On Regularized Kernel Mean Embedding and Stein's Phenomenon

Krikamol Muandet
Empirical Inference Department, MPI-IS

Joint work with

Krik
MPI-IS

Bernhard
MPI-IS

Kenji
ISM

Arthur
UCL

Bharath
Cambridge

Estimation Problem

German tank problem

Estimation Problem

Estimation Theory

An estimation of the value of an unknown parameter from a set of observations of a random variable.

- X : a random variable
- Θ : parameter space
- $\{\mathbb{P}_\theta, \theta \in \Theta\}$: a set of probability laws indexed by θ
- $\delta : \mathcal{X} \to \Theta$: an estimator

Estimation Problem

Estimation Theory

An estimation of the value of an unknown parameter from a set of observations of a random variable.

- X : a random variable
- Θ : parameter space
- $\{\mathbb{P}_\theta, \theta \in \Theta\}$: a set of probability laws indexed by θ
- $\delta : \mathcal{X} \to \Theta$: an estimator

Objective

Given observations $x_1, x_2, \ldots, x_n$, find an approximate value of parameter θ.

$$\delta(x_1, x_2, \ldots, x_n) = \hat{\theta} \quad \text{such that} \quad \hat{\theta} \approx \theta$$

Kernel Mean Embedding

A kernel mean embedding (KME) of a distribution $\mathbb{P}$ is

$$\mu : \mathscr{P} \to \mathcal{H}, \quad \mathbb{P} \mapsto \int_{\mathcal{X}} K(x, \cdot)\, d\mathbb{P}(x) \tag{1}$$

An empirical estimate can be obtained from a random sample $X_1, X_2, \ldots, X_n$ as

$$\widehat{\mu} : \widehat{P}_n \mapsto \frac{1}{n} \sum_{i=1}^{n} K(X_i, \cdot) \tag{2}$$

Kernel Mean Embedding

A kernel mean embedding (KME) of a distribution $\mathbb{P}$ is

$$\mu : \mathscr{P} \to \mathcal{H}, \quad \mathbb{P} \mapsto \int_{\mathcal{X}} K(x, \cdot)\, d\mathbb{P}(x) \tag{1}$$

An empirical estimate can be obtained from a random sample $X_1, X_2, \ldots, X_n$ as

$$\widehat{\mu} : \widehat{P}_n \mapsto \frac{1}{n} \sum_{i=1}^{n} K(X_i, \cdot) \tag{2}$$

The estimator we are using is

$$\delta(K(X, \cdot)) = K(X, \cdot)$$

Kernel Mean Embedding

A kernel mean embedding (KME) of a distribution $\mathbb{P}$ is

$$\mu : \mathscr{P} \to \mathcal{H}, \quad \mathbb{P} \mapsto \int_{\mathcal{X}} K(x, \cdot)\, d\mathbb{P}(x) \tag{1}$$

An empirical estimate can be obtained from a random sample $X_1, X_2, \ldots, X_n$ as

$$\widehat{\mu} : \widehat{P}_n \mapsto \frac{1}{n} \sum_{i=1}^{n} K(X_i, \cdot) \tag{2}$$

The estimator we are using is

$$\delta(K(X, \cdot)) = K(X, \cdot)$$

Is this the best we can do?

Applications of Kernel Mean Embedding

- Maximum mean discrepancy (MMD)
- Kernel Bayes rule
- Dependency measure (HSIC)
- Supervised feature selection
- Kernel-based hypothesis testing
- Hilbert space embedding of conditional distribution
- Kernel belief propagation
- Hilbert space embedding of POMDPs
- Supervised learning on distributions, e.g., support measure machines and domain generalization
- etc.

Stein's Phenomenon

Consider the problem of estimating the mean of the Gaussian distribution

$$X \sim \mathcal{N}(\theta, \sigma^2 I)$$

- If σ^2 is known, the James-Stein estimator is given by

$$\hat{\theta}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{\|X\|^2}\right) X$$

- If $d \geq 3$, then $\hat{\theta}_{JS}$ dominates the maximum likelihood estimator

$$\hat{\theta}_{ML} = X$$

- In other words, the maximum likelihood estimator is inadmissible.
- The James-Stein estimator is itself inadmissible.
- Shrinkage estimator $\hat{\theta}_{shrink} = \gamma\tilde{\theta} + (1-\gamma)\hat{\theta}_{ML}$ where $\gamma \in [0, 1]$.

Admissibility

Theorem

The usual estimator $\delta(K(X,\cdot)) = K(X,\cdot)$ *of kernel mean embedding is* **inadmissible** *with respect to the RKHS norm.*

Proof.

proof sketch:

1. the kernel mean μ is the mean of the Gaussian measure $\mathcal{N}(\mu, C)$ on the Hilbert space with the covariance operator C.
2. By projecting functions in Hilbert space onto the bases formed by the eigenvectors of C, we obtain a sequence of independent random variables. Each of which is distributed according to the Gaussian distribution with unit variance.
3. Apply trick similar to finite dimensional case.

□

Admissibility

Theorem

The usual estimator $\delta(K(X,\cdot)) = K(X,\cdot)$ *of kernel mean embedding is* **inadmissible** *with respect to the RKHS norm.*

Proof.

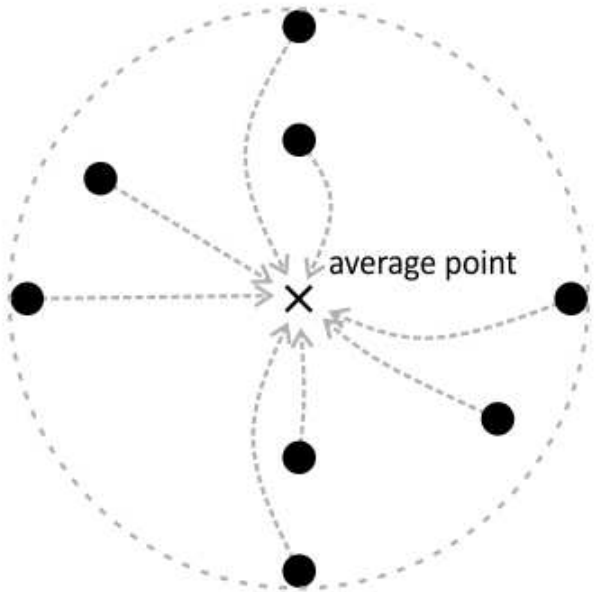
proof sketch:

1. the kernel mean μ is the mean of the Gaussian measure $\mathcal{N}(\mu, C)$ on the Hilbert space with the covariance operator C.
2. By projecting functions in Hilbert space onto the bases formed by the eigenvectors of C, we obtain a sequence of independent random variables. Each of which is distributed according to the Gaussian distribution with unit variance.
3. Apply trick similar to finite dimensional case.

□

How to construct a better estimator?

Regularized Kernel Mean Embedding



The usual estimate can be obtained by minimizing

$$\frac{1}{2}\sum_{i=1}^{n}\|K(X_i,\cdot)-g\|_{\mathcal{H}}^2 \tag{3}$$

Regularized Kernel Mean Embedding

First, we assume that $g = \sum_{i=1}^{n} \beta_i K(X_i, \cdot)$. Then, we consider the regularized version of (3):

$$\frac{1}{2} \sum_{i=1}^{n} \left\| K(X_i, \cdot) - \sum_{j=1}^{n} \beta_j K(X_j, \cdot) \right\|_{\mathcal{H}}^2 + \frac{\lambda}{2} \boldsymbol{\beta}^\top \boldsymbol{\beta} \tag{4}$$

Regularized Kernel Mean Embedding

First, we assume that $g = \sum_{i=1}^{n} \beta_i K(X_i, \cdot)$. Then, we consider the regularized version of (3):

$$\frac{1}{2} \sum_{i=1}^{n} \left\| K(X_i, \cdot) - \sum_{j=1}^{n} \beta_j K(X_j, \cdot) \right\|_{\mathcal{H}}^2 + \frac{\lambda}{2} \boldsymbol{\beta}^\top \boldsymbol{\beta} \tag{4}$$

Minimizing (4) w.r.t. $\boldsymbol{\beta}$ yields

$$\boldsymbol{\beta} = (\mathbf{K} + \lambda_n \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n \tag{5}$$

Regularized Kernel Mean Embedding

First, we assume that $g = \sum_{i=1}^{n} \beta_i K(X_i, \cdot)$. Then, we consider the regularized version of (3):

$$\frac{1}{2} \sum_{i=1}^{n} \left\| K(X_i, \cdot) - \sum_{j=1}^{n} \beta_j K(X_j, \cdot) \right\|_{\mathcal{H}}^2 + \frac{\lambda}{2} \boldsymbol{\beta}^\top \boldsymbol{\beta} \tag{4}$$

Minimizing (4) w.r.t. $\boldsymbol{\beta}$ yields

$$\boldsymbol{\beta} = (\mathbf{K} + \lambda_n \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n \tag{5}$$

Thus, the regularized kernel mean embedding (RKME) is

$$\hat{\mu}_{RKME} = \sum_{i=1}^{n} \beta_i K(X_i, \cdot) = \boldsymbol{\Phi}_{\mathbf{x}}^\top (\mathbf{K} + \lambda_n \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n$$

where $\boldsymbol{\Phi}_{\mathbf{x}} = [K(X_1, \cdot), K(X_2, \cdot), \ldots, K(X_n, \cdot)]^\top$.

Experiments

Distributions

1. Gaussian Distribution $\mathcal{N}(\theta, C)$
2. Mixture of Gaussians Distribution $\sum_{k=1}^{m} \pi_k \mathcal{N}(\theta_k, C_k)$

Loss function

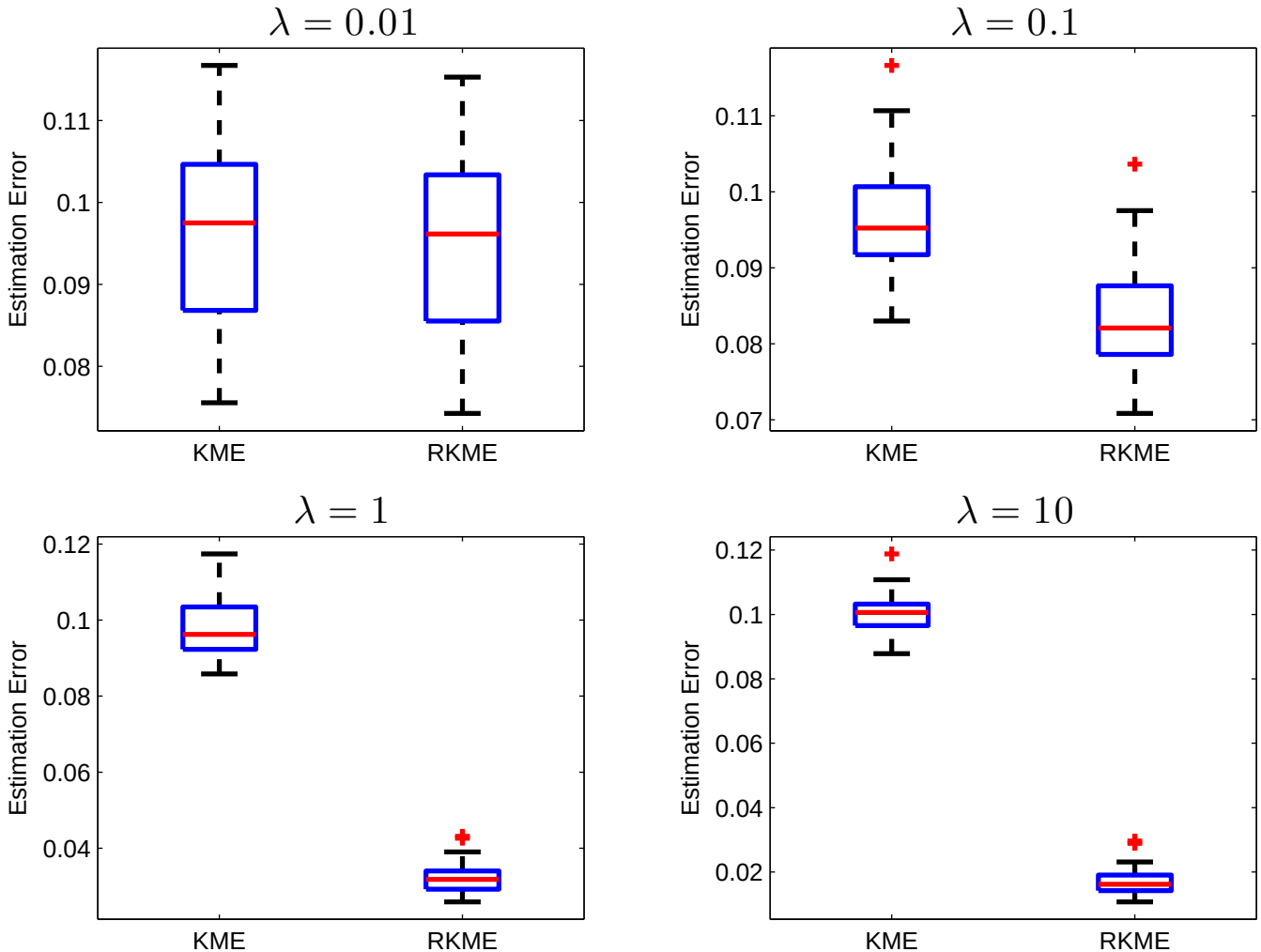
$$L(\beta) = \left\| \sum_{i=1}^{n} \beta_i K(X_i, \cdot) - \mathbb{E}_{\mathbb{P}}[K(X_i, \cdot)] \right\|_{\mathcal{H}}^2$$

Parameters

1. Gaussian RBF kernel $K(X, X') = \exp(-0.5 \cdot \|X - X'\|^2/\sigma^2)$
2. Bandwidth parameter $\sigma^2 = \mathit{median}\{\|X - X'\|^2\}$
3. Regularization (shrinkage) parameter λ

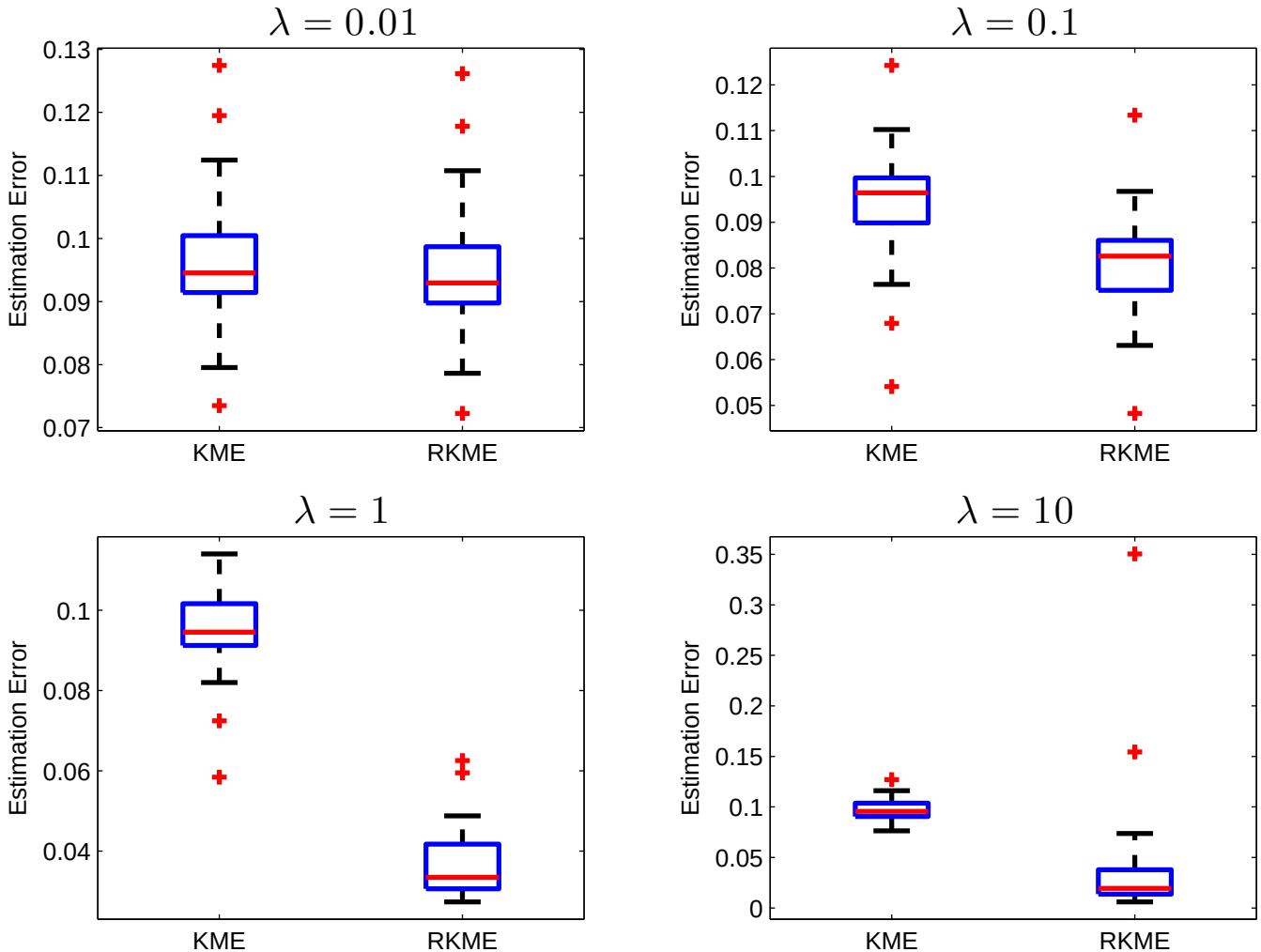
Experimental Results

Single Gaussian distributions (n=10,dim=20)



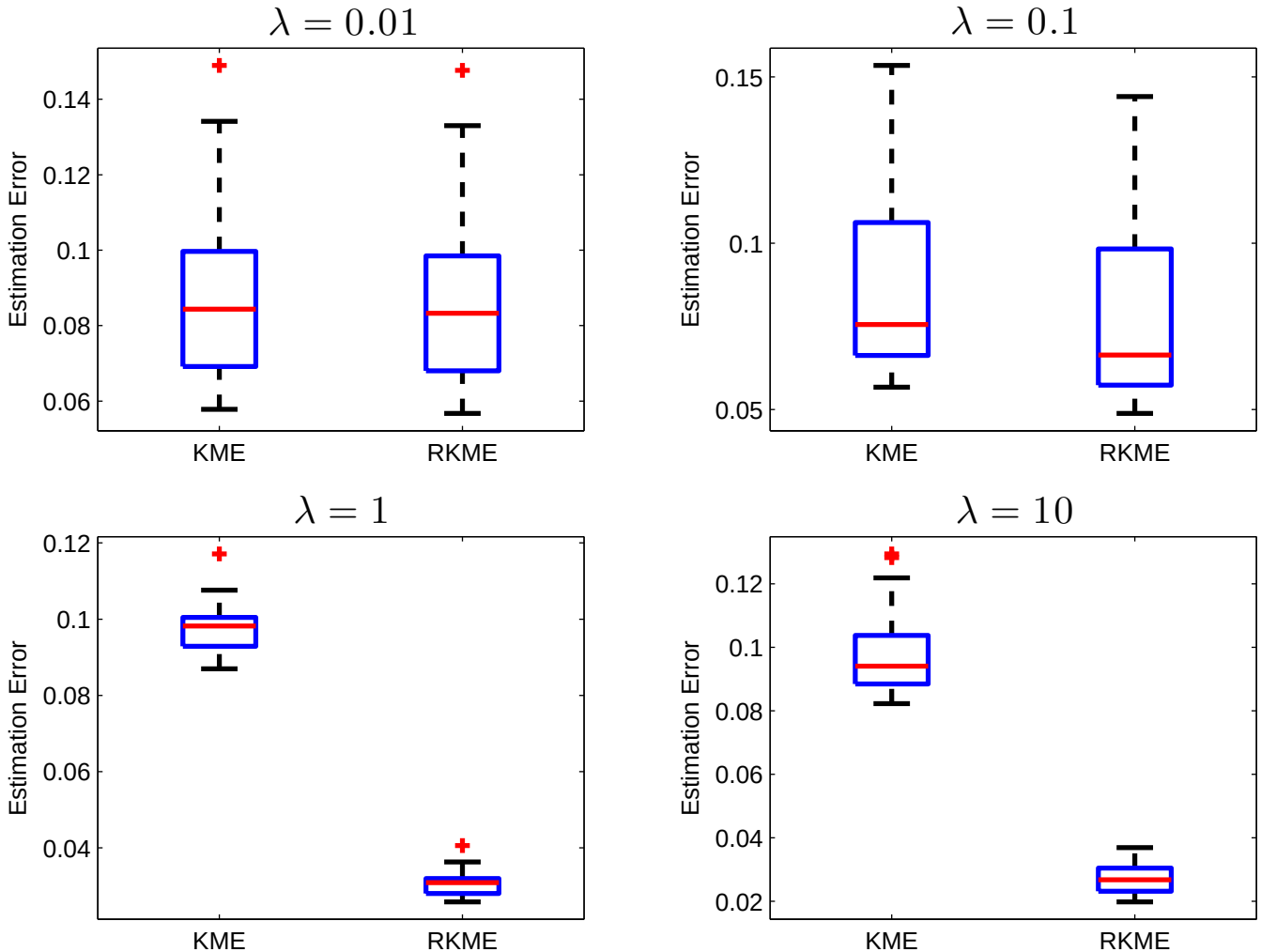
Experimental Results

Multiple Gaussian distributions (n=10,dim=20)



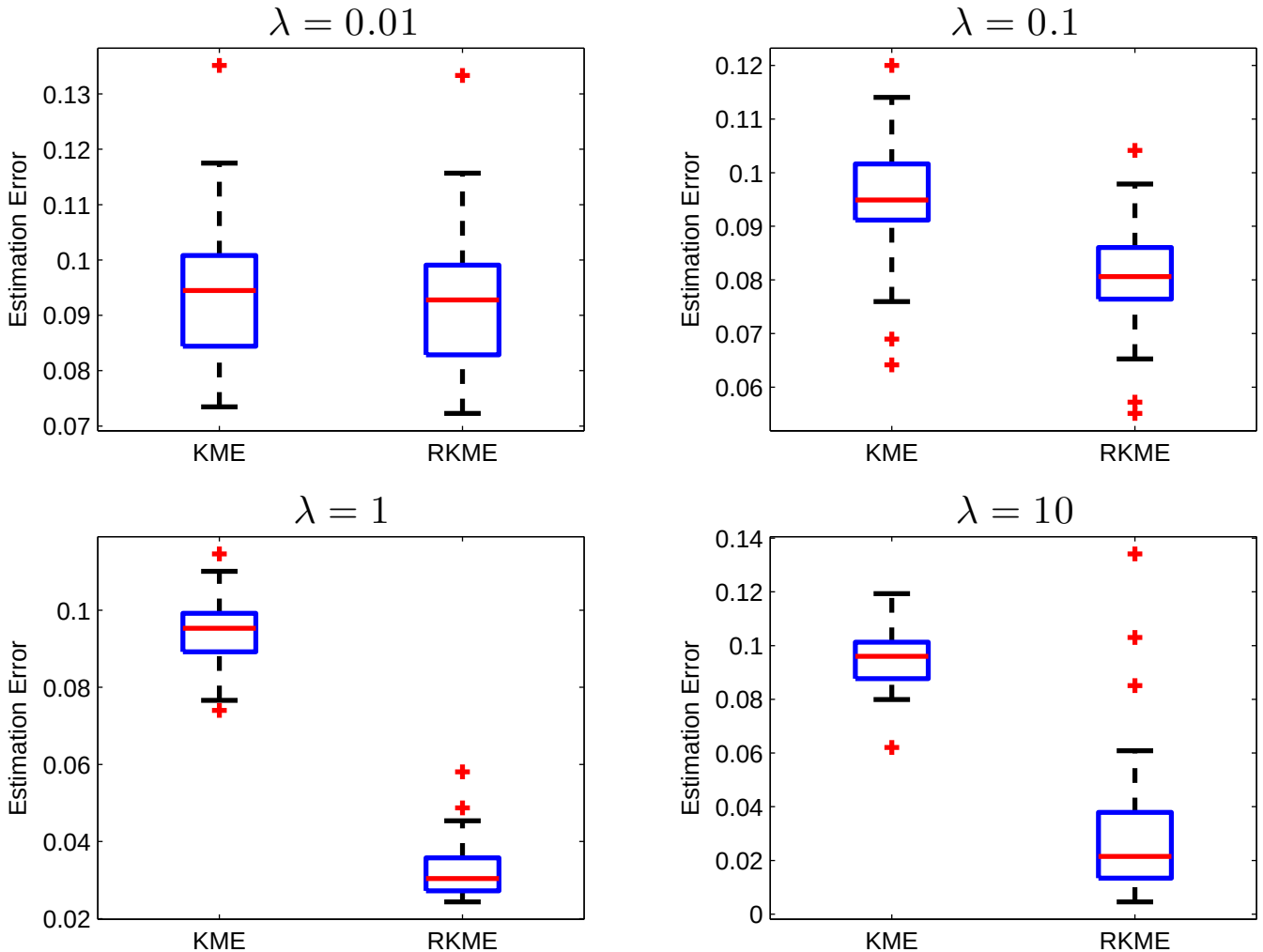
Experimental Results

Single MoG distributions (n=10,dim=20)



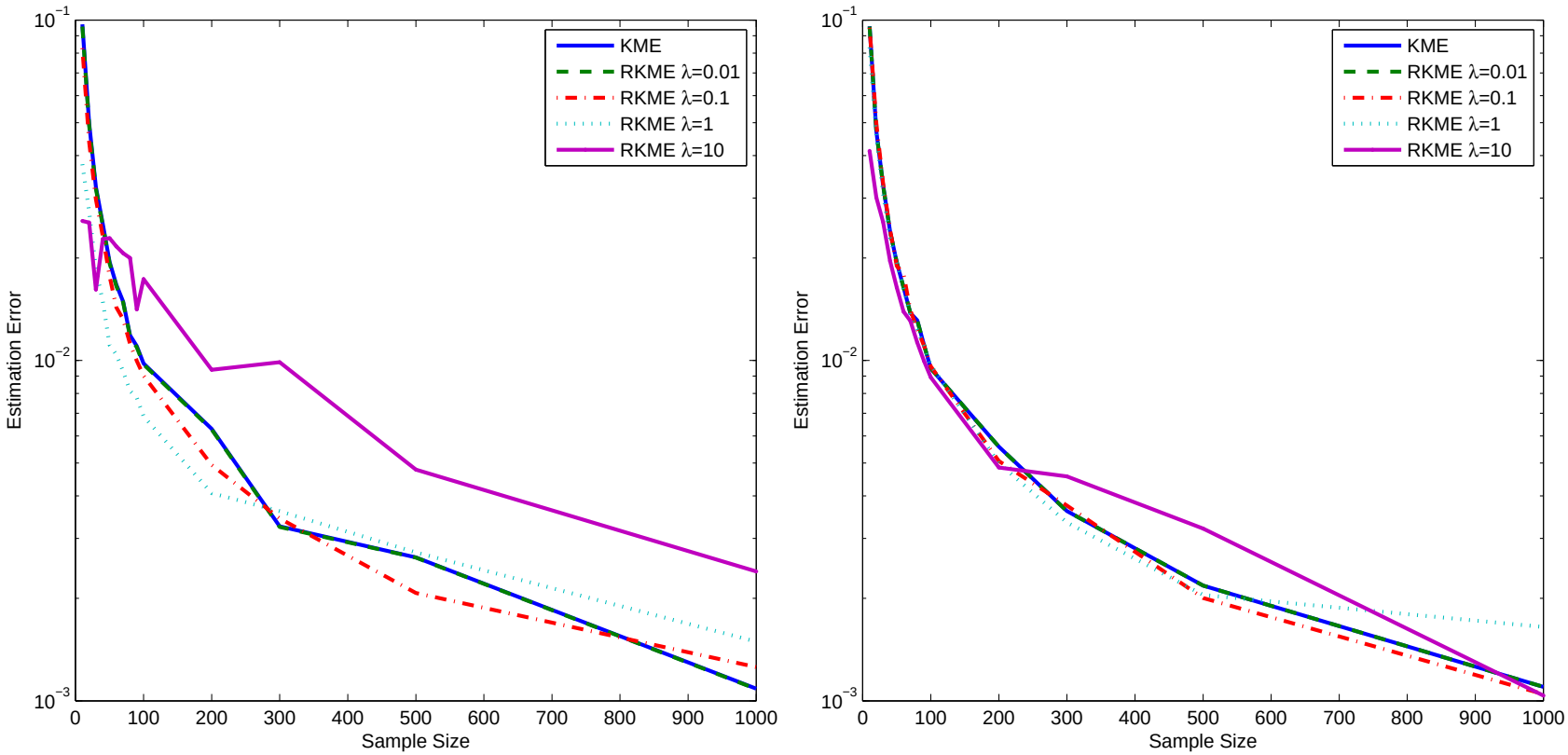
Experimental Results

Multiple MoG distributions (n=10,dim=20)



Experimental Results

Sample Size (number of experiments = 30)



Open Questions and Ongoing Works

- Is there an empirical Bayes estimator?
- Bayesian estimator of kernel mean embedding
- Covariance operator
- Different regularizers
- Theoretical analysis of the regularized kernel mean embedding
- Experiments on real-world data sets

Question?

Appendix

Let $K = UDU^\top$ be the spectral decomposition of K. The regularized KME can be written as

$$\begin{aligned}
\widehat{\mu}_{RKME} &= \mathbf{\Phi}_{\mathbf{x}}^\top (\mathbf{K} + \lambda_n \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n \\
&= \mathbf{\Phi}_{\mathbf{x}}^\top (UDU^\top + \lambda_n \mathbf{I})^{-1} UDU^\top \mathbf{1}_n \\
&= \mathbf{\Phi}_{\mathbf{x}}^\top (U(D + \lambda_n \mathbf{I})U^\top)^{-1} UDU^\top \mathbf{1}_n \\
&= \mathbf{\Phi}_{\mathbf{x}}^\top U(D + \lambda_n \mathbf{I})^{-1} DU^\top \mathbf{1}_n \\
&= \mathbf{\Phi}_{\mathbf{x}}^\top \sum_{i=1}^{n} \mathbf{u}_i \left(\frac{d_i}{d_i + \lambda_n} \right) \mathbf{u}_i^\top \mathbf{1}_n
\end{aligned}$$

As $\lambda > 0$, we have that $d_i/(d_i + \lambda_n) < 1$. Hence, the RKME is a shrinkage estimator.