A Kernel Approache to Covariate Shift

Arthur Gretton

Gatsby Computational Neuroscience Unit

November 2011

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:
  - Brain computer interfaces
  - Spam detection
  - Medical diagnosis

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:
  - Brain computer interfaces
  - Spam detection
  - Medical diagnosis

Does this make sense?

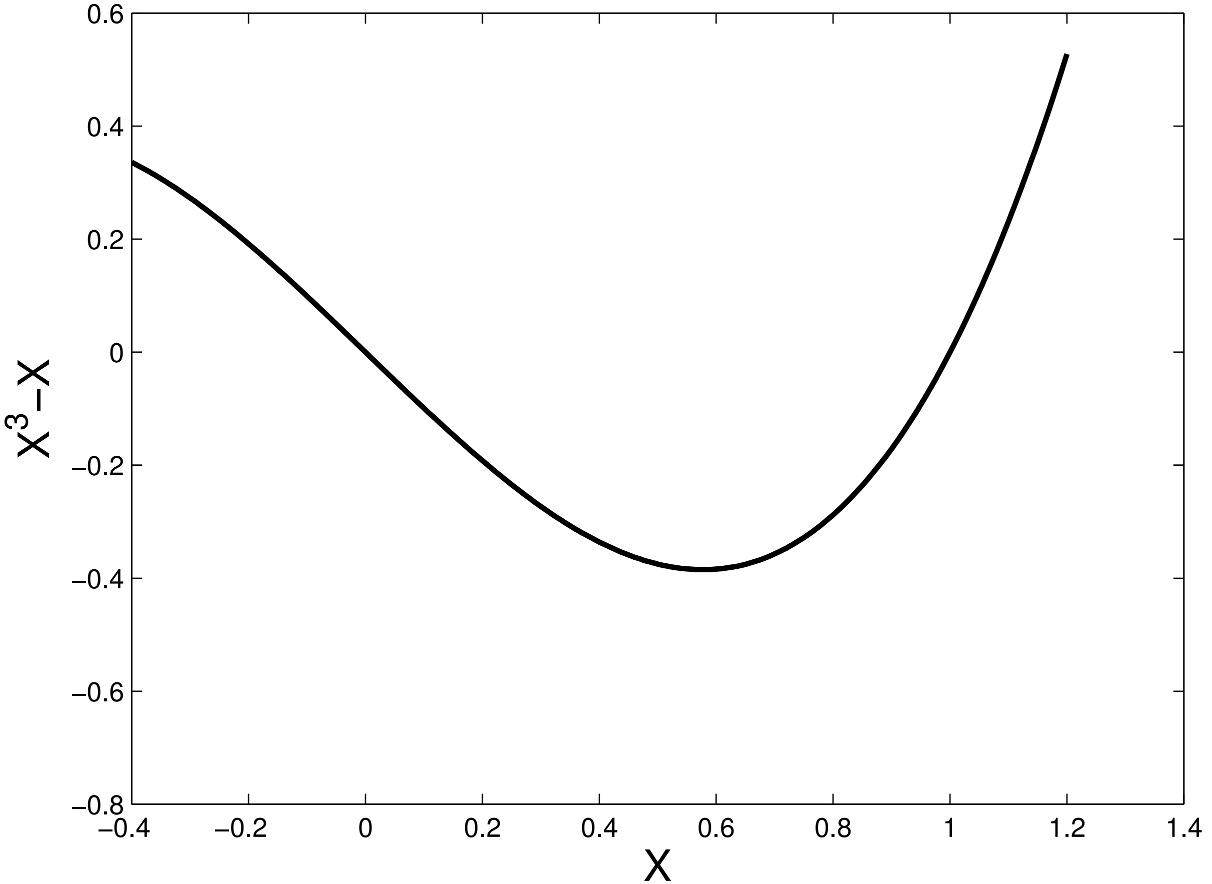
Transfer learning and covariate shift

- 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:
  - Brain computer interfaces
  - Spam detection
  - Medical diagnosis
- Assumption: $\mathbf{P}_{\text{tr}}(x, y) = \mathbf{P}(y|x)\mathbf{P}_{\text{tr}}(x)$ and $\mathbf{P}_{\text{te}}(x, y) = \mathbf{P}(y|x)\mathbf{P}_{\text{te}}(x)$

Conditional probs unchanged: covariate shift

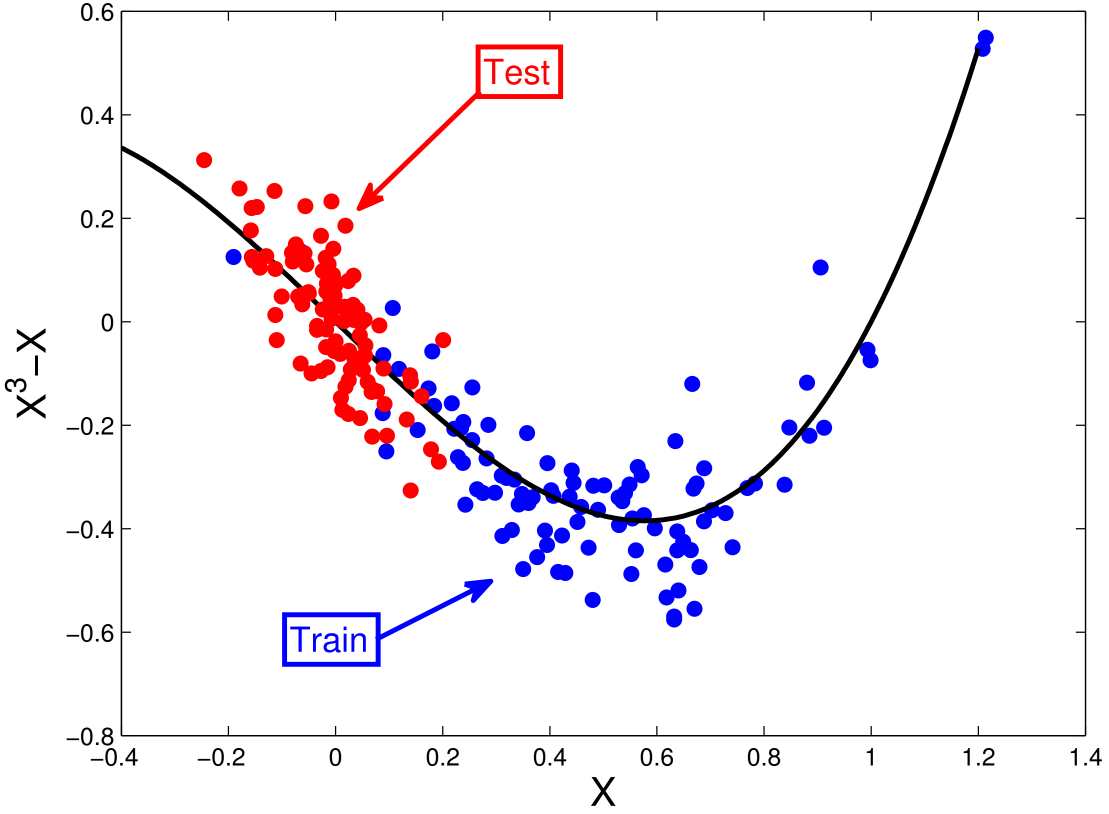
A toy example

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



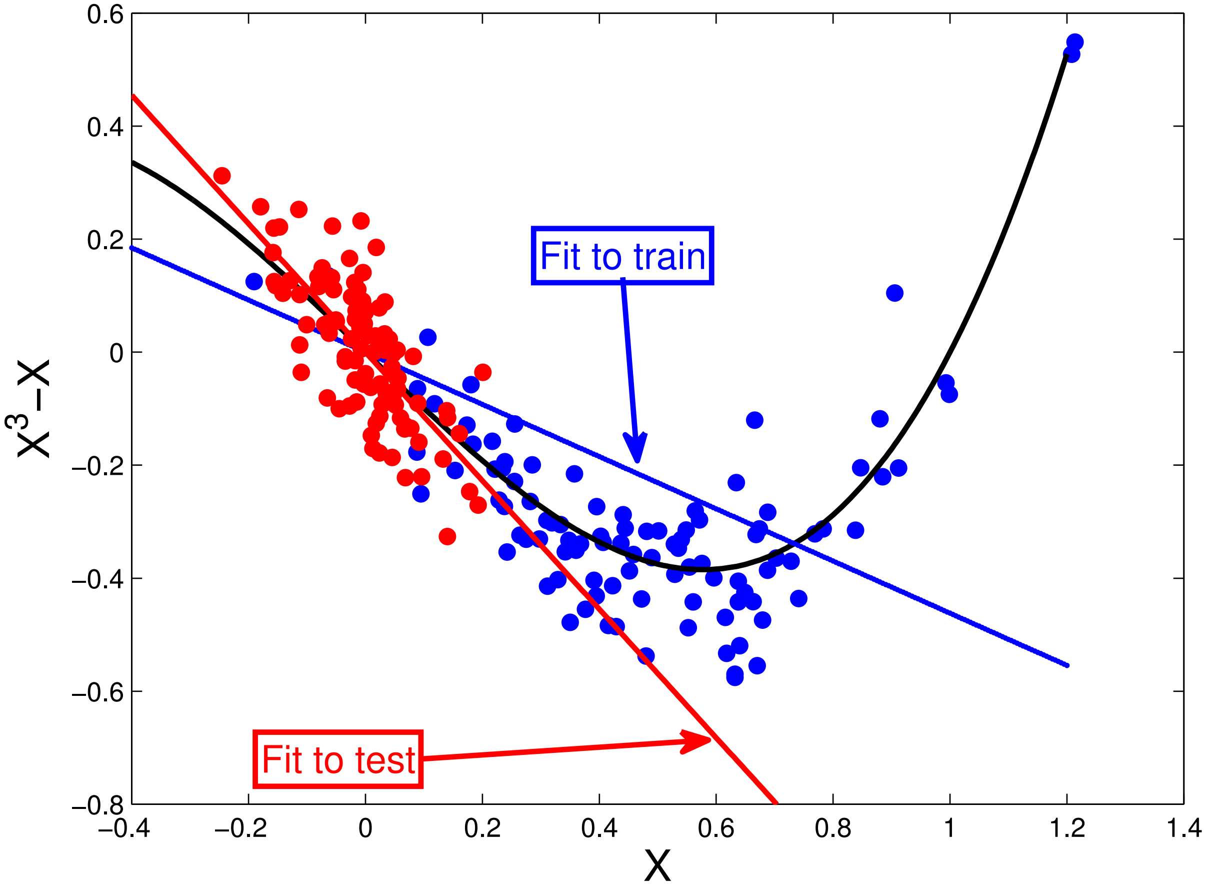
A toy example

- 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:

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

- Importance weighting:

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

Importance weighting

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

$$
\mathrm{var}_{\mathsf{P}_{\mathrm{tr}}}\left(l(x,y,\theta)\frac{\mathsf{P}_{\mathrm{te}}(x,y)}{\mathsf{P}_{\mathrm{tr}}(x,y)}\right)
$$

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

Importance weighting

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

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

Importance weighting

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

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

Importance weighting

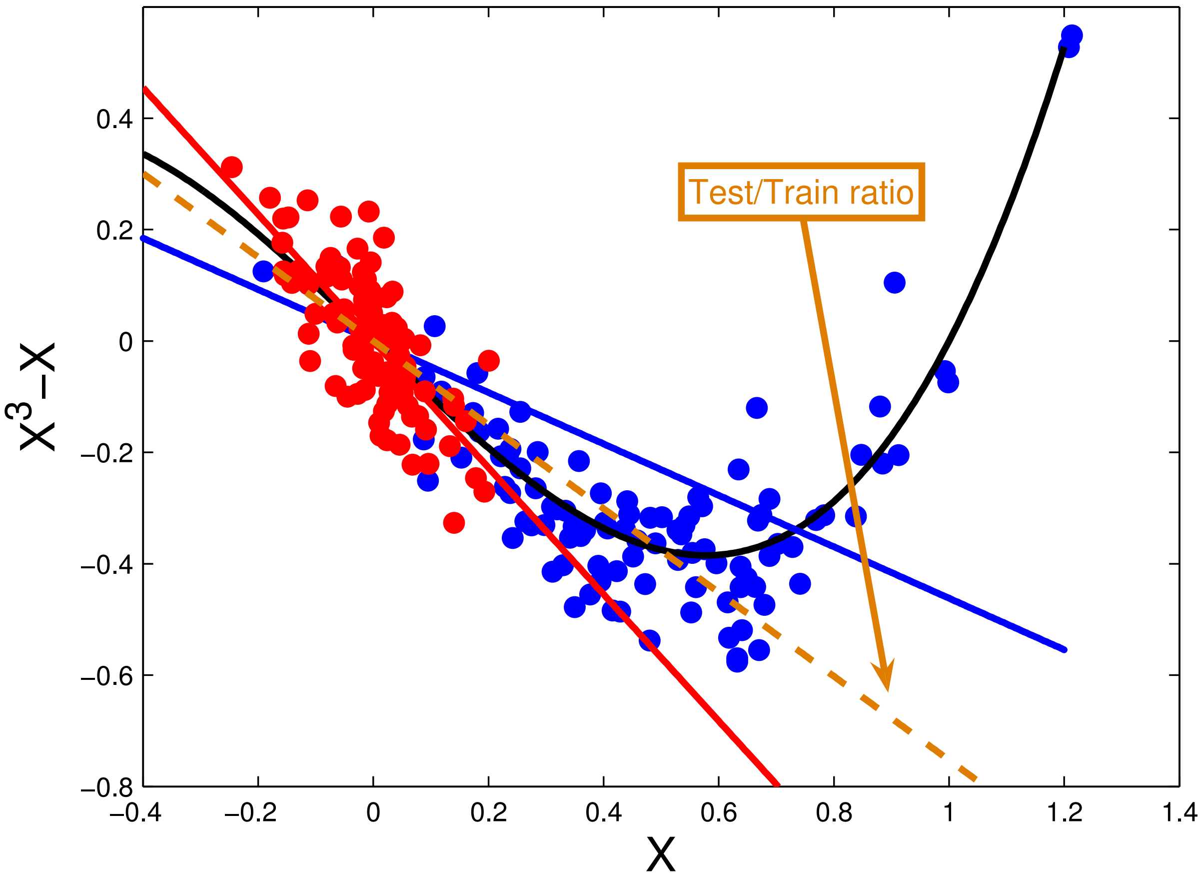
- Variance of importance weighted risk [Robert and Casella, 2004]

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

- P_{tr} should have heavier tails than P_{te}

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

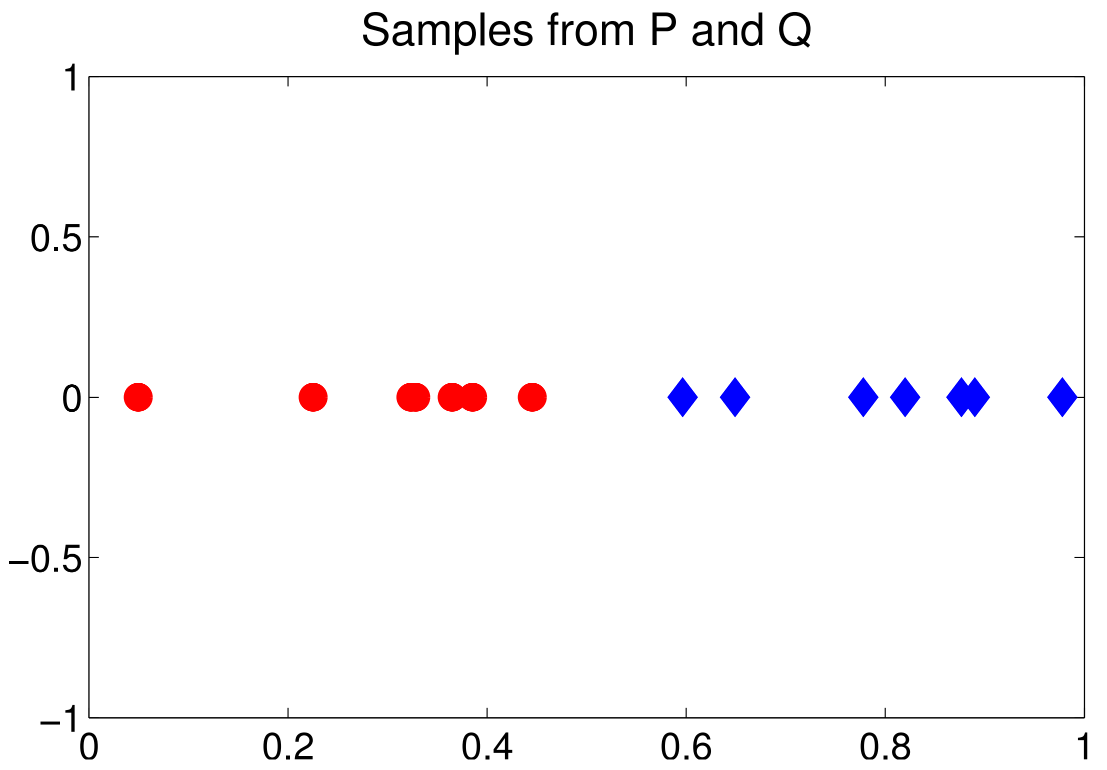
Alternatives to density estimation

- 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 $\beta\mathbf{P}_{\text{tr}}$ 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 $\beta\mathbf{P}_{\text{tr}}$ and $\mathbf{P}_{\text{te}}$ [Huang et al., 2007, Gretton et al., 2008]

Kernel distribution metric for transfer learning

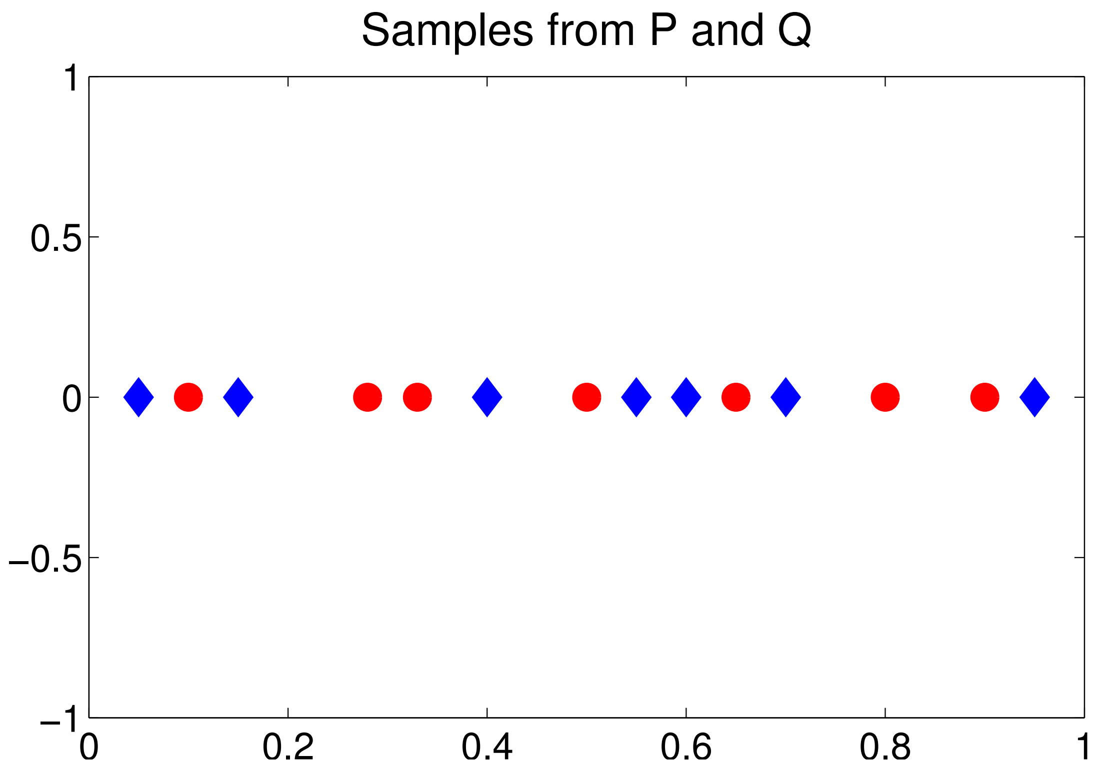
A distance between distributions

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



A distance between distributions

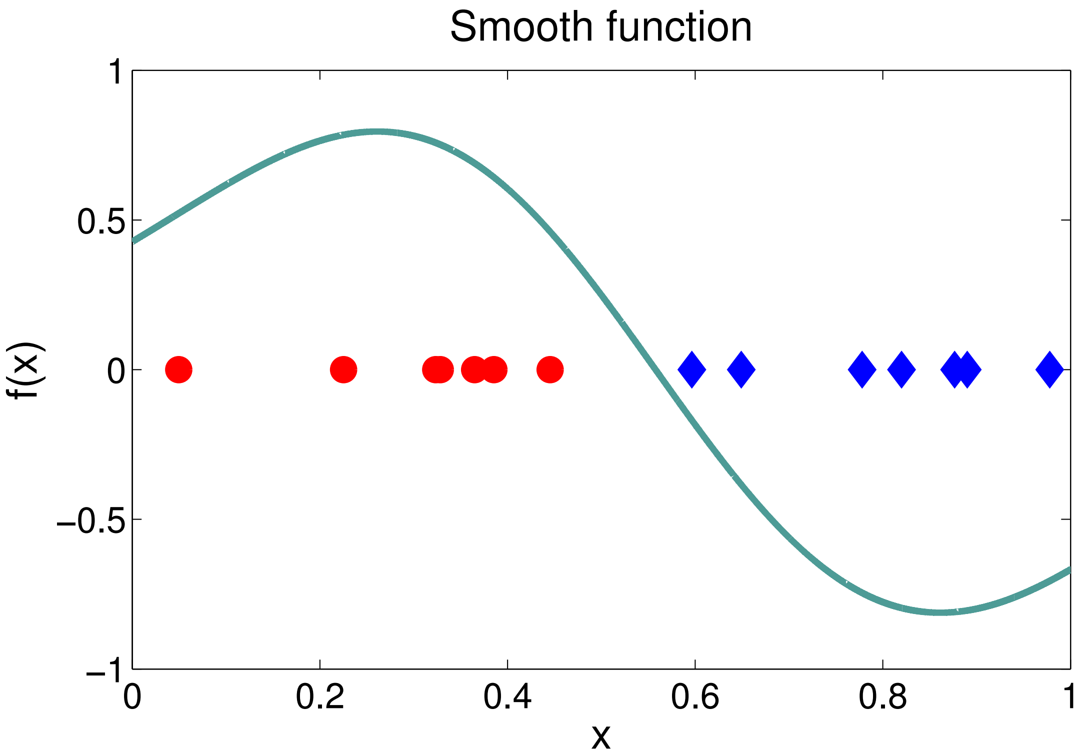
- Are **P** and **Q** different?



A distance between distributions

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

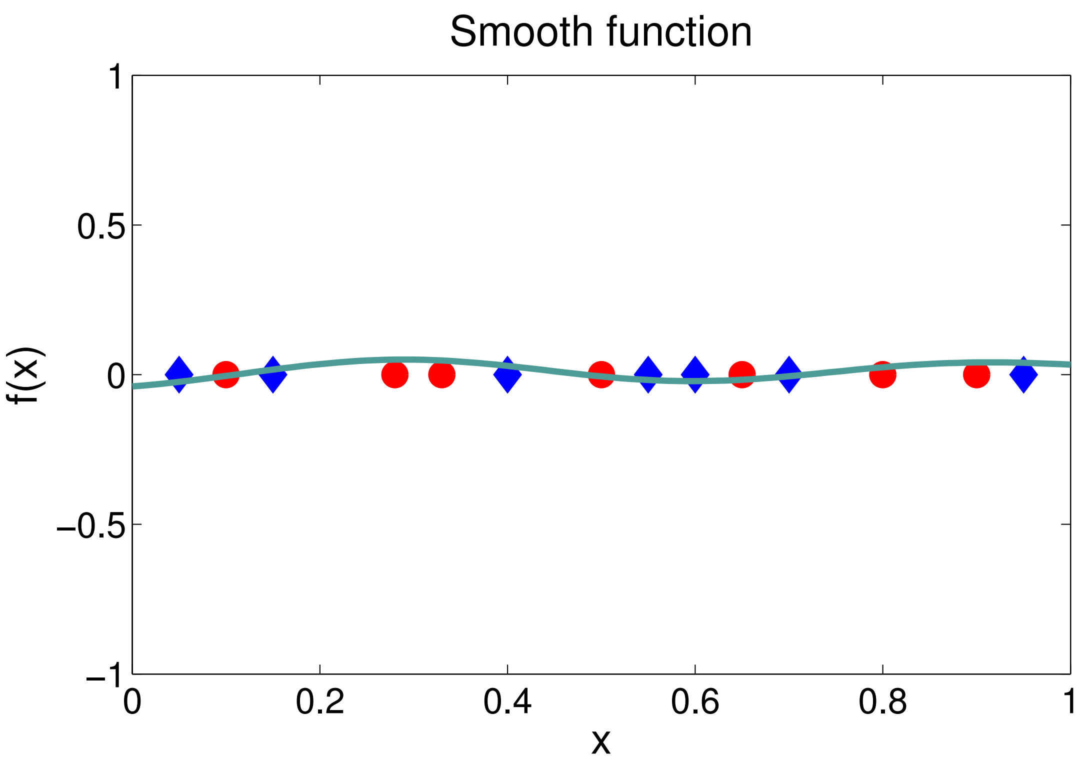
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathsf{y})\right].$$



A distance between distributions

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

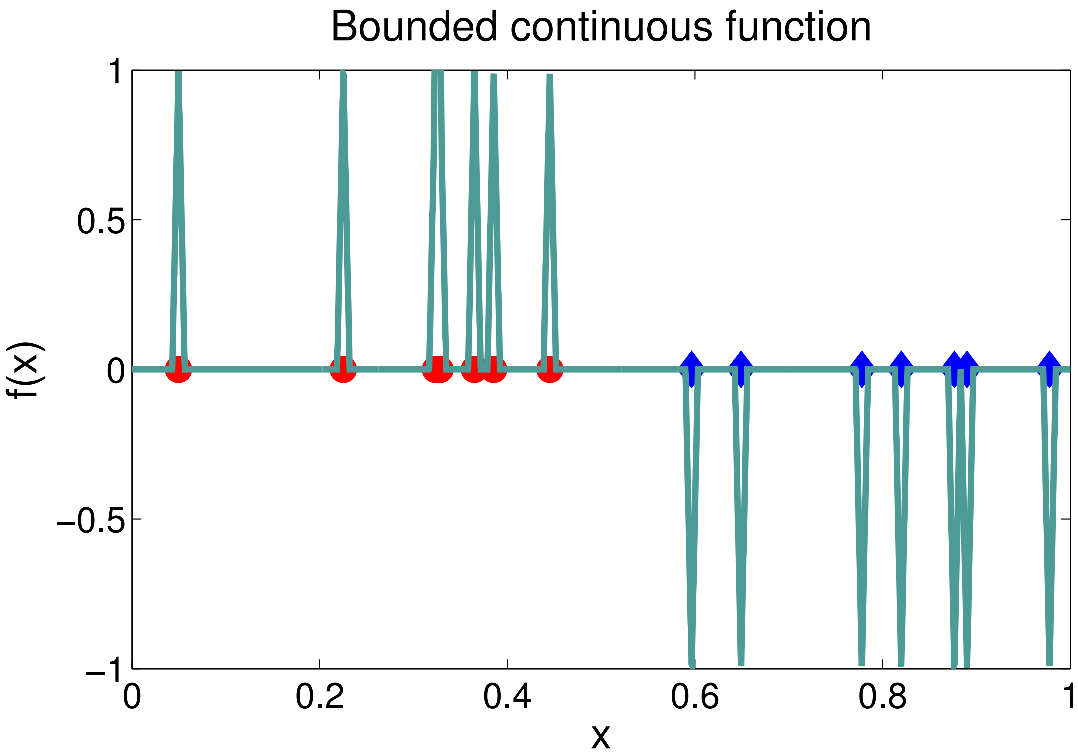
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathrm{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathrm{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

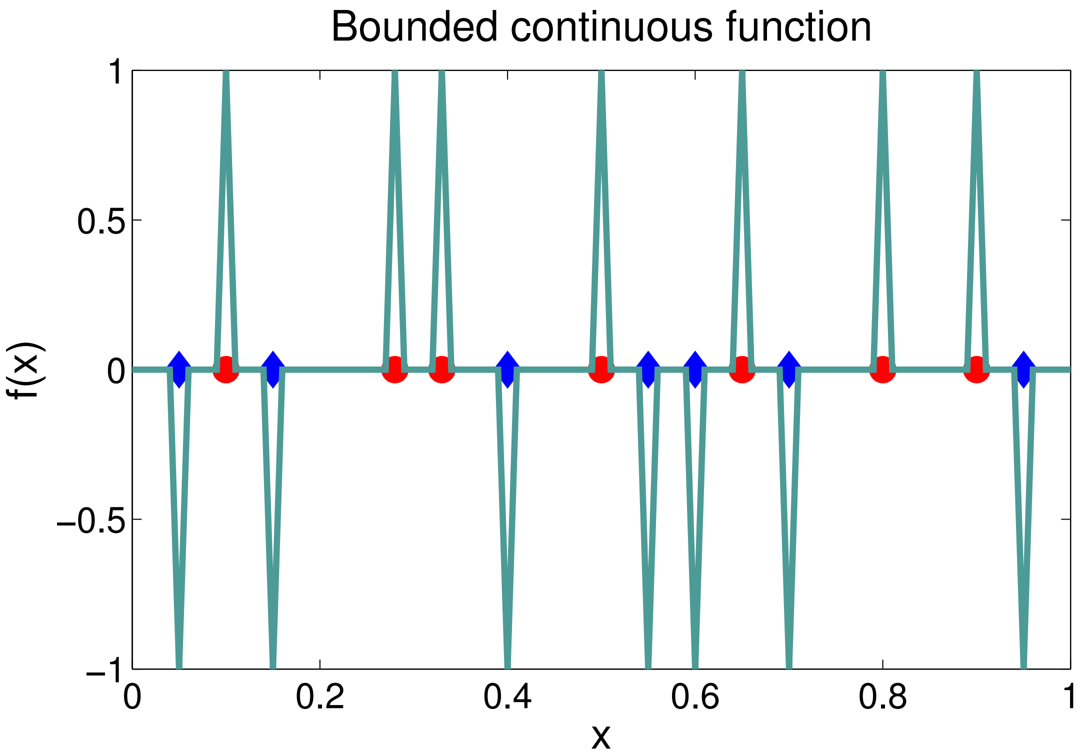
$$\text{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

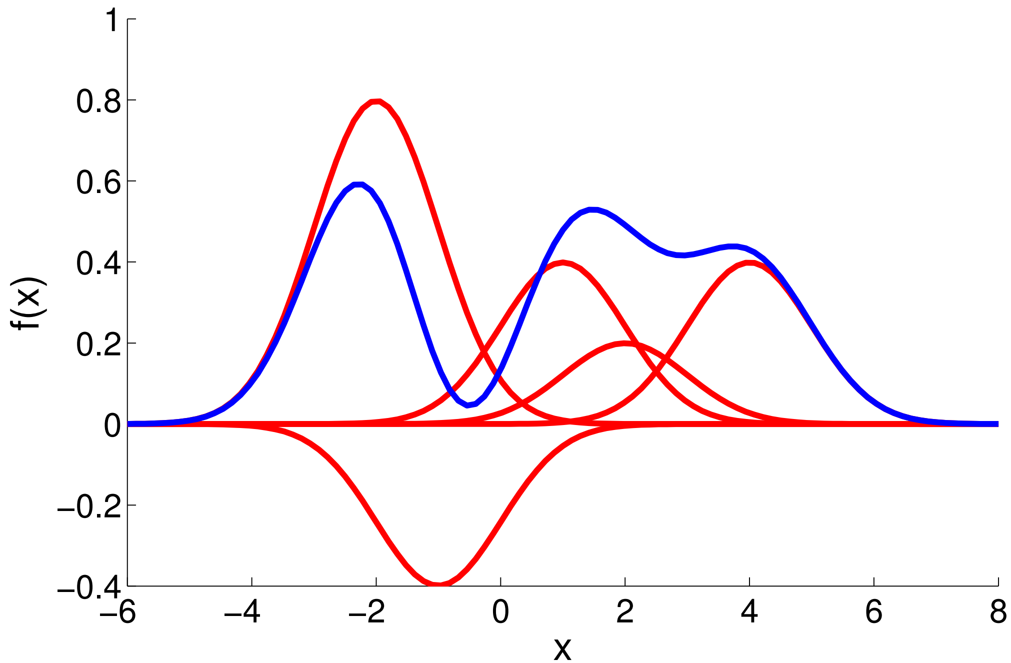
- What if the function is not smooth?

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



Constructing the smooth function

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



Kernel mean matching for transfer learning

- Reweight training points to minimize MMD: Kernel Mean Matching (KMM)

$$\underset{\beta}{\text{minimize}} \;\; \text{MMD}(\mathbf{P}_{\text{te}}(x), \beta(x)\mathbf{P}_{\text{tr}}(x); F)$$

$$\text{subject to } \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathbf{P}_{\text{tr}}}[\beta(x)] = 1.$$

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

Kernel mean matching for transfer learning

- Reweight training points to minimize MMD: Kernel Mean Matching (KMM)

$$\underset{\beta}{\text{minimize}} \;\; \mathrm{MMD}(\mathbf{P}_{\text{te}}(x), \beta(x)\mathbf{P}_{\text{tr}}(x); F)$$

$$\text{subject to } \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathbf{P}_{\text{tr}}}[\beta(x)] = 1.$$

- 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:

$$\min_{\beta} \left(\frac{1}{{n_{\text{tr}}}^2} \beta^\top K \beta - \frac{2}{{n_{\text{tr}}}^2} \kappa^\top \beta \right) + \text{const.}$$

Kernel mean matching for transfer learning

- Reweight training points to minimize MMD: Kernel Mean Matching (KMM)

$$\begin{aligned}\underset{\beta}{\text{minimize}}\ & \text{MMD}(\mathbf{P}_{\text{te}}(x), \beta(x)\mathbf{P}_{\text{tr}}(x); F) \\ \text{subject to}\ & \beta(x) \geq 0 \text{ and } \mathbf{E}_{\mathbf{P}_{\text{tr}}}[\beta(x)] = 1.\end{aligned}$$

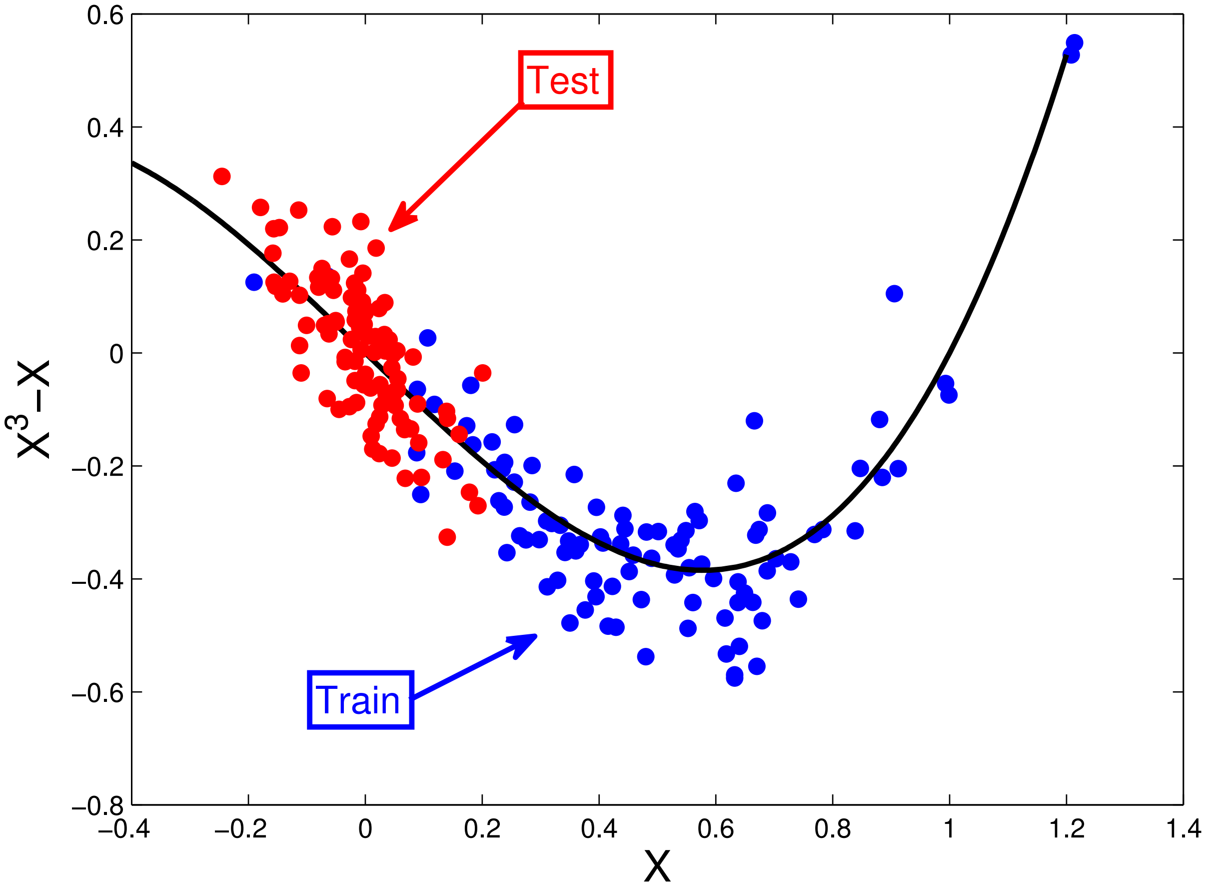
- If $\mathbf{P}_{\text{te}} \ll \mathbf{P}_{\text{tr}}$, characteristic kernel, solution is $\mathbf{P}_{\text{te}}(x) = \beta_{\text{imp}}(x)\mathbf{P}_{\text{tr}}(x)$
- Empirical:

$$\min_{\beta} \left(\frac{1}{{n_{\text{tr}}}^2} \beta^\top K \beta - \frac{2}{{n_{\text{tr}}}^2} \kappa^\top \beta \right) + \text{const.}$$

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

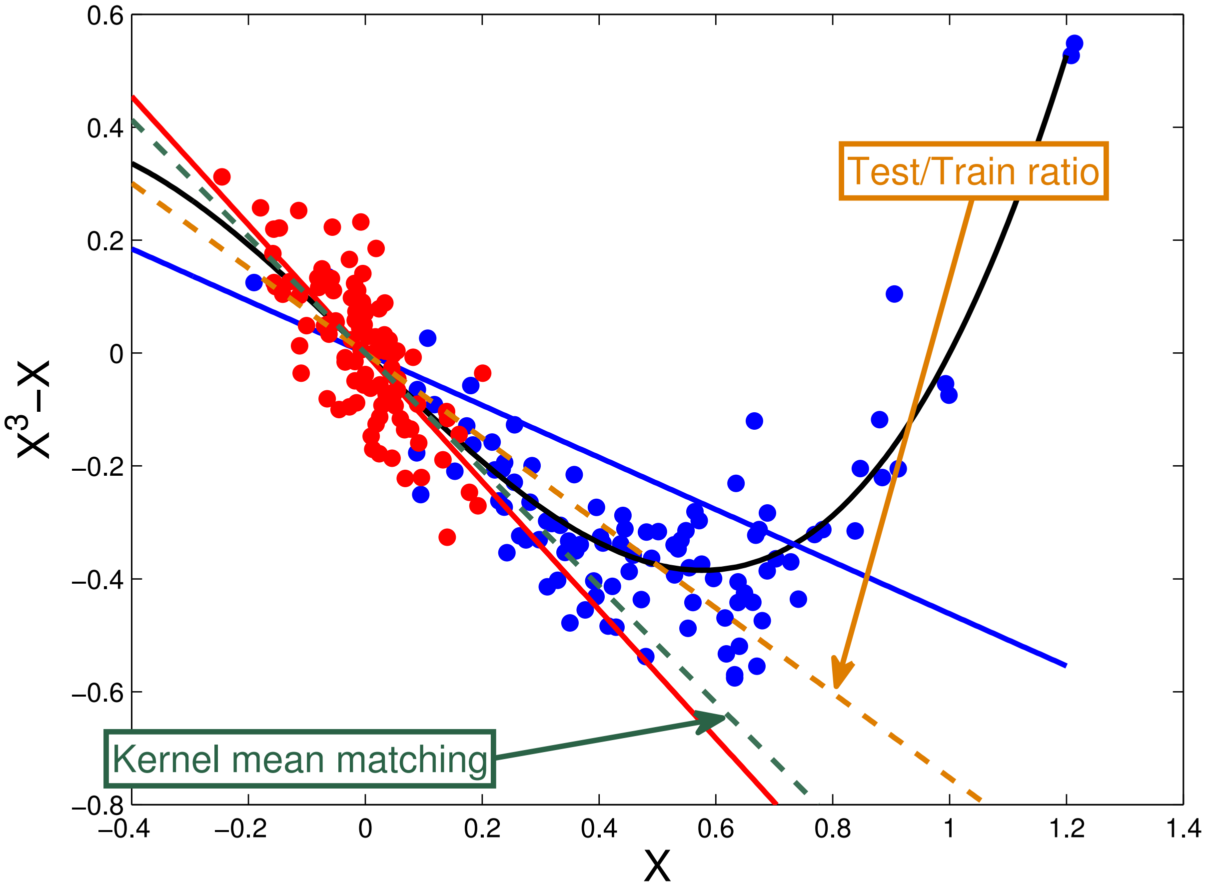
Kernel mean matching for transfer learning

- Compare KMM and importance sampling



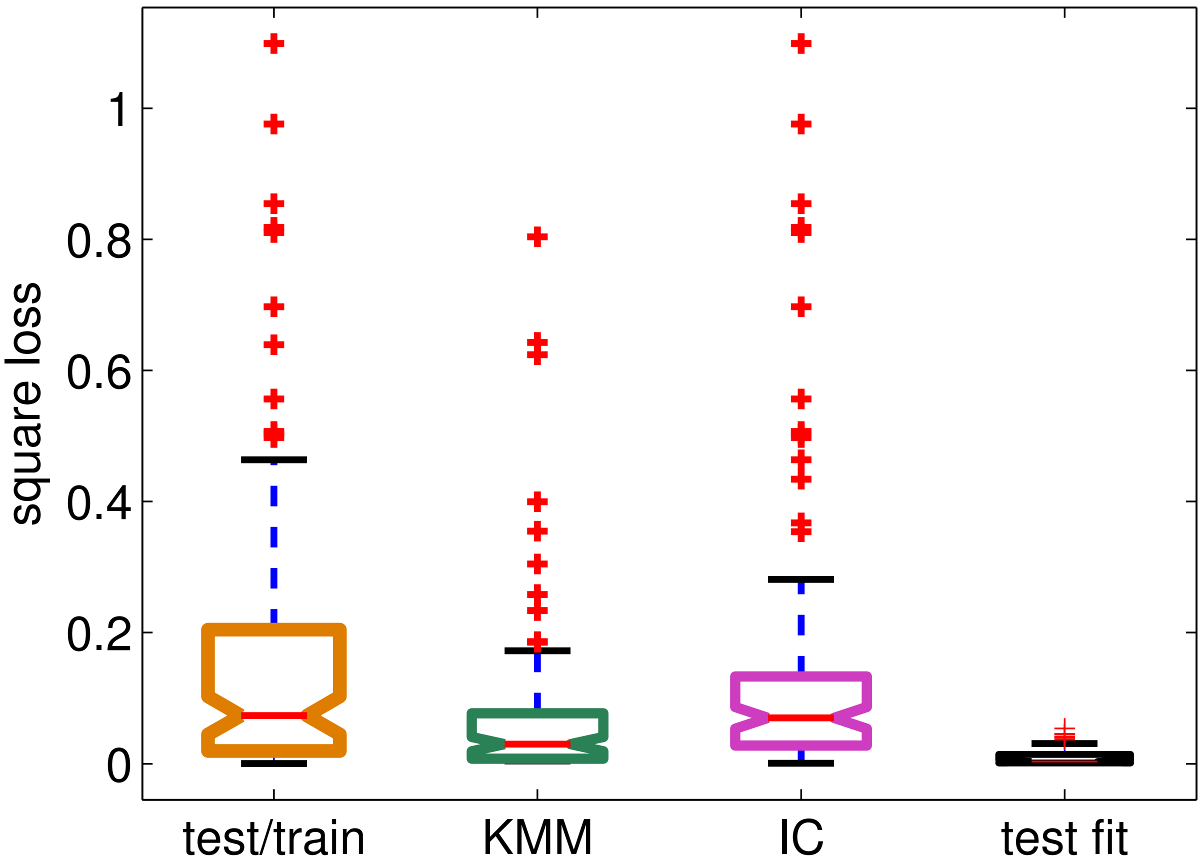
Kernel mean matching for transfer learning

- Compare KMM and importance sampling



Kernel mean matching for transfer learning

- 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]
- $\mathbf{P}\left(S|x, \theta_{\text{shift}}\right)$ classifies training $(s = 1)$ vs test $(s = 0)$
- 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

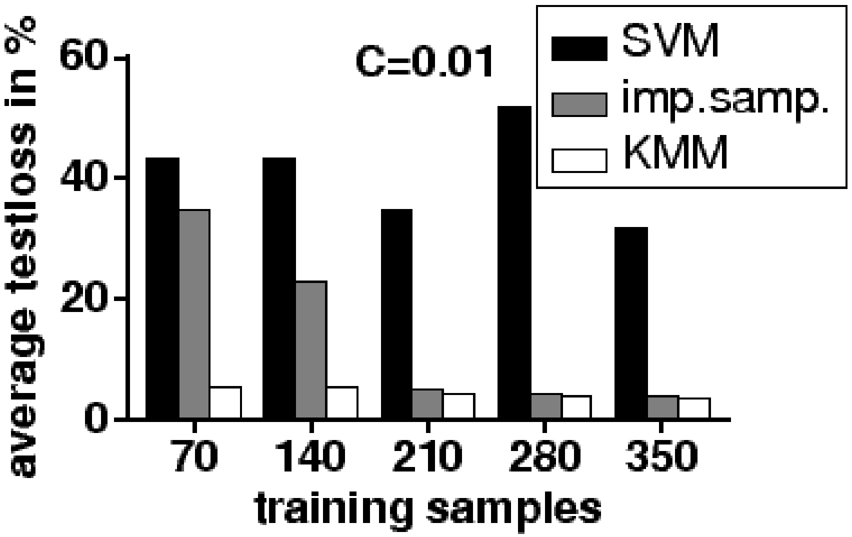
Experiments

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
  - $P(s_i = 1|x_i) \propto \exp(-0.05\|x_i - \overline{x}\|^2)$

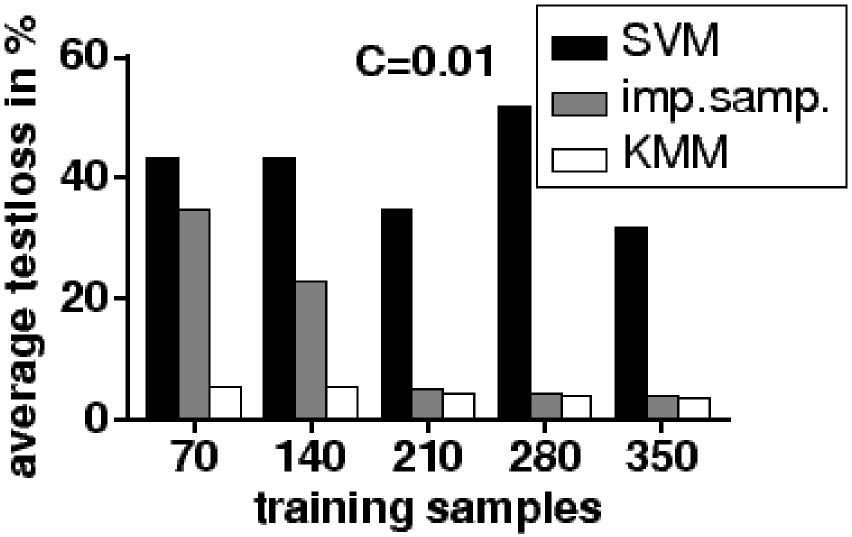
Breast Cancer data

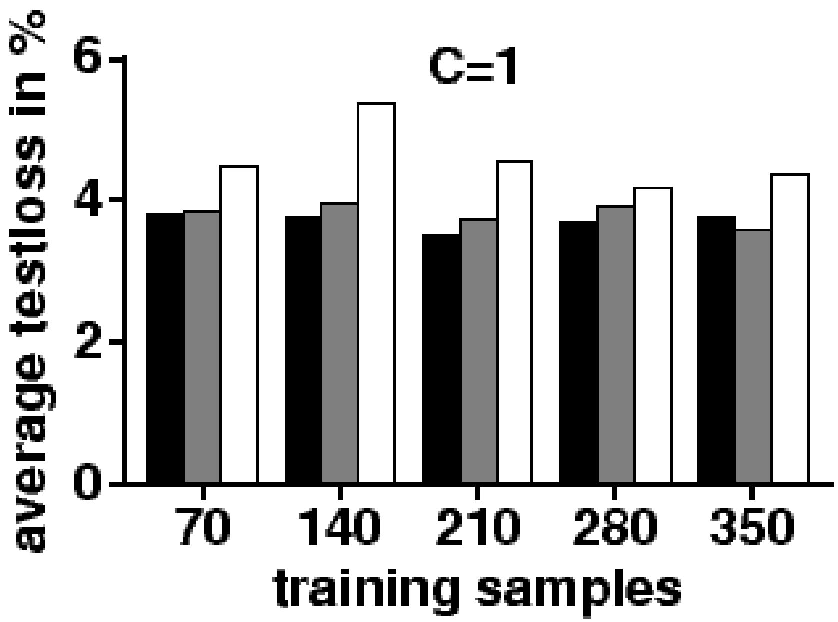
- Reweighting greatly improves performance
- KMM outperforms IS at small sample sizes



Breast Cancer data

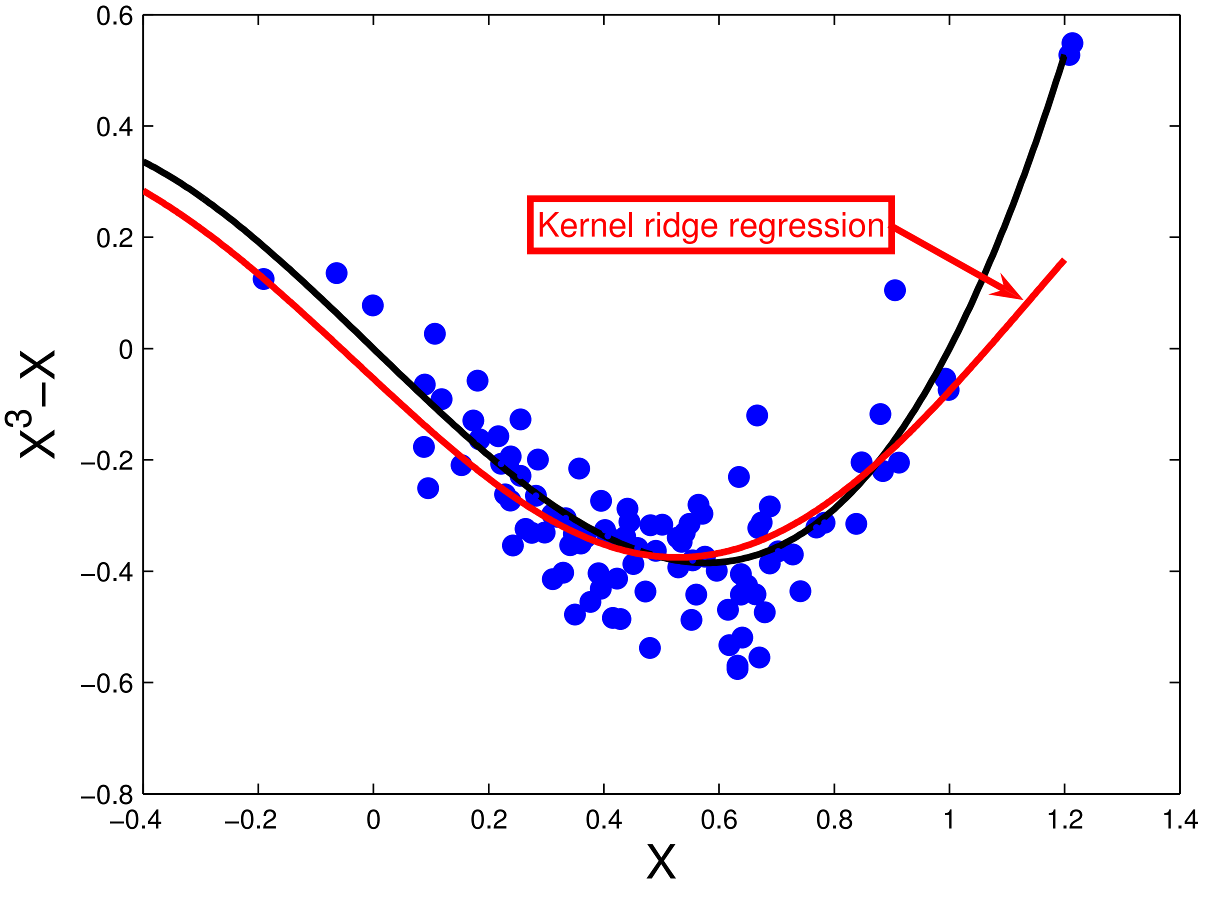
- KMM slightly decreases performance
- IS does not help





Toy example revisited

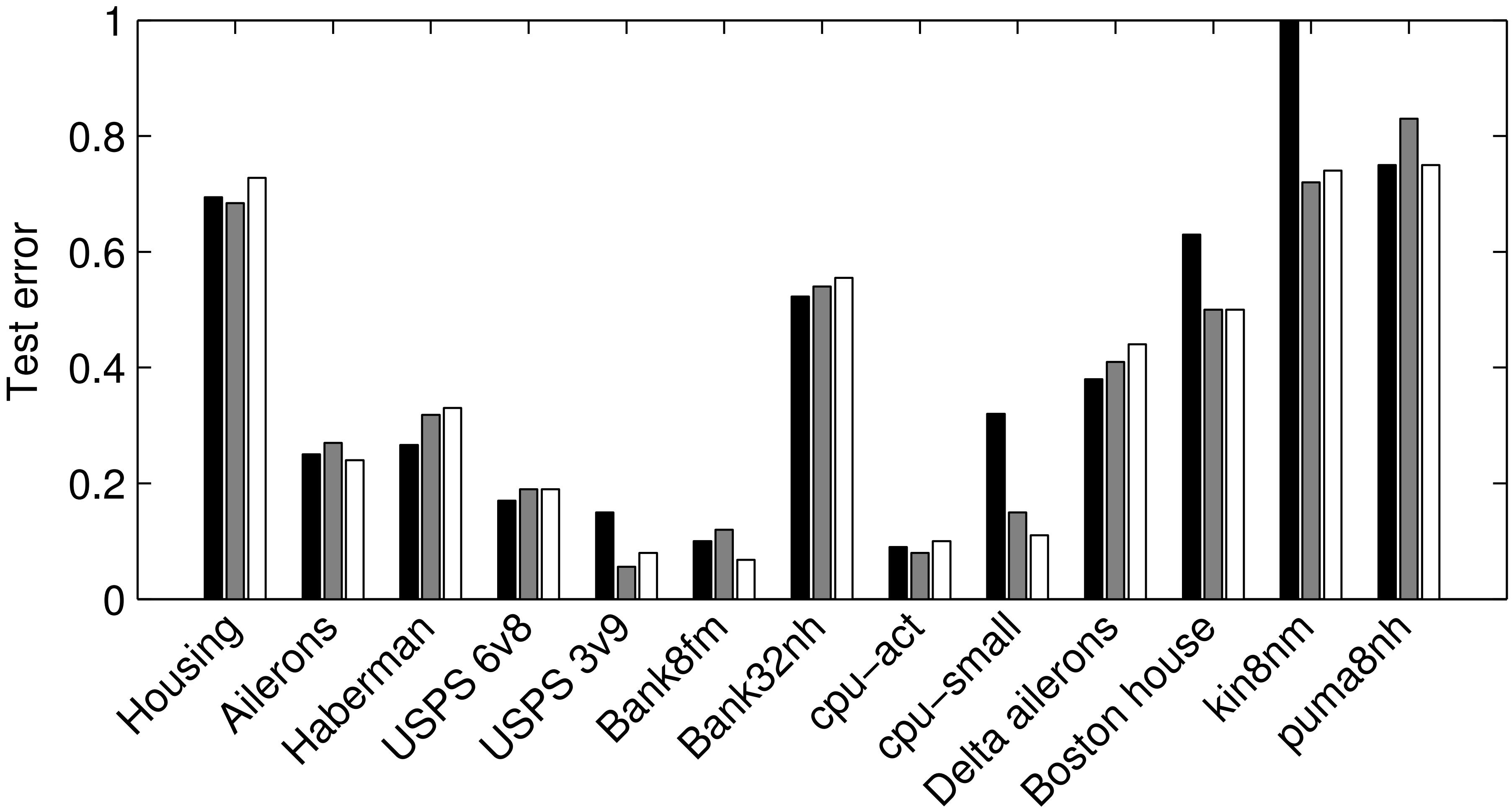
- Kernel ridge regression result



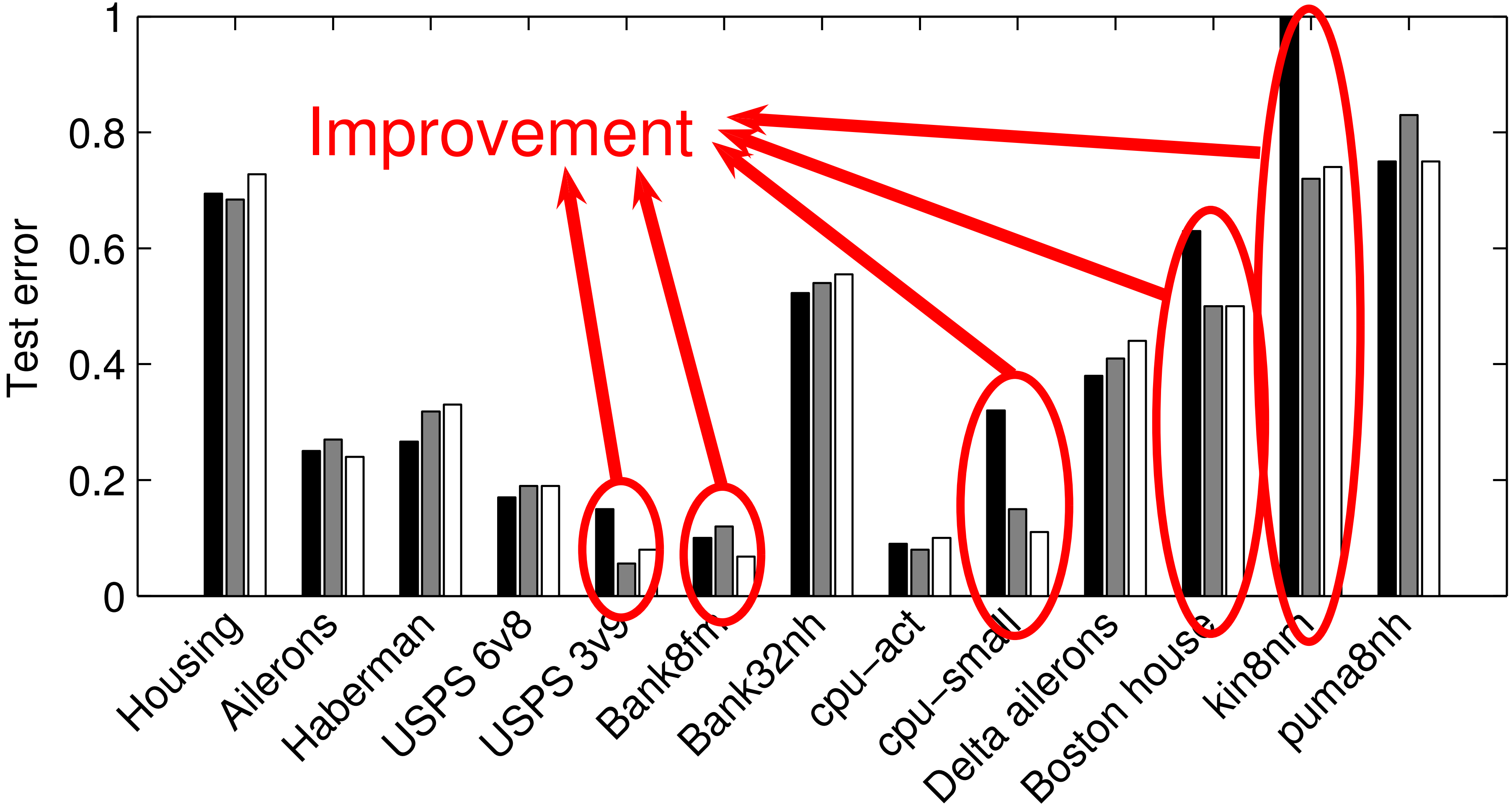
Large scale experiments

- 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

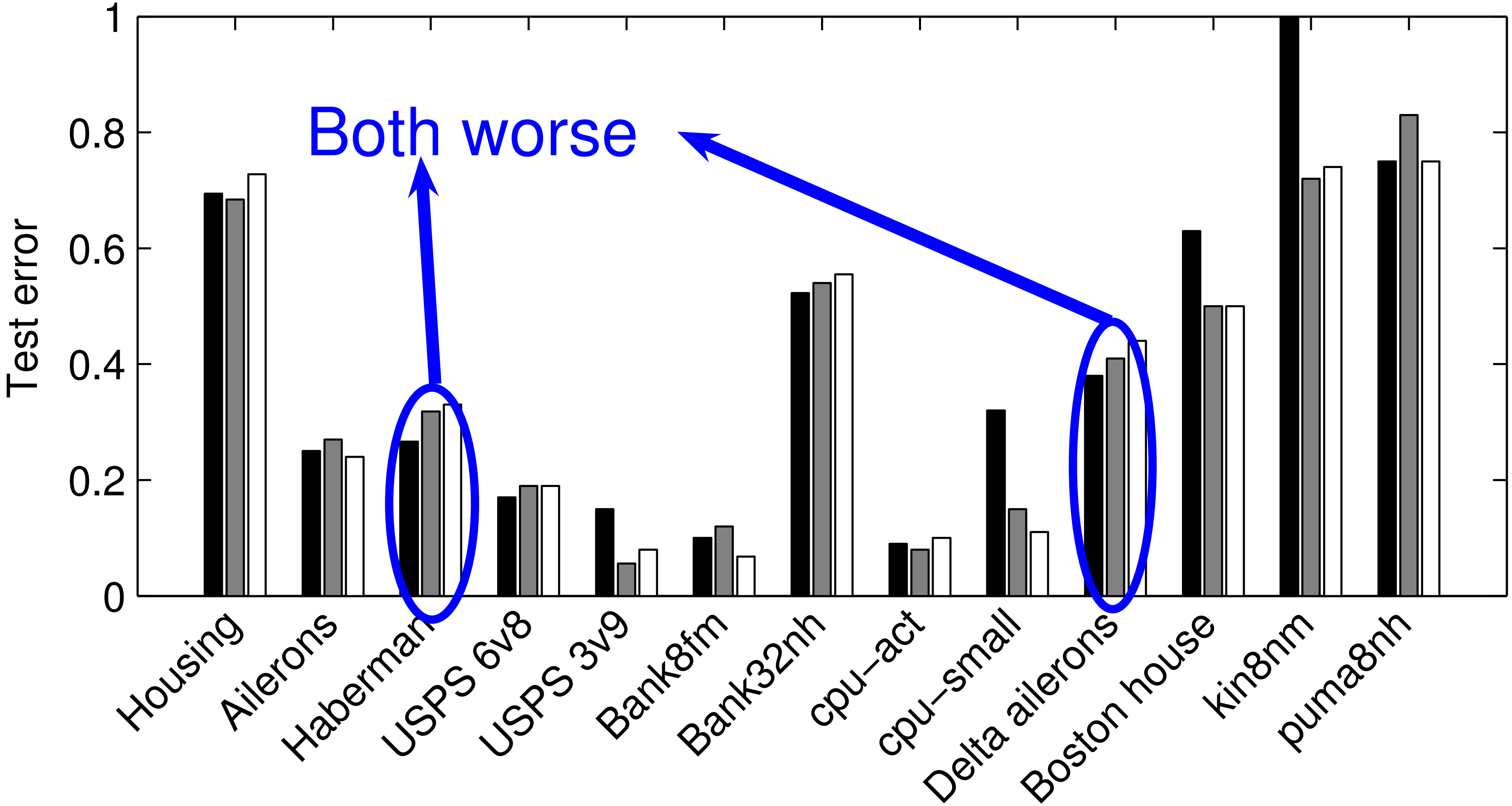
Large scale experiments



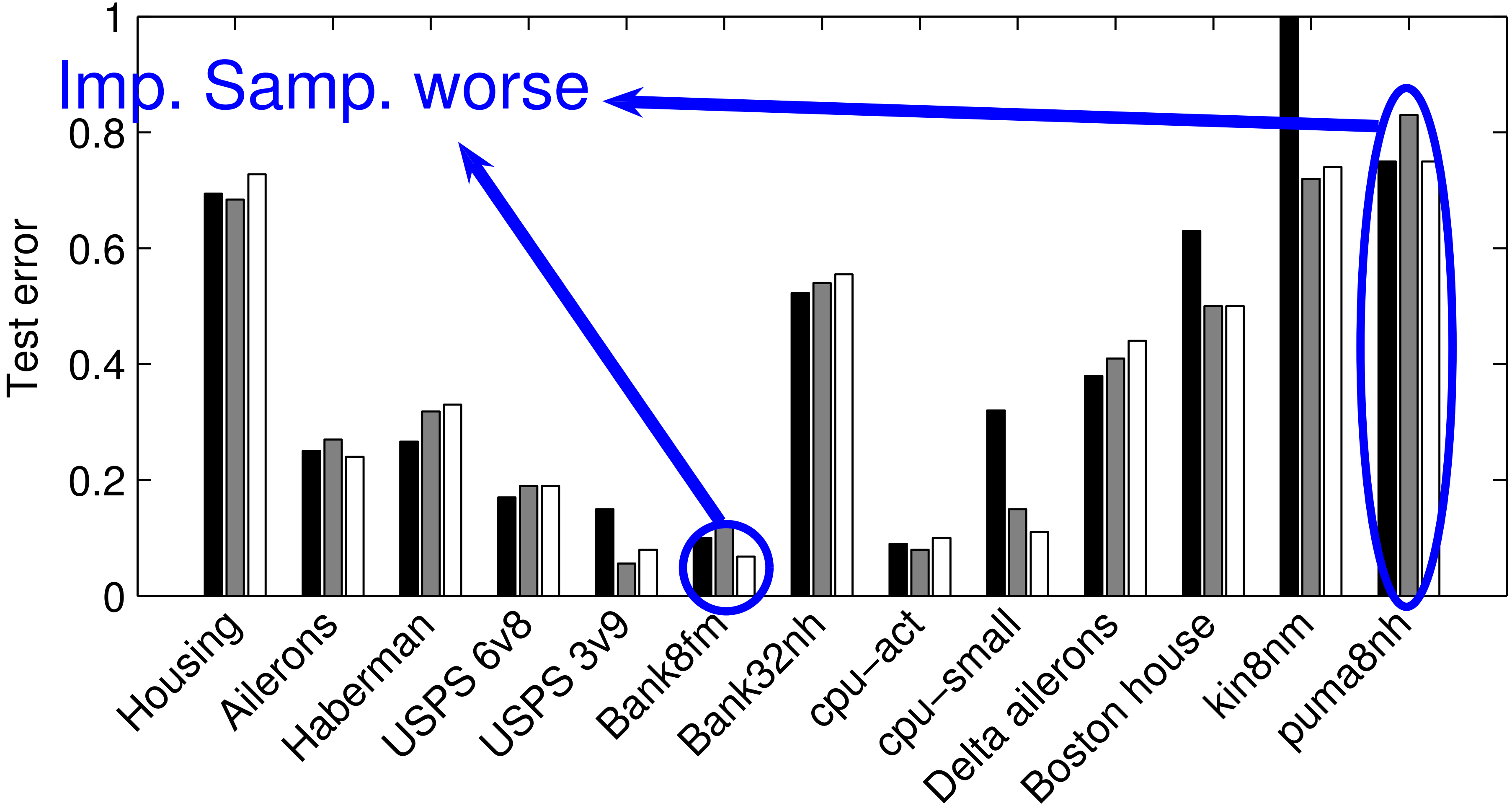
Large scale experiments



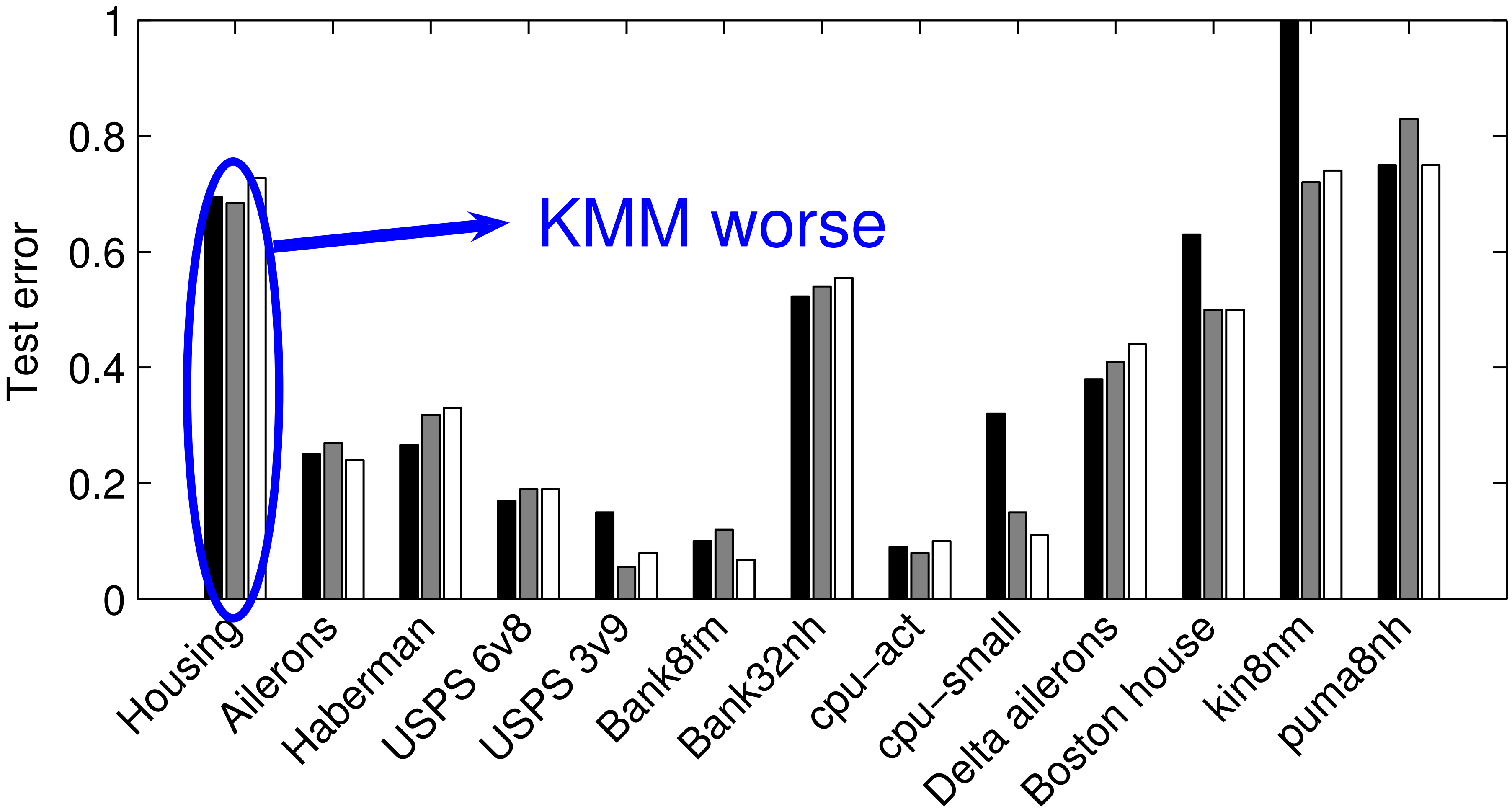
Large scale experiments



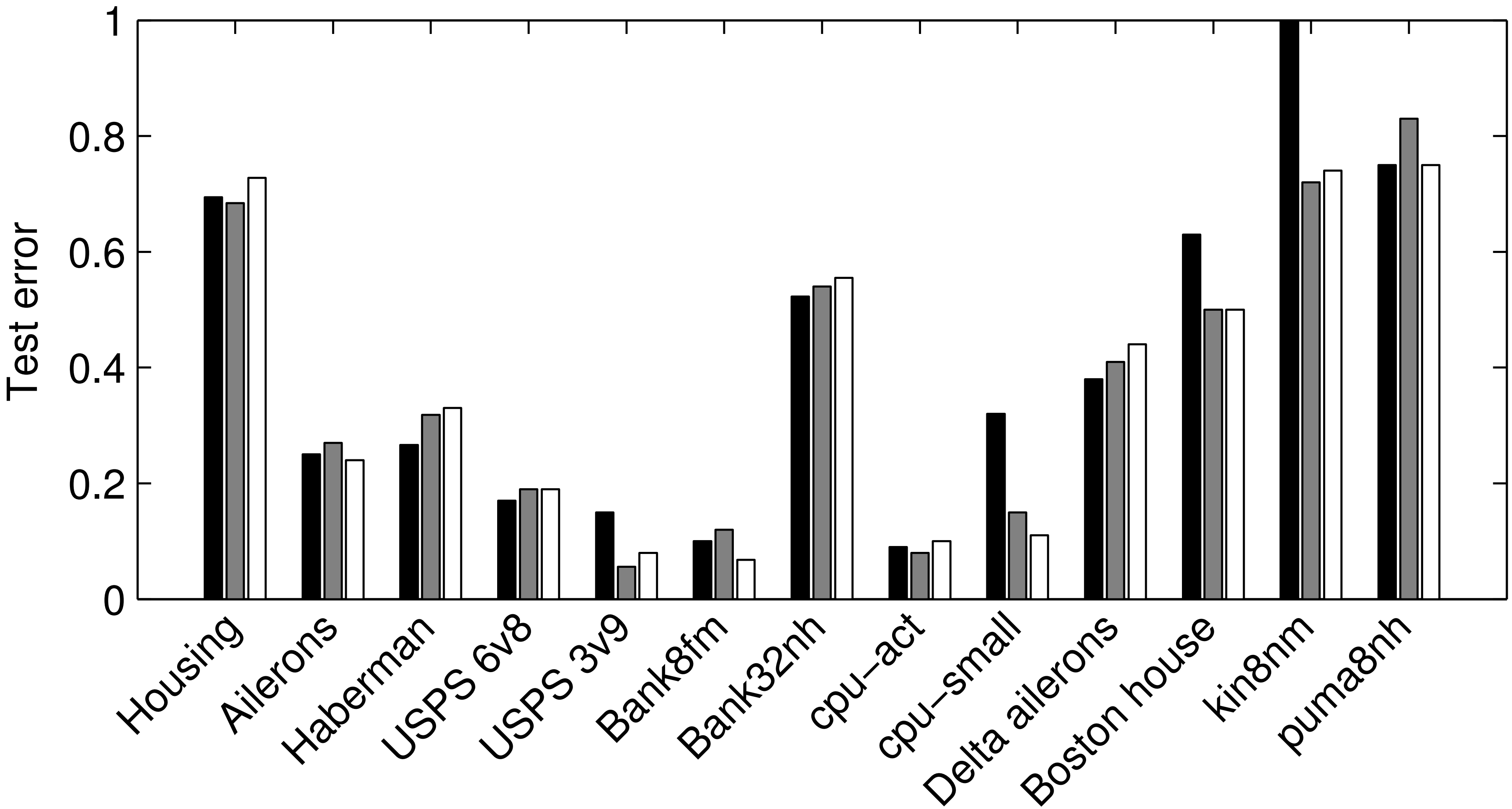
Large scale experiments



Large scale experiments



Large scale experiments



Conclusion

- KMM: perform covariate shift without density estimation
- Large performance advantage for "simple" learning algorithms
- Mixed results for powerful learning algorithms
- Model selection remains an issue

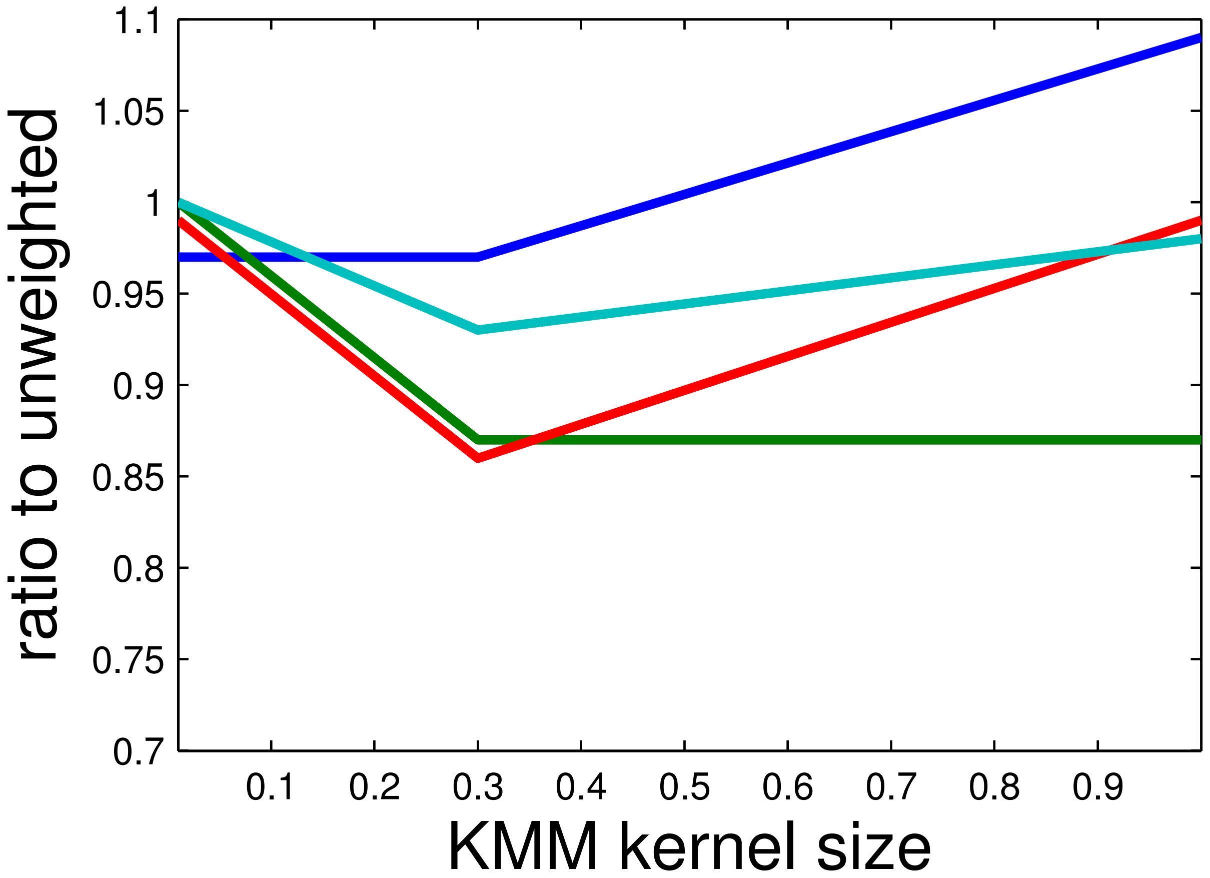
Acknowledgements

- 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?

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? [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? [Sriperumbudur et al., 2010]
- Does knowing something about the learning problem help?

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

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.

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 Machine Learning Research*, 10:1391–1445, 2009.

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:

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

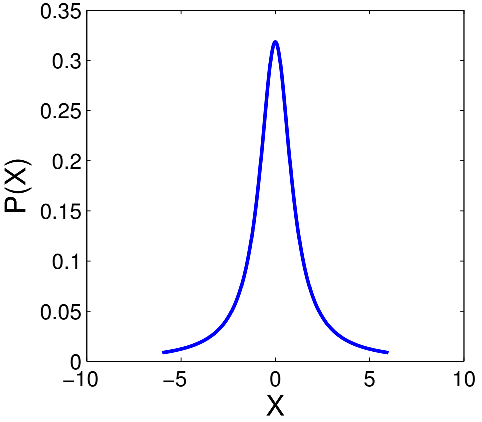
  – $\phi_{\mathbf{P}}$ characteristic function of $\mathbf{P}$

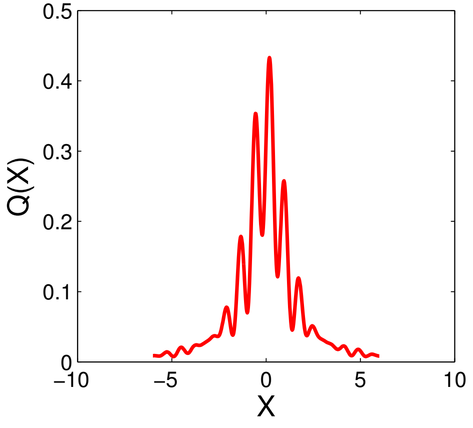
  – $f^{\wedge}$ is Fourier transform, $f^{\vee}$ is inverse Fourier transform

  – $\mu_x := \int k(\cdot, x)\, d\mathbf{P}(x)$

Characteristic Kernels (2)

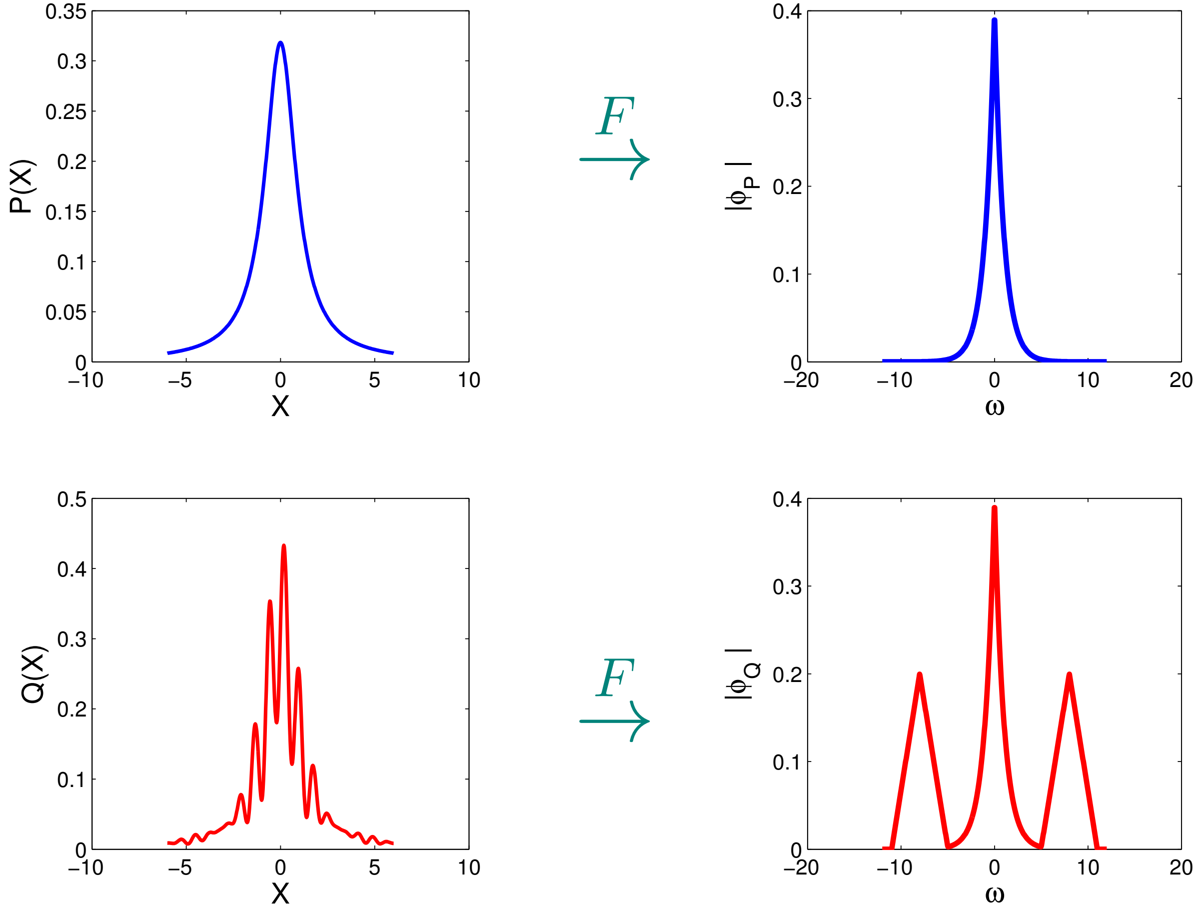
- Example: **P** differs from **Q** at (roughly) one frequency





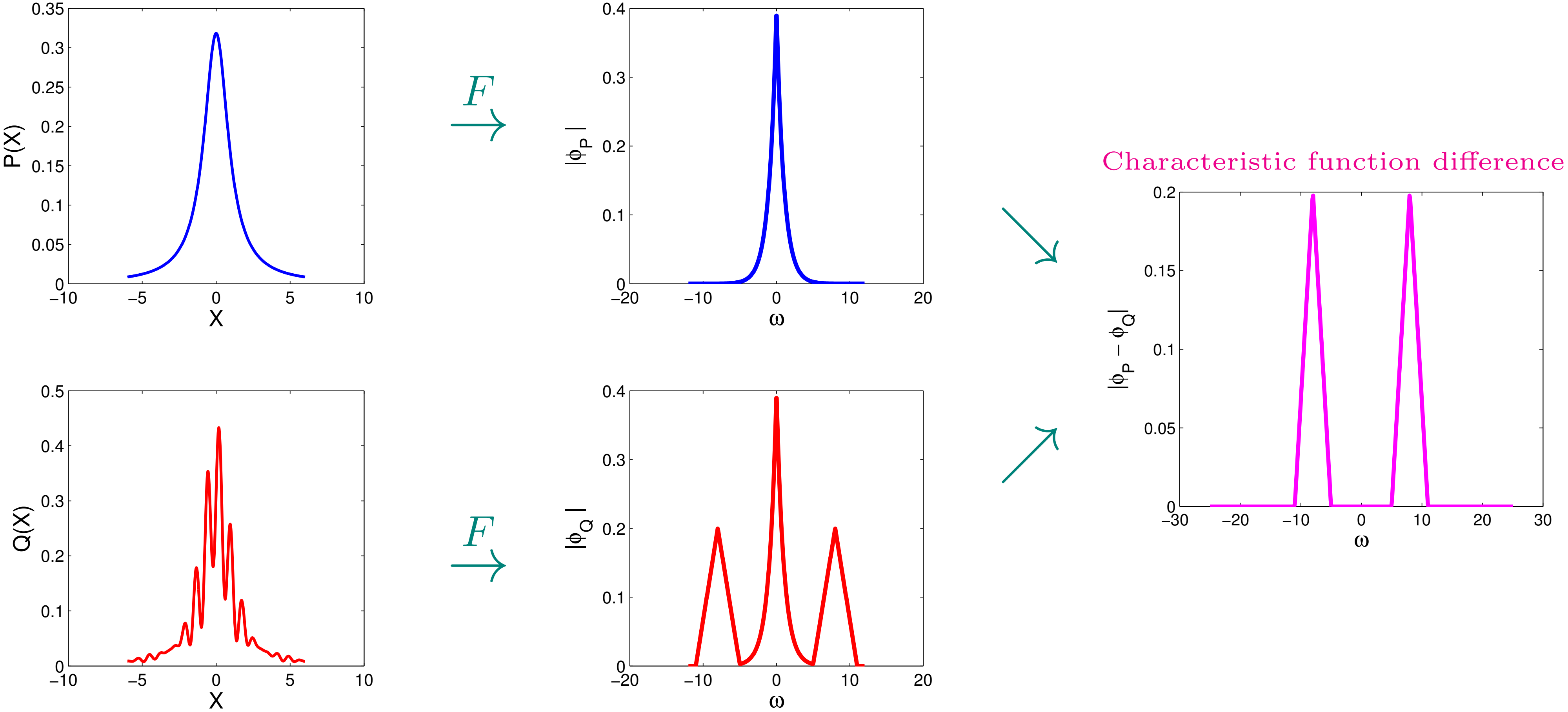
Characteristic Kernels (2)

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



Characteristic Kernels (2)

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

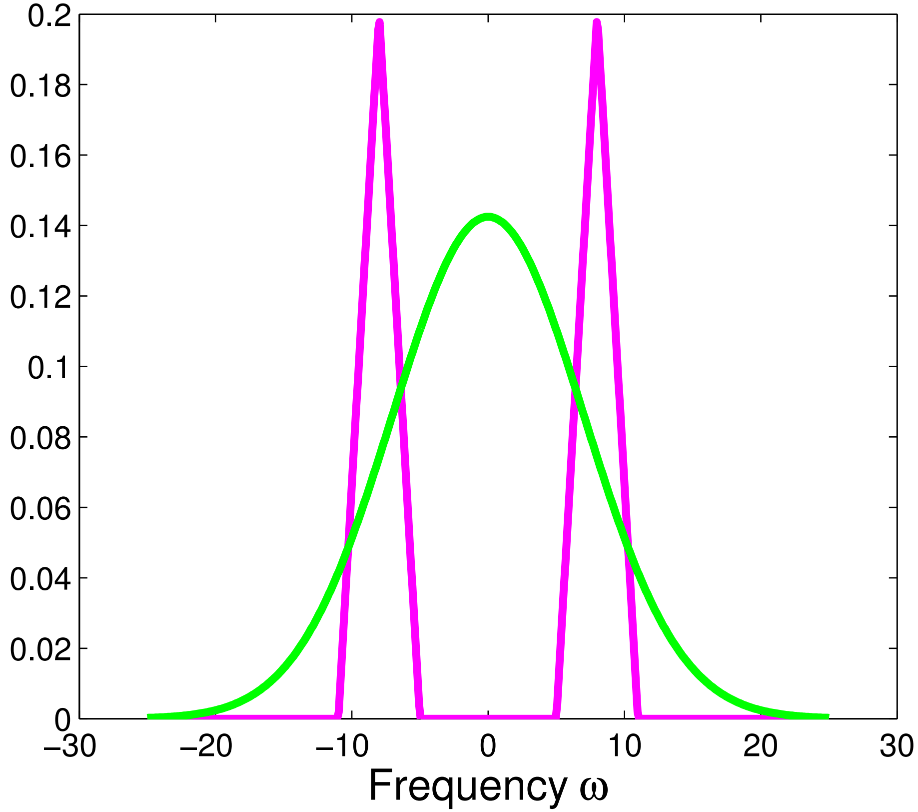


Characteristic Kernels (3)

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

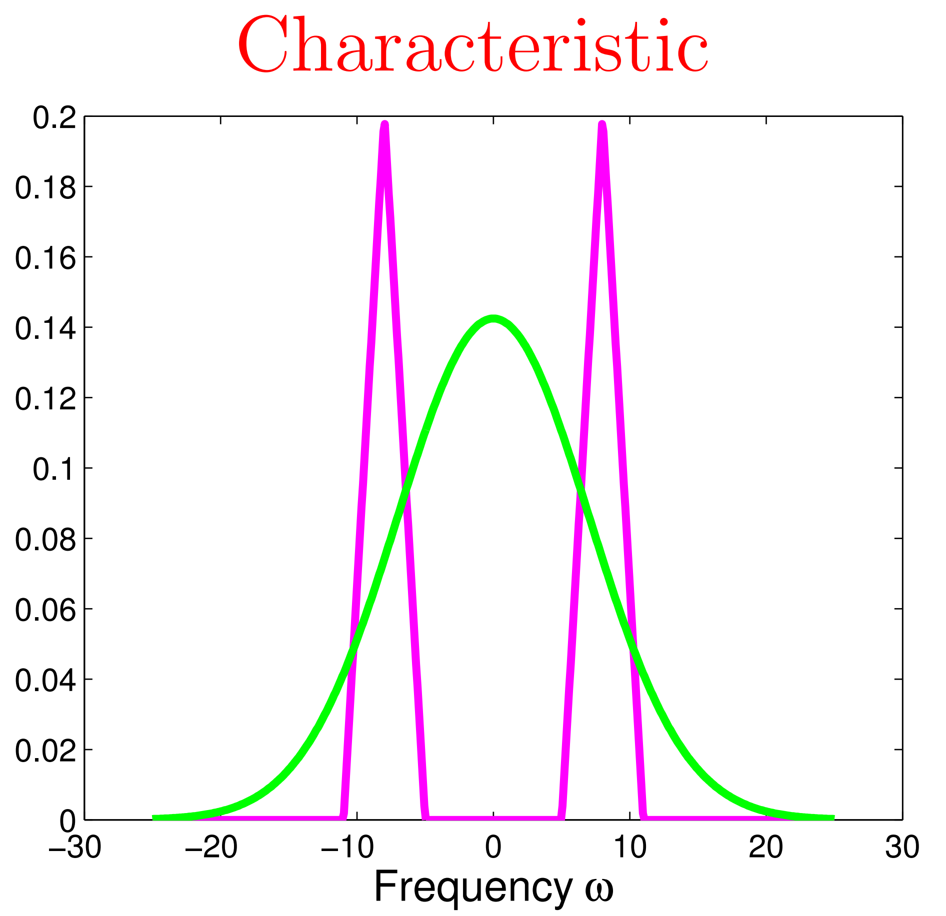
Gaussian kernel

Difference $|\phi_P - \phi_Q|$



Characteristic Kernels (3)

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

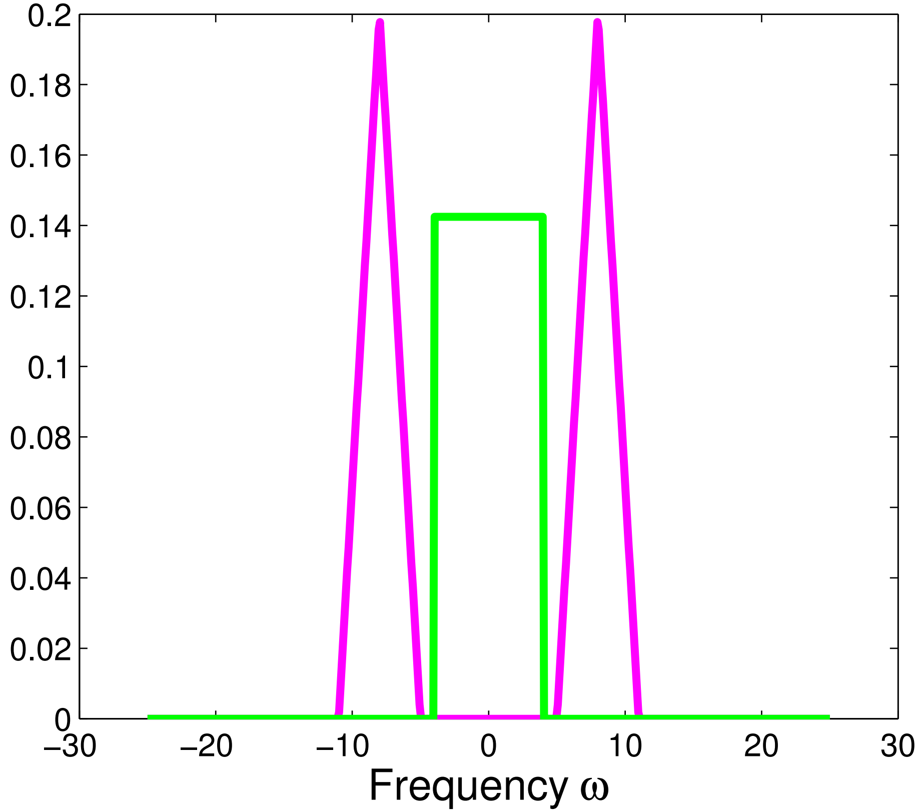


Characteristic Kernels (4)

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

Sinc kernel

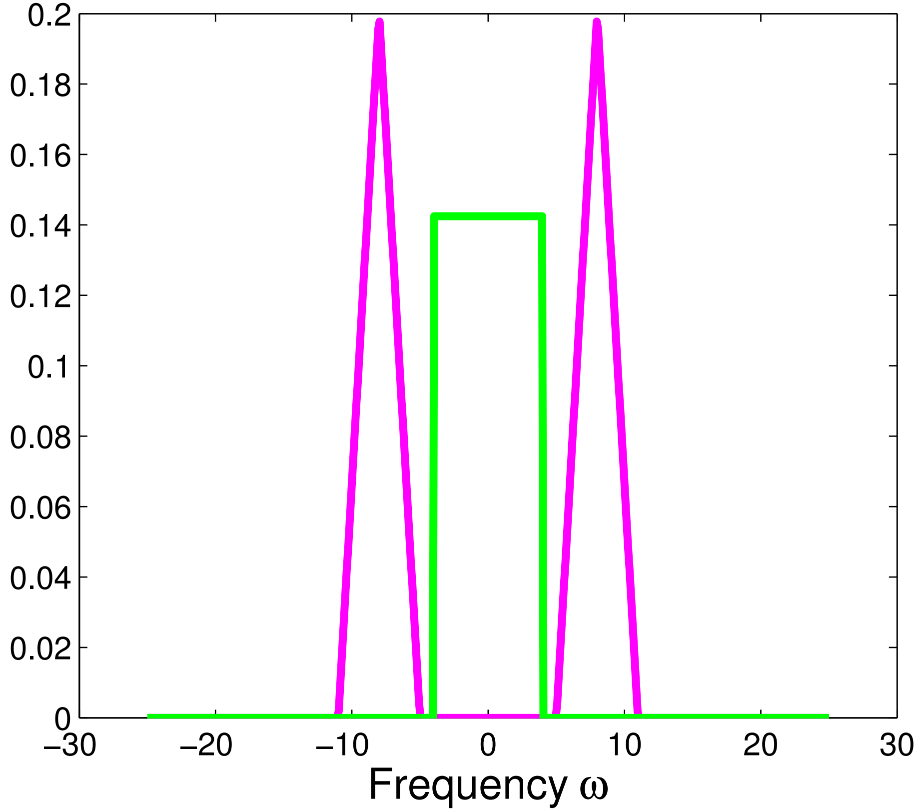
Difference $|\phi_P - \phi_Q|$



Characteristic Kernels (4)

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

NOT characteristic

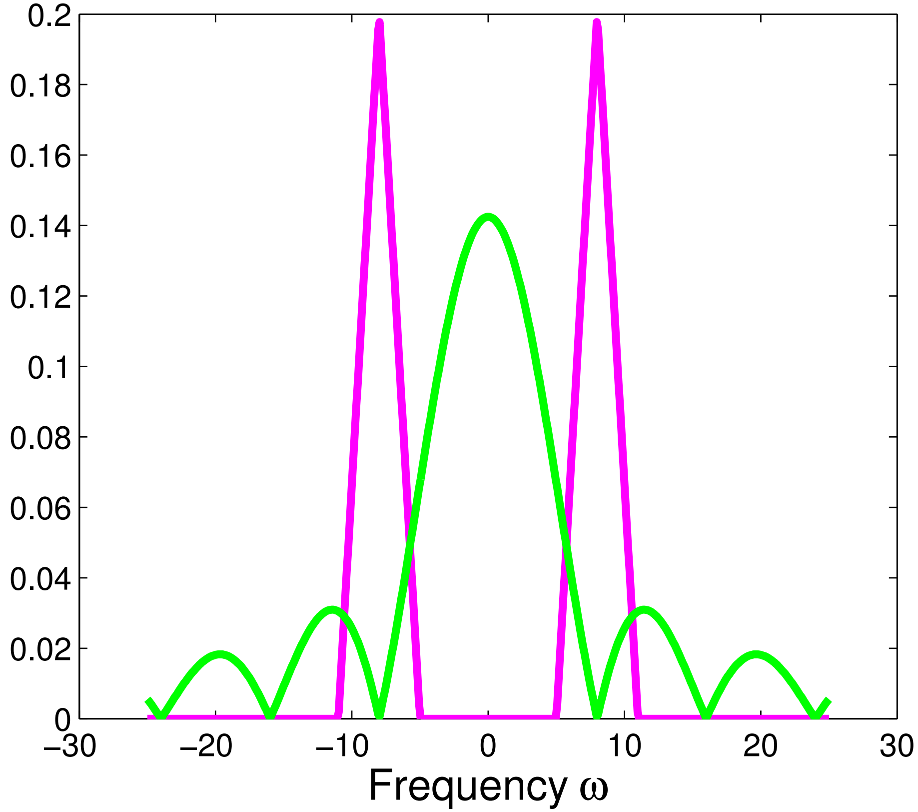


Characteristic Kernels (5)

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

B-Spline kernel

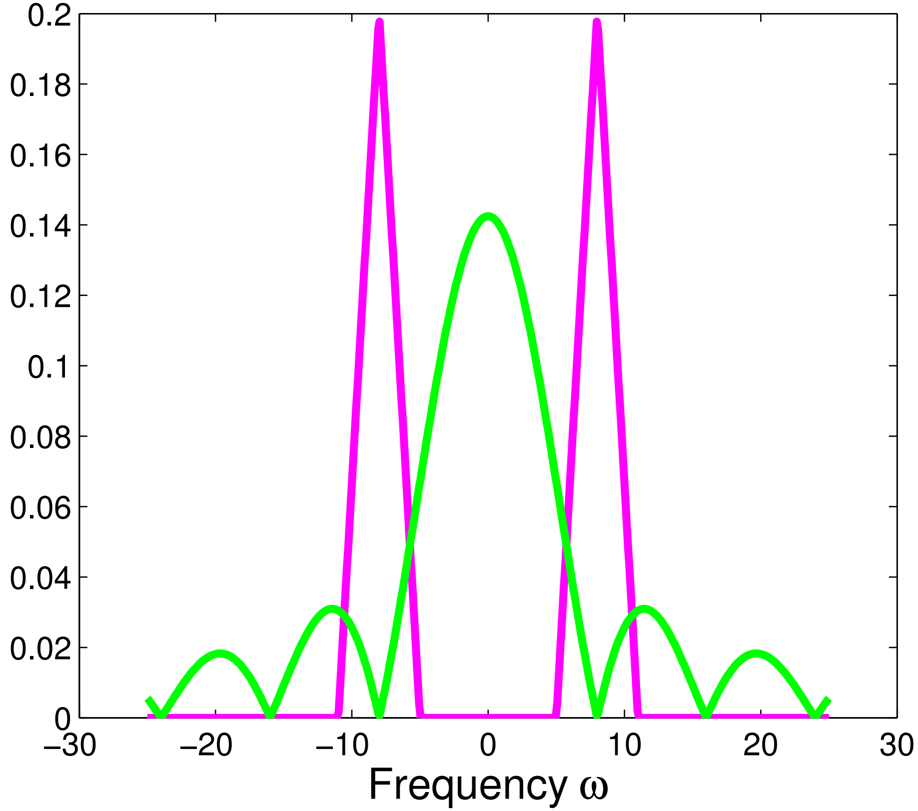
Difference $|\phi_P - \phi_Q|$



Characteristic Kernels (5)

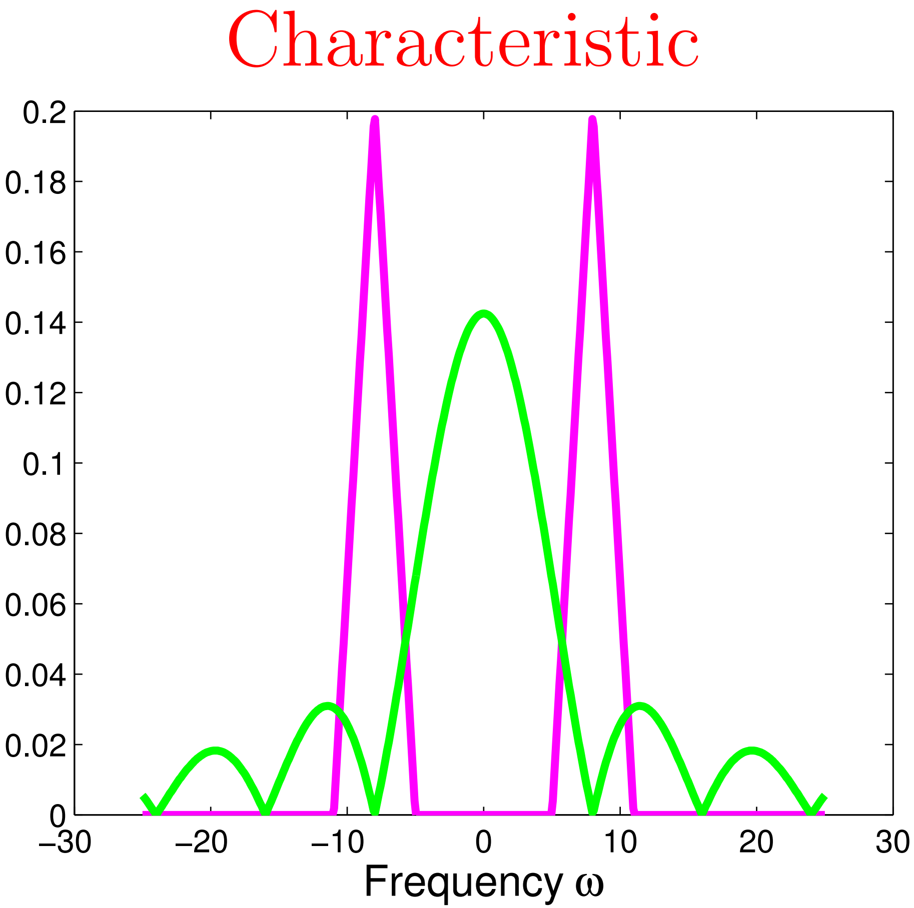
- Example: **P** differs from **Q** at (roughly) one frequency

???



Characteristic Kernels (5)

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



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]

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

Summary: Characteristic Kernels

- Characteristic kernel: (MMD = 0 iff **P** = **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?]

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

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)