Towards Outlier-Robust Statistical Inference on Kernel-Enriched Domains

Zoltán Szabó – CMAP, École Polytechnique

Joint work with:

- Matthieu Lerasle @ Paris-Sud University; CNRS
- Timothée Mathieu @ Paris-Sud University
- Guillaume Lecué @ ENSAE ParisTech

Pre-AISTATS ML Seminar
RIKEN Center for AI Project, Tokyo
Apr. 15, 2019

Kernel

- Kernel: $K(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}, \quad (\forall x, y \in \mathcal{X})$.

Kernel

- Kernel: $K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}, \quad (\forall x, y \in \mathcal{X})$.
  - Examples:

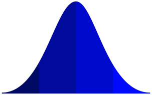
$$K(x, y) = e^{-\gamma \|x-y\|_2^2} \qquad \leftarrow \text{bounded},$$
$$K(x, y) = e^{\gamma \langle x, y \rangle} \qquad \leftarrow \text{unbounded}.$$

Kernel, RKHS

- Kernel: $K(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}, \quad (\forall x, y \in \mathcal{X})$.
  - Examples:

$$K(x,y) = e^{-\gamma \|x-y\|_2^2} \qquad \leftarrow \text{bounded},$$
$$K(x,y) = e^{\gamma \langle x,y \rangle} \qquad \leftarrow \text{unbounded}.$$

- RKHS: $\mathcal{H}_K = \overline{\{\sum_{i=1}^n \alpha_i K(\cdot, x_i)\}} \subset \mathbb{R}^{\mathcal{X}}$. $\varphi(x) = \underbrace{K(\cdot, x)} \in \mathcal{H}_K$.

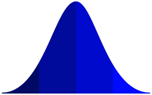


Kernel, RKHS, mean embedding

- Kernel: $K(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}, \quad (\forall x, y \in \mathcal{X})$.
  - Examples:

$$K(x,y) = e^{-\gamma\|x-y\|_2^2} \qquad \leftarrow\text{bounded,}$$
$$K(x,y) = e^{\gamma\langle x,y\rangle} \qquad \leftarrow\text{unbounded.}$$

- RKHS: $\mathcal{H}_K = \overline{\{\sum_{i=1}^n \alpha_i K(\cdot, x_i)\}} \subset \mathbb{R}^{\mathcal{X}}$. $\varphi(x) = \underbrace{K(\cdot, x)} \in \mathcal{H}_K$.



- We represent distributions in RKHSs: $\mu_{\mathbb{P}} := \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_K$.

Data types with kernels: $(\mathcal{X}, K)$

- Trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002], time series [Cuturi, 2011], strings [Lodhi et al., 2002],
- mixture models, hidden Markov models or linear dynamical systems [Jebara et al., 2004],
- sets [Haussler, 1999, Gärtner et al., 2002], fuzzy domains [Guevara et al., 2017], distributions [Hein and Bousquet, 2005, Martins et al., 2009, Muandet et al., 2011],
- groups [Cuturi et al., 2005] $\xrightarrow{\text{spec.}}$ permutations [Jiao and Vert, 2018],
- graphs [Vishwanathan et al., 2010, Kondor and Pan, 2016].

Back to mean embeddings: $\mu_{\mathbb{P}}$

Very natural representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x).$$

Very natural representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x).$$

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(y) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, y)}(x).$$

Very natural representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x).$$

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(y) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, y)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(y) = \int_{\mathbb{R}^d} e^{i\langle y, x\rangle} \mathrm{d}\mathbb{P}(x).$$

Very natural representation

$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x)$.

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(y) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, y)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(y) = \int_{\mathbb{R}^d} e^{i\langle y, x \rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment-generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(y) = \int_{\mathbb{R}^d} e^{\langle y, x \rangle} \mathrm{d}\mathbb{P}(x).$$

Very natural representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x).$$

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(y) = \mathbb{E}_{x\sim\mathbb{P}} \chi_{(-\infty,y)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(y) = \int_{\mathbb{R}^d} e^{i\langle y,x\rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment-generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(y) = \int_{\mathbb{R}^d} e^{\langle y,x\rangle} \mathrm{d}\mathbb{P}(x).$$

Trick

φ: on any kernel-endowed domain!

Mean embedding ($\exists$)

- $\mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x)$ exists $\Leftrightarrow \int_{\mathcal{X}} \underbrace{\|\varphi(x)\|_{\mathcal{H}_K}}_{\sqrt{K(x,x)}} \mathrm{d}\mathbb{P}(x) < \infty.$

Mean embedding ($\exists$), MMD

- $\mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x)$ exists $\Leftrightarrow \int_{\mathcal{X}} \underbrace{\|\varphi(x)\|_{\mathcal{H}_K}}_{\sqrt{K(x,x)}} \mathrm{d}\mathbb{P}(x) < \infty$.
- Maximum mean discrepancy (MMD):

$$\text{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K}$$

Mean embedding (∃), MMD

- $\mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x)$ exists $\Leftrightarrow \int_{\mathcal{X}} \underbrace{\|\varphi(x)\|_{\mathcal{H}_K}}_{\sqrt{K(x,x)}} \mathrm{d}\mathbb{P}(x) < \infty$.
- Maximum mean discrepancy (MMD):

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K} = \sup_{f \in B_K} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_K}.$$

Mean embedding (∃), MMD

- $\mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x)$ exists $\Leftrightarrow \int_{\mathcal{X}} \underbrace{\|\varphi(x)\|_{\mathcal{H}_K}}_{\sqrt{K(x,x)}} \mathrm{d}\mathbb{P}(x) < \infty$.
- Maximum mean discrepancy (M MD):

$$\begin{aligned} \mathrm{MMD}(\mathbb{P},\mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K} \\ &= \sup_{f \in B_K} \underbrace{\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_K}}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}. \end{aligned}$$

Until now

We have defined $\mu_{\mathbb{P}}$ and $\mathrm{MMD}(\mathbb{P},\mathbb{Q})$.

- Applications:
  - two-sample testing [Borgwardt et al., 2006, Gretton et al., 2012],
  - domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017],
  - kernel Bayesian inference [Song et al., 2011, Fukumizu et al., 2013]
  - approximate Bayesian computation [Park et al., 2016], probabilistic programming [Schölkopf et al., 2015],
  - model criticism [Lloyd et al., 2014, Kim et al., 2016], goodness-of-fit [Balasubramanian et al., 2017],
  - distribution classification [Muandet et al., 2011, Lopez-Paz et al., 2015], [Zaheer et al., 2017], distribution regression [Szabó et al., 2016], [Law et al., 2018],
  - topological data analysis [Kusano et al., 2016].
- Review [Muandet et al., 2017].

Easy to estimate them!

Given: $(x_n)_{n\in[N]} \sim \mathbb{P}$ samples. $\hat{\mu}_{\mathbb{P}} = \frac{1}{N}\sum_{n\in[N]} K(\cdot, x_n)$.

Easy to estimate them!

Given: $(x_n)_{n\in[N]} \sim \mathbb{P}$, $(y_n)_{n\in[N]} \sim \mathbb{Q}$ samples. $\hat{\mu}_{\mathbb{P}} = \frac{1}{N}\sum_{n\in[N]} K(\cdot, x_n)$.

- $\widehat{\text{MMD}}(\mathbb{P},\mathbb{Q}) = \|\hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}}\|_{\mathcal{H}_K}$

$$= \sqrt{\frac{1}{N^2}\sum_{i,j\in[N]}\left[K(x_i,x_j) + K(y_i,y_j) - 2K(x_i,y_j)\right]} \quad \text{(V-stat)},$$

Easy to estimate them!

Given: $(x_n)_{n\in[N]} \sim \mathbb{P}$, $(y_n)_{n\in[N]} \sim \mathbb{Q}$ samples. $\hat{\mu}_{\mathbb{P}} = \frac{1}{N}\sum_{n\in[N]} K(\cdot, x_n)$.

- $\widehat{\text{MMD}}(\mathbb{P},\mathbb{Q}) = \|\hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}}\|_{\mathcal{H}_K}$

$$= \sqrt{\frac{1}{N^2}\sum_{i,j\in[N]}\left[K(x_i,x_j) + K(y_i,y_j) - 2K(x_i,y_j)\right]} \quad \text{(V-stat)},$$

$$\overset{or}{=} \frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\neq j}}\left[K(x_i,x_j) + K(y_i,y_j)\right] - \frac{2}{N^2}\sum_{i,j\in[N]} K(x_i,y_j).$$

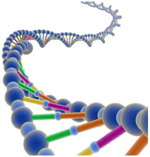
Goal of our work

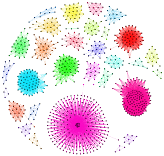
Designing outlier-robust mean embedding and MMD estimators.

Goal of our work

Designing outlier-robust mean embedding and MMD estimators.

- Interest: unbounded kernels.
  - exponential kernel: $K(x,y) = e^{\gamma\langle x,y\rangle}$.
  - polynomial kernel: $K(x,y) = (\langle x,y\rangle + \gamma)^p$.
  - string, time series or graph kernels.

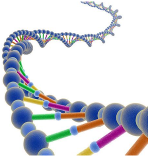


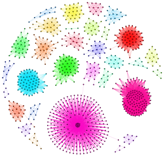


Goal of our work

Designing **outlier-robust** mean embedding and MMD estimators.

- Interest: unbounded kernels.
  - exponential kernel: $K(x,y) = e^{\gamma\langle x,y\rangle}$.
  - polynomial kernel: $K(x,y) = (\langle x,y\rangle + \gamma)^p$.
  - string, time series or graph kernels.





Issue with average

A single outlier can ruin it.

Existing work

- Robust KDE [Kim and Scott, 2012]:

$$\mu_{\mathbb{P}} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot, x)\|_{\mathcal{H}_K}^{2} \, \mathrm{d}\mathbb{P}(x)$$

Existing work

- Robust KDE [Kim and Scott, 2012]:

$$\mu_{\mathbb{P}} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot, x)\|_{\mathcal{H}_K}^{2} \, \mathrm{d}\mathbb{P}(x),$$

$$\mu_{\mathbb{P},L} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} L\left(\|f - K(\cdot, x)\|_{\mathcal{H}_K}\right) \mathrm{d}\mathbb{P}(x)$$

Existing work

- Robust KDE [Kim and Scott, 2012]:

$$\mu_{\mathbb{P}} = \arg\min_{f\in\mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot,x)\|^2_{\mathcal{H}_K} \, \mathrm{d}\mathbb{P}(x),$$

$$\mu_{\mathbb{P},L} = \arg\min_{f\in\mathcal{H}_K} \int_{\mathcal{X}} L\left(\|f - K(\cdot,x)\|_{\mathcal{H}_K}\right) \mathrm{d}\mathbb{P}(x),$$

$$\hat{\mu}_{\mathbb{P},N,L} = \arg\min_{f\in\mathcal{H}_K} \frac{1}{N} \sum_{n\in[N]} L\left(\|f - K(\cdot,x_n)\|_{\mathcal{H}_K}\right)$$

Existing work

- Robust KDE [Kim and Scott, 2012]:

$$\mu_{\mathbb{P}} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot, x)\|^2_{\mathcal{H}_K} \, \mathrm{d}\mathbb{P}(x),$$

$$\mu_{\mathbb{P},L} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} L\left(\|f - K(\cdot, x)\|_{\mathcal{H}_K}\right) \mathrm{d}\mathbb{P}(x),$$

$$\hat{\mu}_{\mathbb{P},N,L} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{n \in [N]} L\left(\|f - K(\cdot, x_n)\|_{\mathcal{H}_K}\right),$$

$$\hat{\mu}_{\mathbb{P},N,L,t} : \text{ iterative approximation of } \hat{\mu}_{\mathbb{P},N,L}, \xrightarrow{t\to\infty} \hat{\mu}_{\mathbb{P},N,L}.$$

Existing work

- Robust KDE [Kim and Scott, 2012]:

$$\mu_{\mathbb{P}} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot, x)\|^2_{\mathcal{H}_K} \, \mathrm{d}\mathbb{P}(x),$$

$$\mu_{\mathbb{P},L} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} L\left(\|f - K(\cdot, x)\|_{\mathcal{H}_K}\right) \mathrm{d}\mathbb{P}(x),$$

$$\hat{\mu}_{\mathbb{P},N,L} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{n \in [N]} L\left(\|f - K(\cdot, x_n)\|_{\mathcal{H}_K}\right),$$

$$\hat{\mu}_{\mathbb{P},N,L,t} : \text{ iterative approximation of } \hat{\mu}_{\mathbb{P},N,L}, \xrightarrow{t \to \infty} \hat{\mu}_{\mathbb{P},N,L}.$$

- Adaptation to KCCA [Alam et al., 2018], relaxation to Hilbert spaces [Sinova et al., 2018].

Existing work

- Robust KDE [Kim and Scott, 2012]:

$$\mu_{\mathbb{P}} = \arg\min_{f\in\mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot,x)\|^2_{\mathcal{H}_K}\, \mathrm{d}\mathbb{P}(x),$$

$$\mu_{\mathbb{P},L} = \arg\min_{f\in\mathcal{H}_K} \int_{\mathcal{X}} L\left(\|f - K(\cdot,x)\|_{\mathcal{H}_K}\right) \mathrm{d}\mathbb{P}(x),$$

$$\hat{\mu}_{\mathbb{P},N,L} = \arg\min_{f\in\mathcal{H}_K} \frac{1}{N} \sum_{n\in[N]} L\left(\|f - K(\cdot,x_n)\|_{\mathcal{H}_K}\right),$$

$$\hat{\mu}_{\mathbb{P},N,L,t}: \text{ iterative approximation of } \hat{\mu}_{\mathbb{P},N,L}, \xrightarrow{t\to\infty} \hat{\mu}_{\mathbb{P},N,L}.$$

- Adaptation to KCCA [Alam et al., 2018], relaxation to Hilbert spaces [Sinova et al., 2018].
- Consistency: For finiteD features [Sinova et al., 2018]

$$\hat{\mu}_{\mathbb{P},N,L} \xrightarrow{N\to\infty} \mu_{\mathbb{P},L}. \quad \text{(empirical M-estimator in } \mathbb{R}^d\text{)}$$

Our building block: Median-Of-meaNs in 1d, $(x_n)_{n\in[N]}$

Goal

Estimate mean while being resistant to contemination.

Our building block: Median-Of-meaNs in 1d, $(x_n)_{n\in[N]}$

Goal

Estimate mean while being resistant to contemination.

MON:

1. Partition: $\underbrace{x_1,\ldots,x_{N/Q}}_{S_1}, \quad \ldots \quad ,\underbrace{x_{N-N/Q+1},\ldots,x_N}_{S_Q}.$

Our building block: Median-Of-meaNs in 1d, $(x_n)_{n\in[N]}$

Goal

Estimate mean while being resistant to contemination.

MON:

1. Partition: $\underbrace{x_1,\ldots,x_{N/Q}}_{S_1}, \quad \ldots \quad ,\underbrace{x_{N-N/Q+1},\ldots,x_N}_{S_Q}.$
2. Compute average in each block:

$$a_1 = \frac{1}{|S_1|}\sum_{i\in S_1} x_i, \quad \ldots \quad , a_Q = \frac{1}{|S_Q|}\sum_{i\in S_Q} x_i.$$

Our building block: Median-Of-meaNs in 1d, $(x_n)_{n\in[N]}$

Goal

Estimate mean while being resistant to contemination.

MON:

1. Partition: $\underbrace{x_1, \ldots, x_{N/Q}}_{S_1}, \quad \ldots \quad , \underbrace{x_{N-N/Q+1}, \ldots, x_N}_{S_Q}$.
2. Compute average in each block:
$$a_1 = \frac{1}{|S_1|}\sum_{i\in S_1} x_i, \quad \ldots \quad , a_Q = \frac{1}{|S_Q|}\sum_{i\in S_Q} x_i.$$
3. Estimate $\mathbb{E}X$: $\text{med}_{q\in[Q]} a_q$.

Idea on MMD (mean embedding: similarly)

1. Use the IPM representation:

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f \in B_K} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_K}.$$

Idea on MMD (mean embedding: similarly)

1. Use the IPM representation:

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\rangle_{\mathcal{H}_K}.$$

2. Replace the expectation with MON:

$$\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]} \left\{ \frac{1}{|S_q|}\sum_{j\in S_q} f(x_j) - \frac{1}{|S_q|}\sum_{j\in S_q} f(y_j) \right\}.$$

Idea on MMD (mean embedding: similarly)

1. Use the IPM representation:

$$\text{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\rangle_{\mathcal{H}_K}.$$

2. Replace the expectation with MON:

$$\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]} \left\{ \frac{1}{|S_q|}\sum_{j\in S_q} f(x_j) - \frac{1}{|S_q|}\sum_{j\in S_q} f(y_j) \right\}.$$

For $Q = 1$, we get back the V-stat MMD estimator.

Idea on MMD (mean embedding: similarly)

1. Use the IPM representation:

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f \in B_K} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}_K}.$$

2. Replace the expectation with MON:

$$\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \sup_{f \in B_K} \operatorname*{med}_{q \in [Q]} \left\{ \frac{1}{|S_q|} \sum_{j \in S_q} f(x_j) - \frac{1}{|S_q|} \sum_{j \in S_q} f(y_j) \right\}.$$

For $Q = 1$, we get back the V-stat MMD estimator.

What can we show about this MONK estimator?

Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P}, \mathbb{Q})$

Assumptions:

1. The # of samples contaminated can be (almost) half of the # of blocks:

$$\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}, \quad N_c \leqslant Q(1/2 - \delta), \quad \delta \in (0, 1/2].$$

Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P}, \mathbb{Q})$

Assumptions:

1. The # of samples contaminated can be (almost) half of the # of blocks:

$$\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}, \quad N_c \leqslant Q(1/2 - \delta), \quad \delta \in (0, 1/2].$$

Clean data: $N_c = 0$, $\delta = \frac{1}{2}$.

Assumptions:

1. The # of samples contaminated can be (almost) half of the # of blocks:

$$\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}, \quad N_c \leqslant Q(1/2 - \delta), \quad \delta \in (0, 1/2].$$

   Clean data: $N_c = 0$, $\delta = \frac{1}{2}$.

2. Assume: $\text{Tr}(\Sigma_{\mathbb{P}})$, $\text{Tr}(\Sigma_{\mathbb{Q}})$ make sense, i.e.

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}}\left[(K(\cdot, x) - \mu_{\mathbb{P}}) \otimes (K(\cdot, x) - \mu_{\mathbb{P}})\right], \Sigma_{\mathbb{Q}} \in \mathcal{L}_1(\mathcal{H}_K).$$

Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Assumptions:

1. The # of samples contaminated can be (almost) half of the # of blocks:

$$\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}, \quad N_c \leqslant Q(1/2-\delta), \quad \delta \in (0, 1/2].$$

   Clean data: $N_c = 0$, $\delta = \frac{1}{2}$.

2. Assume: $\mathrm{Tr}(\Sigma_{\mathbb{P}})$, $\mathrm{Tr}(\Sigma_{\mathbb{Q}})$ make sense, i.e.

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\left[(K(\cdot,x)-\mu_{\mathbb{P}})\otimes(K(\cdot,x)-\mu_{\mathbb{P}})\right], \Sigma_{\mathbb{Q}} \in \mathcal{L}_1(\mathcal{H}_K).$$

   Minimal 2nd-order condition.

Consistency & outlier-robustness of $\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Assumptions:

1. The # of samples contaminated can be (almost) half of the # of blocks:

$$\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}, \quad N_c \leqslant Q(1/2-\delta), \quad \delta \in (0, 1/2].$$

   Clean data: $N_c = 0$, $\delta = \frac{1}{2}$.

2. Assume: $\text{Tr}(\Sigma_{\mathbb{P}})$, $\text{Tr}(\Sigma_{\mathbb{Q}})$ make sense, i.e.

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\left[(K(\cdot,x)-\mu_{\mathbb{P}})\otimes(K(\cdot,x)-\mu_{\mathbb{P}})\right], \Sigma_{\mathbb{Q}} \in \mathcal{L}_1(\mathcal{H}_K).$$

   Minimal 2nd-order condition. Note: $\|A\| \leqslant \|A\|_{HS} \overset{(*)}{\leqslant} \|A\|_1$.

Consistency & outlier-robustness of $\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\boxed{\left|\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \text{MMD}(\mathbb{P},\mathbb{Q})\right|} \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\text{Tr}(\Sigma_{\mathbb{P}})+\text{Tr}(\Sigma_{\mathbb{Q}})}{N}}\right)}{\delta}.$$

Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}(\Sigma_{\mathbb{P}})+\mathrm{Tr}(\Sigma_{\mathbb{Q}})}{N}}\right)}{\delta}.$$

Discussion:

1. N-dependence: $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ is optimal for MMD estimation [Tolstikhin et al., 2016].

Consistency & outlier-robustness of $\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \text{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\text{Tr}(\Sigma_{\mathbb{P}})+\text{Tr}(\Sigma_{\mathbb{Q}})}{N}}\right)}{\delta}.$$

Discussion:

2. Σ-dependence:
   - Optimal sub-Gaussian deviation bound for mean estimation under minimal 2nd-order condition even on $\mathbb{R}^d$ [Lugosi and Mendelson, 2019] – long-lasting open question.

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \text{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\text{Tr}(\Sigma_{\mathbb{P}})+\text{Tr}(\Sigma_{\mathbb{Q}})}{N}}\right)}{\delta}.$$

Discussion:

2. Σ-dependence:
   - Optimal sub-Gaussian deviation bound for mean estimation under minimal 2nd-order condition even on $\mathbb{R}^d$ [Lugosi and Mendelson, 2019] – long-lasting open question.
   - They rely on tournament procedure: numerically hard.

Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}(\Sigma_{\mathbb{P}})+\mathrm{Tr}(\Sigma_{\mathbb{Q}})}{N}}\right)}{\delta}.$$

Discussion:

2. Σ-dependence:
   - Optimal sub-Gaussian deviation bound for mean estimation under minimal 2nd-order condition even on $\mathbb{R}^d$ [Lugosi and Mendelson, 2019] – long-lasting open question.
   - They rely on tournament procedure: numerically hard.
   - Most practical convex relaxation [Hopkins, 2018]: $\mathcal{O}\left(N^{24}\right)$.

Consistency & outlier-robustness of $\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \text{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\text{Tr}\left(\Sigma_{\mathbb{P}}\right)+\text{Tr}\left(\Sigma_{\mathbb{Q}}\right)}{N}}\right)}{\delta}.$$

Discussion:

3. δ-dependence:
   - Larger δ means less outliers,
     - the bound becomes tighter,
     - one needs less blocks.

Consistency & outlier-robustness of $\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \text{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\text{Tr}(\Sigma_{\mathbb{P}})+\text{Tr}(\Sigma_{\mathbb{Q}})}{N}}\right)}{\delta}.$$

Discussion:

3. δ-dependence:
   - Larger δ means less outliers,
     - the bound becomes tighter,
     - one needs less blocks.
   - optimal?

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}\left(\Sigma_{\mathbb{P}}\right)+\mathrm{Tr}\left(\Sigma_{\mathbb{Q}}\right)}{N}}\right)}{\delta}.$$

Discussion:

4. breakdown point – asymptotic concept:
   - median ⇒ Using Q blocks is resistant to $Q/2$ outliers.
   - Q can grow with N, as (almost) $N/2$.
   - Breakdown point can be 25%.

Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}\left(\Sigma_{\mathbb{P}}\right)+\mathrm{Tr}\left(\Sigma_{\mathbb{Q}}\right)}{N}}\right)}{\delta}.$$

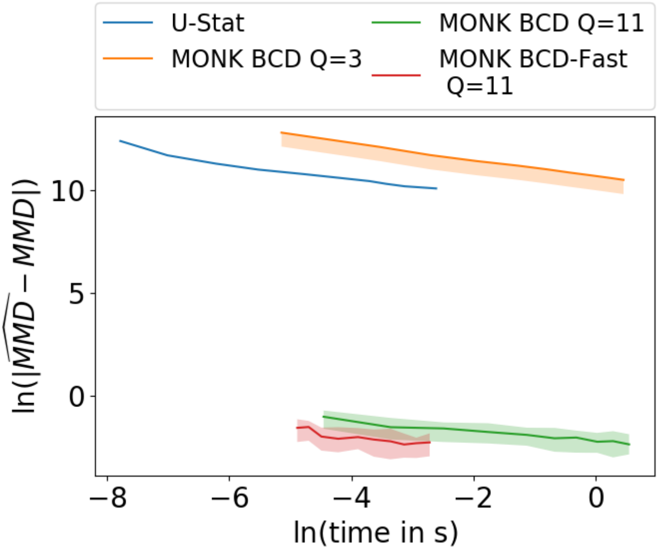
Discussion:

5. Unknown Q:
   - One choose Q adaptively by the Lepski method.
   - Same guarantee but with increased computional cost.

Numerical demo: quadratic kernel, $N_c = 5$ outliers

1. No outliers / bounded kernel: MONK is a safe alternative.
2. Relevant case: outliers & unbounded kernel.
   - $\mathbb{P} := \mathcal{N}\left(\mu_1, \sigma_1^2\right) \neq \mathbb{Q} := \mathcal{N}\left(\mu_2, \sigma_2^2\right)$. $\mu_m, \sigma_m \sim U[0,1]$, fixed.
   - $N \in \{200, 400, \ldots, 2000\}$.
   - 5-5 corrupted samples: $(x)_{n=N-4}^{N} = 2000$, $(y_n)_{n=N-4}^{N} = 4000$.
   - $(\mathbb{P}, \mathbb{Q}, K)$: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q})$ is analytic.
   - Performance:
     - 100 MC simulations,
     - median and quartiles.

Numerical demo: quadratic kernel, $N_c = 5$ outliers

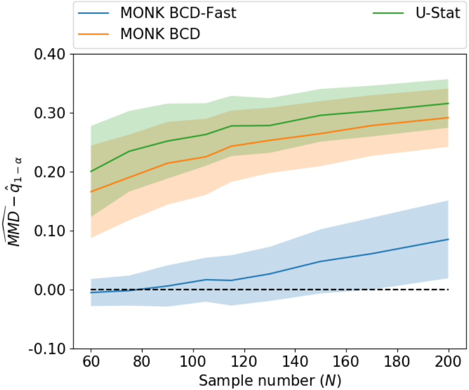


DNA analysis: 2-sample testing

- Discrimination of 2 DNA categories (EI, IE).
- Subsequent String Kernel (K).
- Significance level: $\alpha = 0.05$.
- Performance:
  - 4000 MC simulations,
  - mean $\pm$ std of $\widehat{\text{MMD}} - \hat{q}_{1-\alpha}$.
- $\hat{q}_{1-\alpha}$: Using 150 bootstrap permutations.

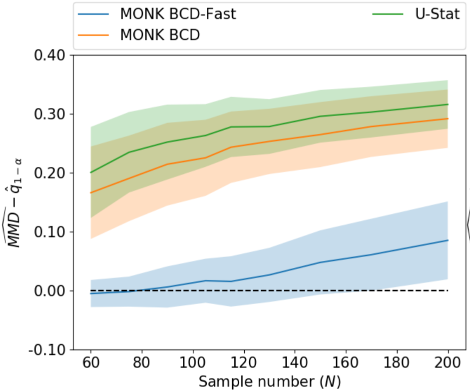
DNA analysis: plots

Inter-class: EI-IE

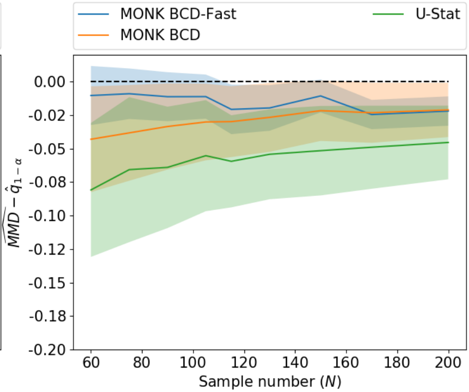


DNA analysis: plots

Inter-class: EI-IE,



Intra-class: EI-EI (IE-IE)



Summary

- Focus: Outlier-robust mean embedding & MMD estimation.
- Technique: median-of-means.
- Finite-sample guarantees (optimality), excessive resistance to contamination.

Summary

- Focus: Outlier-robust mean embedding & MMD estimation.
- Technique: median-of-means.
- Finite-sample guarantees (optimality), excessive resistance to contamination.
- Preprint, code:

  MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means, TR (http://arxiv.org/abs/1802.04784).

  https://bitbucket.org/TimotheeMathieu/monk-mmd

Thank you for the attention!

Computational complexity of MMD estimators

N: sample number, Q: number of blocks, T: number of iterations.

Method	Complexity
U-Stat	$\mathcal{O}\left(N^2\right)$
MONK BCD	$\mathcal{O}\left(N^3 + T\left[N^2 + Q\log(Q)\right]\right)$
MONK BCD-Fast	$\mathcal{O}\left(\frac{N^3}{Q^2} + T\left[\frac{N^2}{Q} + Q\log(Q)\right]\right)$

Pseudo-code: 2-sample testing

Input: Two samples: $(X_n)_{n\in[N]}$, $(Y_n)_{n\in[N]}$. Number of bootstrap permutations: $B \in \mathbb{Z}^+$. Level of the test: $\alpha \in (0,1)$. Kernel function with hyperparameter $\theta \in \Theta$: K_θ.

Split the dataset randomly into 3 equal parts:

$$[N] = \dot{\bigcup_{i=1}^{3}} I_i, \quad |I_1| = |I_2| = |I_3|.$$

Tune the hyperparameters using the 1st part of the dataset:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} J_\theta := \widehat{\text{MMD}}_\theta\left((X_n)_{n\in I_1}, (Y_n)_{n\in I_1}\right).$$

Estimate the $(1-\alpha)$-quantile of $\widehat{\text{MMD}}_{\hat{\theta}}$ under the null, using B bootstrap permutations from $(X_n)_{n\in I_2} \cup (Y_n)_{n\in I_2}$: $\hat{q}_{1-\alpha}$.

Compute the test statistic on the third part of the dataset:

$$T_{\hat{\theta}} = \widehat{\text{MMD}}_{\hat{\theta}}\left((X_n)_{n\in I_3}, (Y_n)_{n\in I_3}\right).$$

Output: $T_{\hat{\theta}} - \hat{q}_{1-\alpha}$.

Alam, M. A., Fukumizu, K., and Wang, Y.-P. (2018).
Influence function and robust variant of kernel canonical correlation analysis.
Neurocomputing, 304:12–29.

Balasubramanian, K., Li, T., and Yuan, M. (2017).
On the optimality of kernel-embedding based goodness-of-fit tests.
Technical report.
(https://arxiv.org/abs/1709.08148).

Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2017).
Domain generalization by marginal transfer learning.
Technical report.
(https://arxiv.org/abs/1711.07910).

Borgwardt, K., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006).

Integrating structured biological data by kernel maximum mean discrepancy.
Bioinformatics, 22:e49–57.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Neural Information Processing Systems (NIPS)*, pages 625–632.

Cuturi, M. (2011).
Fast global alignment kernels.
In *International Conference on Machine Learning (ICML)*, pages 929–936.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:1169–1198.

Fukumizu, K., Song, L., and Gretton, A. (2013).
Kernel Bayes' rule: Bayesian inference with positive definite kernels.

Journal of Machine Learning Research, 14:3753–3783.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research, 13:723–773.

Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.
In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Haussler, D. (1999).
Convolution kernels on discrete structures.

Technical report, Department of Computer Science, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hein, M. and Bousquet, O. (2005).
Hilbertian metrics and positive definite kernels on probability measures.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.

Hopkins, S. B. (2018).
Mean estimation with sub-gaussian rates in polynomial time.
Technical report.
(https://arxiv.org/abs/1809.07425).

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Jiao, Y. and Vert, J.-P. (2018).

The Kendall and Mallows kernels for permutations.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(7):1755–1769.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*, pages 291–298.

Kim, B., Khanna, R., and Koyejo, O. O. (2016).
Examples are not enough, learn to criticize! criticism for interpretability.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2280–2288.

Kim, J. and Scott, C. D. (2012).
Robust kernel density estimation.
Journal of Machine Learning Research, 13:2529–2565.

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.

In *Neural Information Processing Systems (NIPS)*, pages 2982–2990.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016).
Persistence weighted Gaussian kernel for topological data analysis.
In *International Conference on Machine Learning (ICML)*, pages 2004–2013.

Law, H. C. L., Sutherland, D. J., Sejdinovic, D., and Flaxman, S. (2018).
Bayesian approaches to distribution regression.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2014).
Automatic construction and natural-language description of nonparametric regression models.

In *AAAI Conference on Artificial Intelligence*, pages 1242–1250.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).
Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015).
Towards a learning theory of cause-effect inference.
International Conference on Machine Learning (ICML; PMLR), 37:1452–1461.

Lugosi, G. and Mendelson, S. (2019).
Sub-gaussian estimators of the mean of a random vector.
Annals of Statistics, 47(2):783–794.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2009).
Nonextensive information theoretic kernels on measures.

The Journal of Machine Learning Research, 10:935–975.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2011).
Learning from distributions via support measure machines.
In *Neural Information Processing Systems (NIPS)*, pages 10–18.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017).
Kernel mean embedding of distributions: A review and beyond.
Foundations and Trends in Machine Learning, 10(1-2):1–141.

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016).
K2-ABC: Approximate Bayesian computation with kernel embeddings.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 51, pages 51:398–407.

Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S., and Peters, J. (2015).
Computing functions of random variables via reproducing kernel Hilbert space representations.
Statistics and Computing, 25(4):755–766.

Sinova, B., González-Rodríguez, G., and Aelst, S. V. (2018).
M-estimators of location for functional data.
Bernoulli, 24:2328–2357.

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. (2011).
Kernel belief propagation.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715.

Szabó, Z., Sriperumbudur, B., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.
Journal of Machine Learning Research, 17(152):1–40.

Tolstikhin, I., Sriperumbudur, B. K., and Schölkopf, B. (2016).
Minimax estimation of maximal mean discrepancy with radial kernels.
In *NIPS*, pages 1930–1938.

Vishwanathan, S. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010).
Graph kernels.
Journal of Machine Learning Research, 11:1201–1242.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017).
Deep sets.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 3394–3404.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013).
Domain adaptation under target and conditional shift.
Journal of Machine Learning Research, 28(3):819–827.