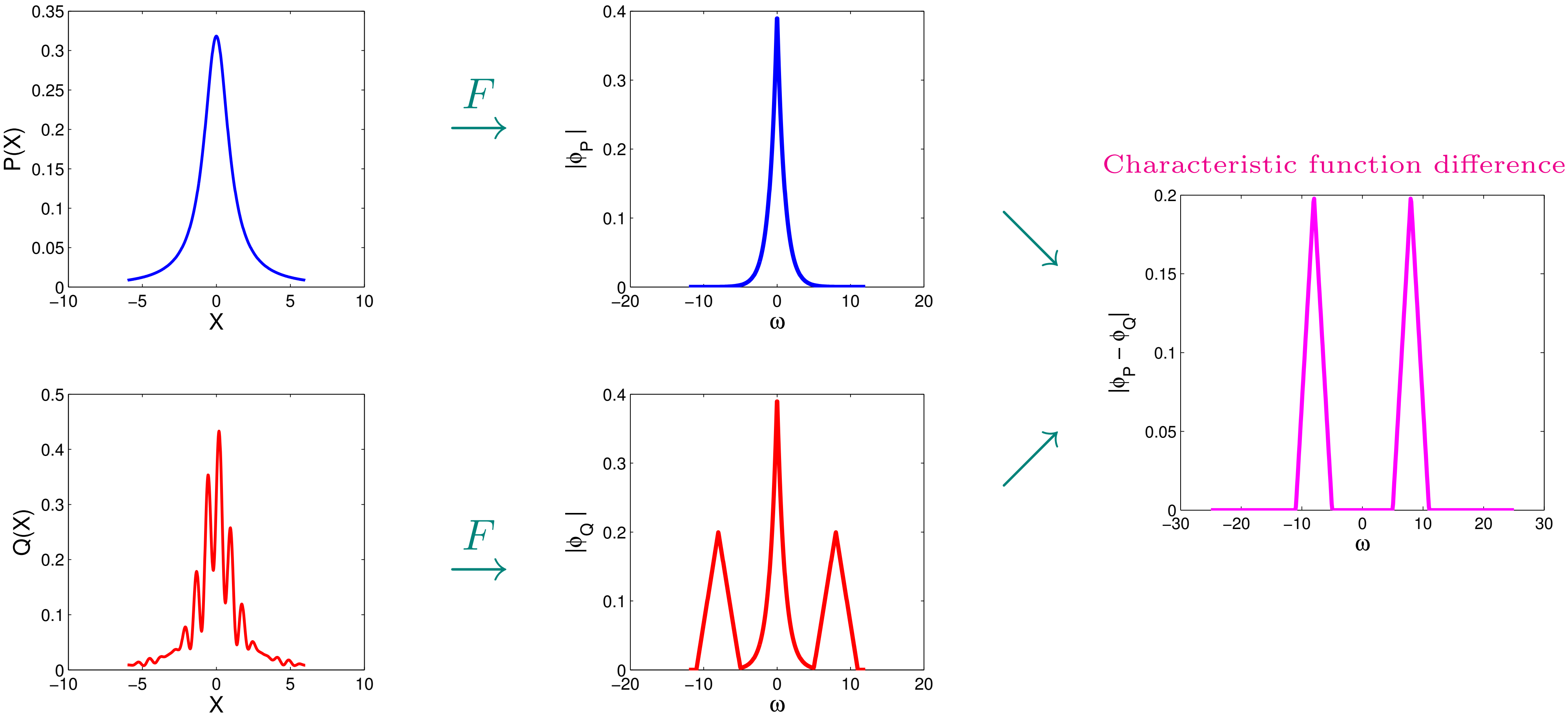
Characteristic Kernels (2)

- Example: **P** differs from **Q** at (roughly) one frequency



Characteristic Kernels (2)

- Example: **P** differs from **Q** at (roughly) one frequency

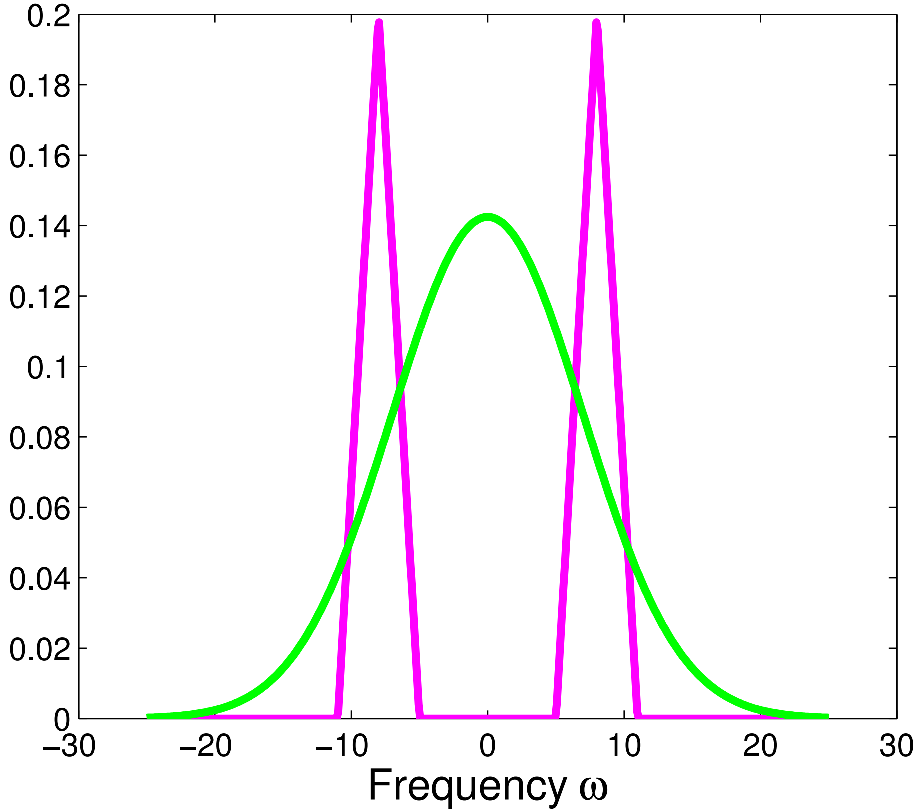


Characteristic Kernels (3)

- Example: **P** differs from **Q** at (roughly) one frequency

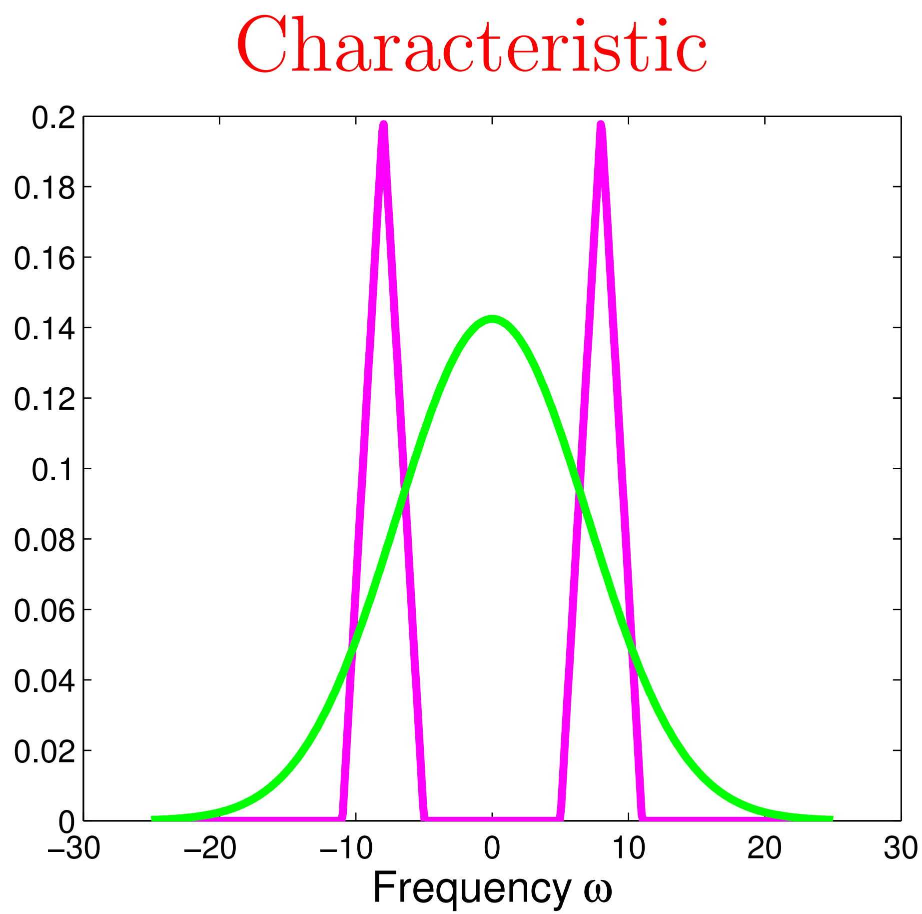
Gaussian kernel

Difference $|\phi_P - \phi_Q|$



Characteristic Kernels (3)

- Example: **P** differs from **Q** at (roughly) one frequency

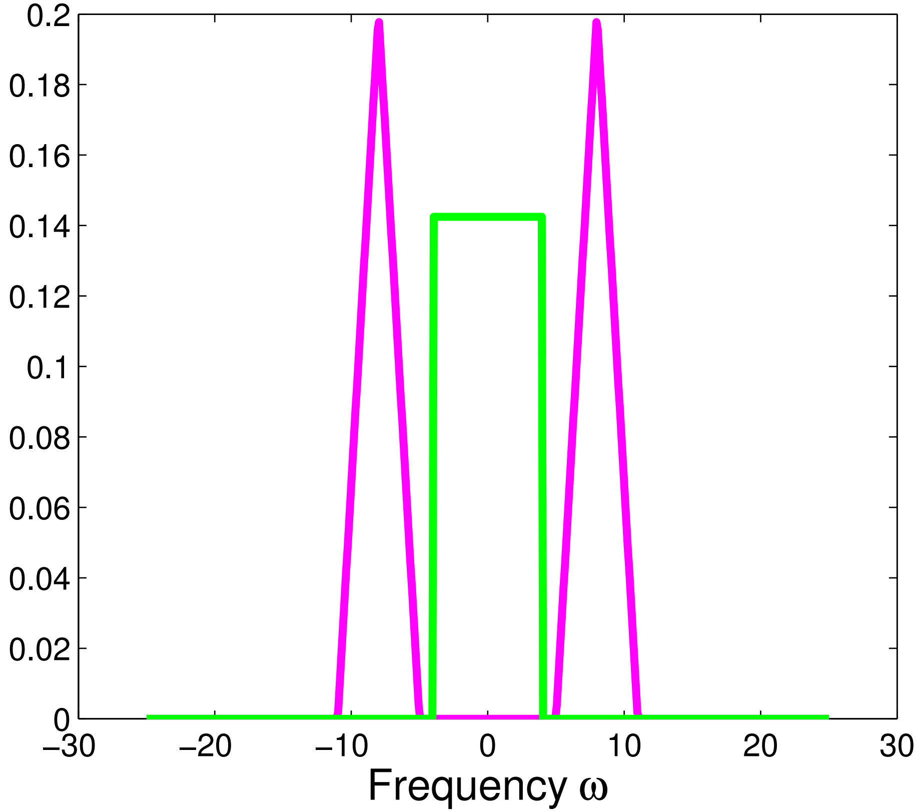


Characteristic Kernels (4)

- Example: **P** differs from **Q** at (roughly) one frequency

Sinc kernel

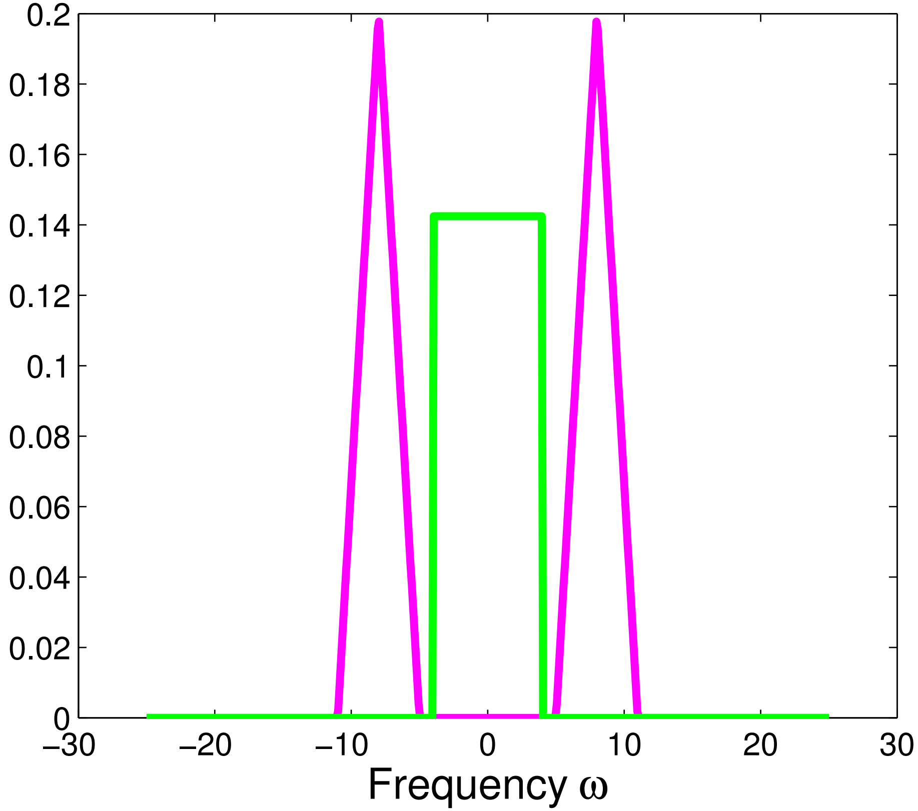
Difference $|\phi_P - \phi_Q|$



Characteristic Kernels (4)

- Example: **P** differs from **Q** at (roughly) one frequency

NOT characteristic

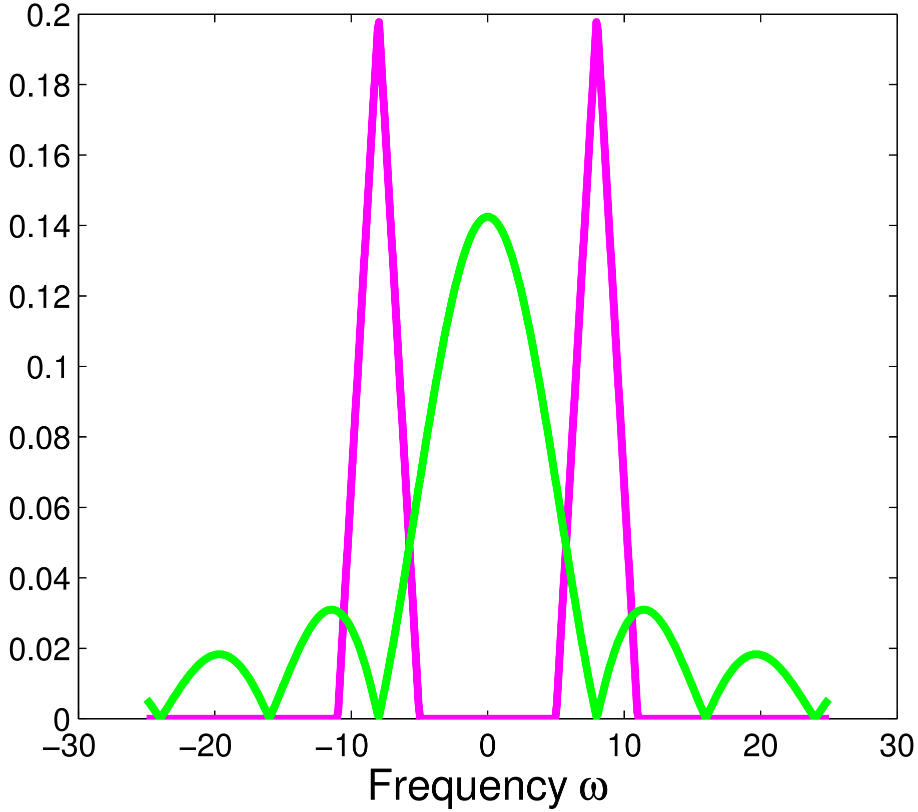


Characteristic Kernels (5)

- Example: **P** differs from **Q** at (roughly) one frequency

B-Spline kernel

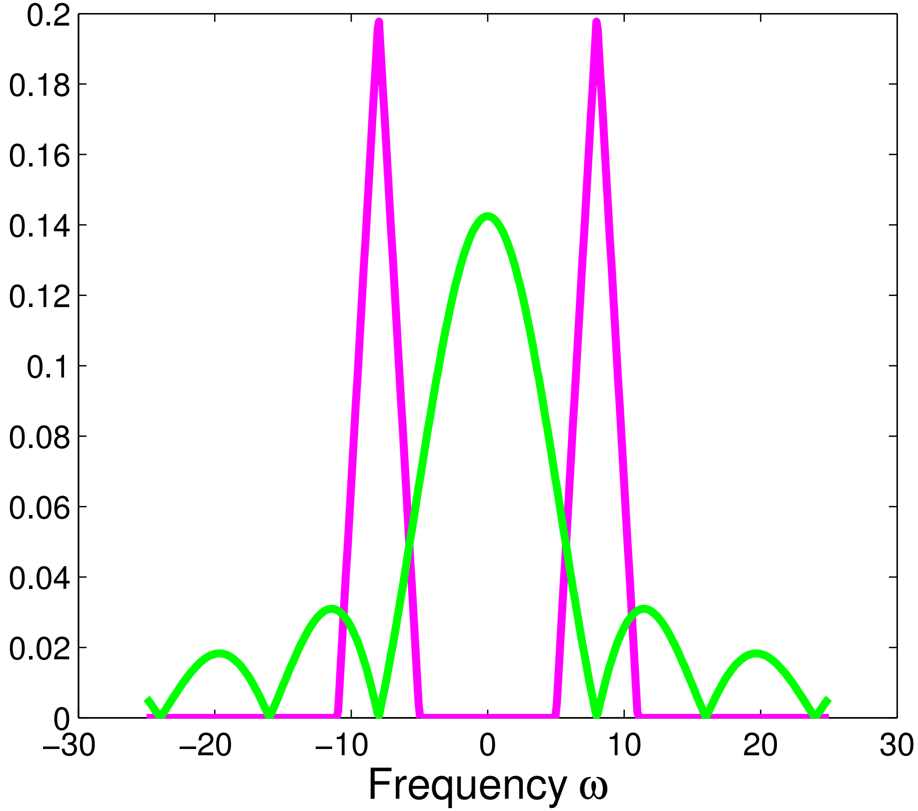
Difference $|\phi_P - \phi_Q|$



Characteristic Kernels (5)

- Example: **P** differs from **Q** at (roughly) one frequency

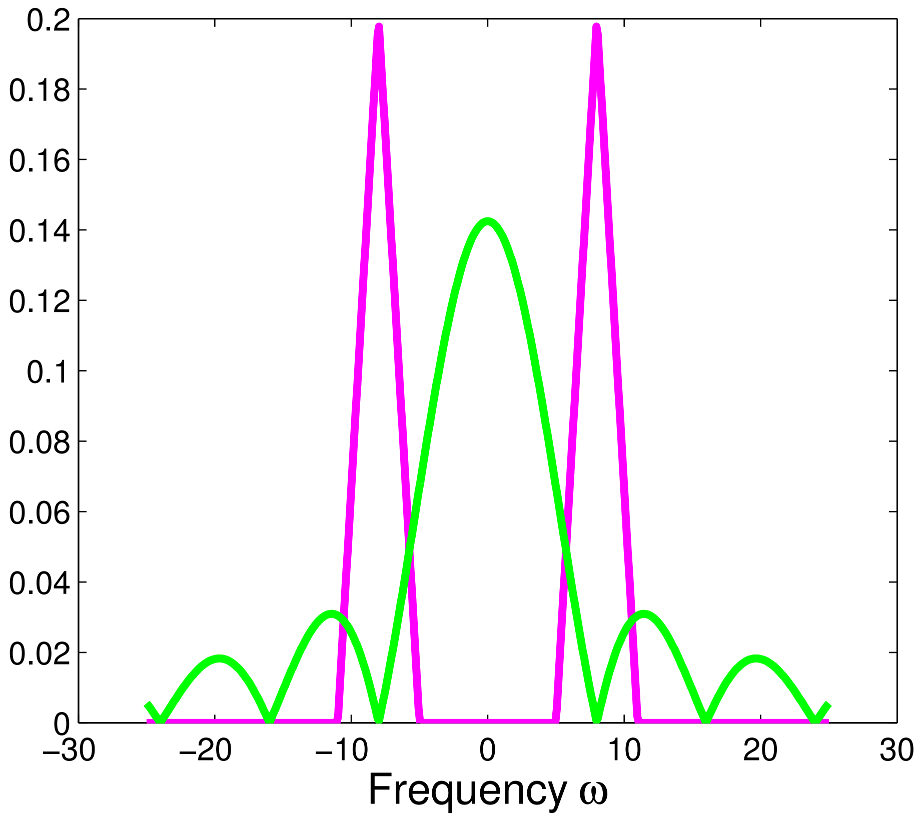
???



Characteristic Kernels (5)

- Example: **P** differs from **Q** at (roughly) one frequency

Characteristic



Summary: Characteristic Kernels

- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ [COLT08]

Summary: Characteristic Kernels

- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ [COLT08]
  - Corollary: continuous, compactly supported k characteristic

Summary: Characteristic Kernels

- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ [COLT08]
  - Corollary: continuous, compactly supported k characteristic
- Alternative property: continuous, strictly P.D., includes NON-translation invariant [COLT09?]

$$k(x, y) = e^{\sigma x^\top y}, \ \sigma > 0$$

Summary: Characteristic Kernels

- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ [COLT08]
  - Corollary: continuous, compactly supported k characteristic
- Alternative property: continuous, strictly P.D., includes NON-translation invariant [COLT09?]
- Similar reasoning wherever extensions of Bochner's theorem exist: [NIPS08a]
  - Locally compact Abelian groups (periodic domains)
  - Compact, non-Abelian groups (orthogonal matrices)
  - The semigroup $\mathbb{R}_n^+$ (histograms)