The MONK

Zoltán Szabó – CMAP, École Polytechnique

Joint work with:

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

MathoDS 3
Hong Kong, China
June 21, 2019

Motivation: Tibet, monks

Mean embedding, MMD

- Mean embedding:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{\text{example: } \mathbb{I}_{(-\infty,\cdot)}(x),\, e^{i\langle \cdot, x\rangle},\, e^{\langle \cdot, x\rangle} \text{ in } \mathbb{R}^d} \mathrm{d}\mathbb{P}(x).$$

- Mean embedding:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{\text{example: } \mathbb{I}_{(-\infty,\cdot)}(x),\, e^{i\langle \cdot, x\rangle},\, e^{\langle \cdot, x\rangle} \text{ in } \mathbb{R}^d} \mathrm{d}\mathbb{P}(x).$$

- Maximum mean discrepancy (M MD)[†]:

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\| = \sup_{f \in B} \underbrace{\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\rangle}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}.$$

[†] Nicknames: energy distance, N-distance.

$\mu_{\mathbb{P}}$, MMD: review [Muandet et al., 2017]

Applications:

- two-sample testing [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Borgwardt et al., 2006, Harchaoui et al., 2007, Gretton et al., 2012, Jitkrittum et al., 2016], and its differential private variant [Raj et al., 2019]; independence [Gretton et al., 2008, Pfister et al., 2017, Jitkrittum et al., 2017a] and goodness-of-fit testing [Jitkrittum et al., 2017b, Balasubramanian et al., 2017], causal discovery [Mooij et al., 2016, Pfister et al., 2017],
- domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017], change-point detection [Harchaoui and Cappé, 2007], post selection inference [Yamada et al., 2018],
- 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],
- topological data analysis [Kusano et al., 2016],
- 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],
- generative adversarial networks [Dziugaite et al., 2015, Li et al., 2015, Binkowski et al., 2018], understanding the dynamics of complex dynamical systems [Klus et al., 2018, Klus et al., 2019], . . .

φ domain: few examples

- 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].

φ domain: few examples

- 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].

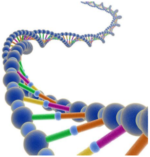
Key: kernels

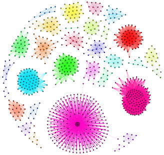
- $K(x,y) = \langle \varphi(x), \varphi(y) \rangle$, $\varphi(x) = K(\cdot, x)$,
- $\mathcal{H}_K = \overline{\text{span}}\{\varphi(x) : x \in \mathcal{X}\} \ni \mu_{\mathbb{P}}$.

Goal of our work

Designing outlier-robust mean embedding and MMD estimators.

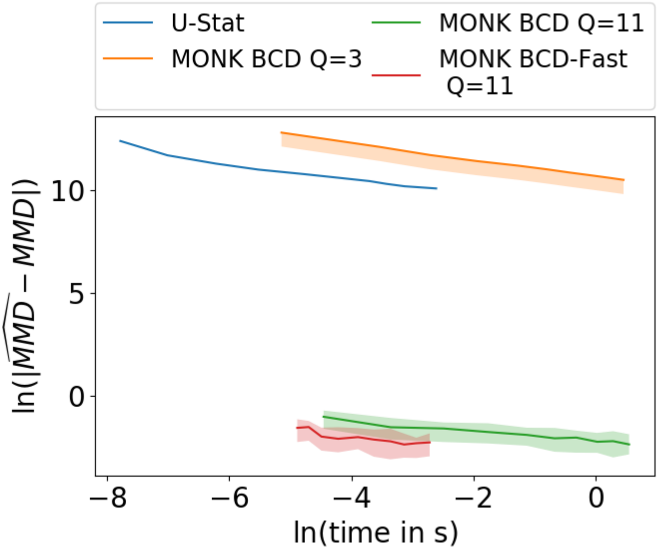
- Interest: unbounded kernels.
    - exponential kernel: $K(x,y) = e^{\gamma \langle x,y \rangle}$.
    - polynomial kernel: $K(x,y) = (\langle x,y \rangle + \gamma)^p$.
    - string, time series or graph kernels.





Issue with average

A single outlier can ruin it.


U-Stat
MONK BCD Q=3
MONK BCD Q=11
MONK BCD-Fast Q=11
$\ln(|\widehat{MMD} - MMD|)$
10
5
0
−8
−6
−4
−2
0
ln(time in s)

Existing work @ kernel land

- Robust KDE [Kim and Scott, 2012]:

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

$$\mu_{\mathbb{P},L} = \arg\min_f \int_{\mathcal{X}} L\left(\|f - K(\cdot, x)\|\right) \mathrm{d}\mathbb{P}(x).$$

Existing work @ kernel land

- Robust KDE [Kim and Scott, 2012]:

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

$$\mu_{\mathbb{P},L} = \arg\min_f \int_{\mathcal{X}} L\left(\|f - K(\cdot, x)\|\right) \mathrm{d}\mathbb{P}(x).$$

  Consistency ($\hat{\mu}_{\mathbb{P},L} \xrightarrow{?} \mu_{\mathbb{P}}$):
  - As a density estimator [Vandermeulen and Scott, 2013] (L-independent).
  - For finiteD features [Sinova et al., 2018] – M-estimation in $\mathbb{R}^d$.
- Adaptation to KCCA [Alam et al., 2018].

@Statistics: Hanson-Wright inequality (mean estimation)

- Gaussian:
  - Let $\{\mathbf{x}_n\}_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$, $\bar{\mathbf{x}}_N = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n$.
  - For any $\eta \in (0,1)$ with probability $1-\eta$ [Hanson and Wright, 1971]

$$\|\bar{\mathbf{x}}_N - \mathbf{m}\|_2 \leqslant \sqrt{\frac{\text{Tr}(\mathbf{\Sigma})}{N}} + \sqrt{\frac{2\lambda_{\max}(\mathbf{\Sigma})\ln(1/\eta)}{N}}. \tag{1}$$

- Gaussian:
  - Let $\{\mathbf{x}_n\}_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$, $\bar{\mathbf{x}}_N = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n$.
  - For any $\eta \in (0,1)$ with probability $1-\eta$ [Hanson and Wright, 1971]

$$\|\bar{\mathbf{x}}_N - \mathbf{m}\|_2 \leqslant \sqrt{\frac{\mathrm{Tr}(\mathbf{\Sigma})}{N}} + \sqrt{\frac{2\lambda_{\max}(\mathbf{\Sigma})\ln(1/\eta)}{N}}. \tag{1}$$

- Similar bound can be proved for sub-Gaussian variables.

- Gaussian:
  - Let $\{\mathbf{x}_n\}_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$, $\bar{\mathbf{x}}_N = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n$.
  - For any $\eta \in (0,1)$ with probability $1-\eta$ [Hanson and Wright, 1971]

$$\|\bar{\mathbf{x}}_N - \mathbf{m}\|_2 \leqslant \sqrt{\frac{\operatorname{Tr}(\mathbf{\Sigma})}{N}} + \sqrt{\frac{2\lambda_{\max}(\mathbf{\Sigma})\ln(1/\eta)}{N}}. \tag{1}$$

- Similar bound can be proved for sub-Gaussian variables.
- Heavy-tailed case:
  - No hope for similar behaviour with the sample mean.
  - Other estimators achieving (1), up to constant?
  - Under minimal assumptions ($\exists \mathbf{\Sigma}$).

Long-lasting open problem. $\Rightarrow$ Performance baseline.

Idea: 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 $\mathbb{E}X$: $\operatorname{med}_{q\in[Q]} a_q$.

On MMD (mean embedding: similarly)

- Recall:

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f\in B} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle.$$

- Replace the expectation with MON:

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

Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous; $\mathcal{X}$: separable.
2. Excessive outlier robustness (δ, median):

$$\text{Contaminated } \# \text{ of samples} < \frac{\# \text{ of blocks}}{2}.$$

Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous; $\mathcal{X}$: separable.
2. Excessive outlier robustness (δ, median):

   $$\text{Contaminated \# of samples} < \frac{\text{\# of blocks}}{2}.$$

   Formally:

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

Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous; $\mathcal{X}$: separable.
2. Excessive outlier robustness (δ, median):

$$\text{Contaminated \# of samples} < \frac{\text{\# of blocks}}{2}.$$

Formally:

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

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

Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous; $\mathcal{X}$: separable.
2. Excessive outlier robustness (δ, median):
$$\text{Contaminated } \# \text{ of samples} < \frac{\# \text{ of blocks}}{2}.$$
3. Minimal 2nd-order condition:
$$\exists \, \text{Tr}(\Sigma_{\mathbb{P}}), \text{Tr}(\Sigma_{\mathbb{Q}}),$$

Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous; $\mathcal{X}$: separable.
2. Excessive outlier robustness (δ, median):

$$\text{Contaminated \# of samples} < \frac{\text{\# of blocks}}{2}.$$

3. Minimal 2nd-order condition:

$$\exists\, \text{Tr}(\Sigma_{\mathbb{P}}), \text{Tr}(\Sigma_{\mathbb{Q}}),$$
$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\left[(K(\cdot,x) - \mu_{\mathbb{P}}) \otimes (K(\cdot,x) - \mu_{\mathbb{P}})\right].$$

Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous; $\mathcal{X}$: separable.
2. Excessive outlier robustness (δ, median):
$$\text{Contaminated } \# \text{ of samples} < \frac{\# \text{ of blocks}}{2}.$$
3. Minimal 2nd-order condition:
$$\exists\, \mathrm{Tr}(\Sigma_{\mathbb{P}}), \mathrm{Tr}(\Sigma_{\mathbb{Q}}),$$
$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\left[(K(\cdot,x)-\mu_{\mathbb{P}}) \otimes (K(\cdot,x)-\mu_{\mathbb{P}})\right].$$

Note: $\|A\| \leqslant \|A\|_{HS} \overset{(*)}{\leqslant} \|A\|_1$.

Finite-sample guarantee

For $\forall \eta \in (0,1)$ such that $Q = Q(\delta, \eta) \in \left(N_c / \left(\frac{1}{2} - \delta\right), \frac{N}{2}\right)$ with prob. $\geqslant 1 - \eta$

$$\left| \widehat{\text{MMD}}_Q(\mathbb{P}, \mathbb{Q}) - \text{MMD}(\mathbb{P}, \mathbb{Q}) \right| \leqslant f(N, \Sigma_{\mathbb{P}}, \Sigma_{\mathbb{Q}}, \eta, \delta).$$

Finite-sample guarantee

For $\forall \eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta) \in \left(N_c/\left(\frac{1}{2}-\delta\right), \frac{N}{2}\right)$ with prob. $\geqslant 1-\eta$

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

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

Finite-sample guarantee

For $\forall\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta) \in \left(N_c/\left(\frac{1}{2}-\delta\right), \frac{N}{2}\right)$ with prob. $\geqslant 1-\eta$

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

- $\Sigma_{\mathbb{P}}$, $\Sigma_{\mathbb{Q}}$, η-dependence:
  $\max\left(\sqrt{\text{Tr}(\Sigma_{\mathbb{P}})+\text{Tr}(\Sigma_{\mathbb{Q}})}, \sqrt{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}\right)$.
  - optimal [Lugosi and Mendelson, 2019] ($\mathbb{R}^d$, tournament procedures),
  - most practical convex relaxation [Hopkins, 2018]: $\mathcal{O}\left(N^{24}+Nd\right)$,
  - meanwhile [Cherapanamjeri et al., 2019]: $\mathcal{O}(N^4+dN^2)$, $d<\infty$.

Finite-sample guarantee

For $\forall \eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta) \in \left(N_c/\left(\frac{1}{2}-\delta\right), \frac{N}{2}\right)$ with prob. $\geqslant 1-\eta$

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

- δ-dependence: optimal?

Finite-sample guarantee

For $\forall \eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta) \in \left(N_c/\left(\frac{1}{2}-\delta\right), \frac{N}{2}\right)$ with prob. $\geqslant 1-\eta$

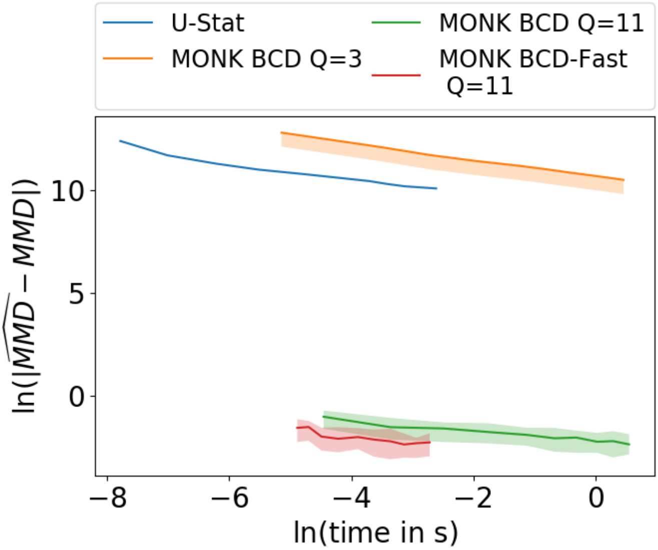
$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leqslant \frac{12\max\left(2\sqrt{\frac{\mathrm{Tr}(\Sigma_{\mathbb{P}})+\mathrm{Tr}(\Sigma_{\mathbb{Q}})}{N}}, \sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}\right)}{\delta}.$$

- Breakdown point can be 25% (asymptotic behavior).

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: $(x_n)_{n=N-4}^{N} = 2000$, $(y_n)_{n=N-4}^{N} = 4000$.
   - $(\mathbb{P}, \mathbb{Q}, K)$: $\text{MMD}(\mathbb{P}, \mathbb{Q})$ is analytic.
   - Performance:
     - 100 MC simulations,
     - median and quartiles.

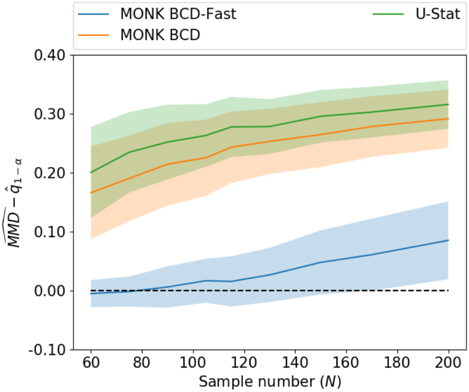
Numerical demo: quadratic kernel, $N_c = 5$ outliers



DNA analysis: 2-sample testing

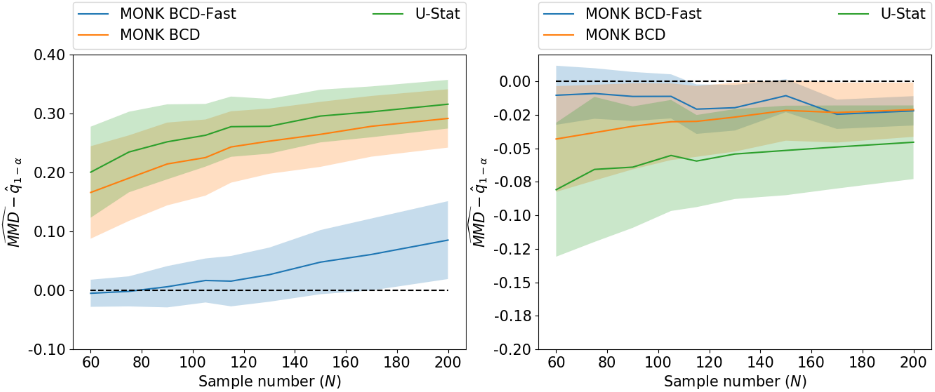
- Discrimination of 2 DNA categories (EI, IE).
- Subsequence String Kernel (K).
- Significance level: $\alpha = 0.05$.
- Performance:
  - 4000 MC simulations,
  - mean $\pm$ std of $\widehat{\text{MMD}} - \hat{q}_{1-\alpha}$.
- $\hat{q}_{1-\alpha}$: Using 150 bootstrap permutations.

Inter-class: EI-IE



Inter-class: EI-IE,

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



Summary

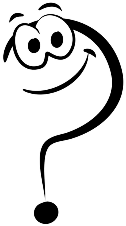
- Goal: outlier-robust mean embedding & MMD estimation.
- MONK estimator: various optimal guarantees (ICML-2019).
- Demo: statistics & gene analysis.
- Code:

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

Summary

- Goal: outlier-robust mean embedding & MMD estimation.
- MONK estimator: various optimal guarantees (ICML-2019).
- Demo: statistics & gene analysis.
- Code:

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



Acks: Guillaume Lecué is supported by a grant of the French National Research Agency (ANR), "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047).

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`).

Baringhaus, L. and Franz, C. (2004).
On a new multivariate two-sample test.
Journal of Multivariate Analysis, 88:190–206.

Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018).
Demystifying MMD GANs.
In *International Conference on Learning Representations (ICLR)*.

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.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019).
Fast mean estimation with sub-Gaussian rates.
Technical report.
(https://arxiv.org/abs/1902.01998).

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.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015).
Training generative neural networks via maximum mean discrepancy optimization.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 258–267.

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.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008).
A kernel statistical test of independence.
In *Neural Information Processing Systems (NIPS)*, pages 585–592.

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.

Hanson, D. and Wright, F. (1971).
A bound on tail probabilities for quadratic forms in independent random variables.
Annals of Mathematical Statistics, 42:1079–1083.

Harchaoui, Z., Bach, F., and Moulines, E. (2007).
Testing for homogeneity with kernel Fisher discriminant analysis.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616.

Harchaoui, Z. and Cappé, O. (2007).
Retrospective mutiple change-point estimation with kernels.
In *IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772.

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.

Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016).
Interpretable distribution features with maximum testing power.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 181–189.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017a).
An adaptive test of independence with analytic kernel embeddings.
In *International Conference on Machine Learning (ICML; PMLR)*, volume 70, pages 1742–1751. PMLR.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017b).
A linear-time kernel goodness-of-fit test.

In *Advances in Neural Information Processing Systems (NIPS)*.

(best paper award = in top 3 out of 3240 submissions).

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.

Klus, S., Bittracher, A., Schuster, I., and Schütte, C. (2019).
A kernel-based approach to molecular conformation analysis.

The Journal of Chemical Physics, 149:244109.

Klus, S., Schuster, I., and Muandet, K. (2018).
Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces.
Technical report.
(https://arxiv.org/abs/1712.01572).

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)*.

Li, Y., Swersky, K., and Zemel, R. (2015).
Generative moment matching networks.
In *International Conference on Machine Learning (ICML; PMLR)*, pages 1718–1727.

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.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016).
Distinguishing cause from effect using observational data: Methods and benchmarks.
Journal of Machine Learning Research, 17:1–102.

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.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2017).

Kernel-based tests for joint independence.
Journal of the Royal Statistical Society: Series B (Statistical Methodology).

Raj, A., Law, H. C. L., Sejdinovic, D., and Park, M. (2019).
A differentially private kernel two-sample test.
In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).*

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.

Székely, G. J. and Rizzo, M. L. (2004).
Testing for equal distributions in high dimension.
InterStat, 5.

Székely, G. J. and Rizzo, M. L. (2005).
A new test for multivariate normality.
Journal of Multivariate Analysis, 93:58–80.

Tolstikhin, I., Sriperumbudur, B. K., and Schölkopf, B. (2016).

Minimax estimation of maximal mean discrepancy with radial kernels.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1930–1938.

Vandermeulen, R. and Scott, C. (2013).
Consistency of robust kernel density estimators.
In *Conference on Learning Theory (COLT; PMLR)*, volume 30, pages 568–591.

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

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2018).
Post selection inference with kernels.

In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 84, pages 152–160.

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.