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:

\[ \mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\chi} \varphi(x) \, d\mathbb{P}(x). \]

Example: \( \mathbb{I}_{(-\infty, \cdot)}(x), e^{i\langle \cdot, x \rangle}, e^{\langle \cdot, x \rangle} \) in \( \mathbb{R}^d \)
Mean embedding, MMD

- Mean embedding:
  \[ \mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_X \varphi(x) \ d\mathbb{P}(x). \]
  example: \( \mathbb{I}_{(-\infty, \cdot)}(x), e^{i \langle \cdot, x \rangle}, e^{\langle \cdot, x \rangle} \) in \( \mathbb{R}^d \)

- Maximum mean discrepancy (MMD)\(^\dagger\):
  \[ \text{MMD}(\mathbb{P}, \mathbb{Q}) = \| \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \| = \sup_{f \in B} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle. \]
  \[ \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{x \sim \mathbb{Q}} f(x) \]

\(^\dagger\) Nicknames: energy distance, N-distance.
 Applications:

- 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], ...
\( \varphi \) 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].
\( \varphi \) 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**

\[
K(x, y) = \langle \varphi(x), \varphi(y) \rangle, \quad \varphi(x) = K(\cdot, x),
\]

\[
\mathcal{H}_K = \text{span} \{ \varphi(x) : x \in \mathcal{X} \} \ni \mu_\mathcal{P}.
\]
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.
Demo: quadratic kernel, 5 outliers

\[ \ln(\mid MMD - \text{MMD}) \]

\[ \ln(\text{time in s}) \]

- U-Stat
- MONK BCD Q=3
- MONK BCD Q=11
- MONK BCD-Fast Q=11
Robust KDE [Kim and Scott, 2012]:

\[
\mu_P = \arg \min_f \int_{\mathcal{X}} \| f - K(\cdot, x) \|^2 \ dP(x),
\]

\[
\mu_{P,L} = \arg \min_f \int_{\mathcal{X}} L(\| f - K(\cdot, x) \|) \ dP(x).
\]
Robust KDE [Kim and Scott, 2012]:

\[
\mu_P = \arg\min_f \int_{\mathcal{X}} \| f - K(\cdot, x) \|^2 \, d\mathbb{P}(x),
\]

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

Consistency (\( \hat{\mu}_{P,L} \rightarrow \mu_P \)):

- 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].
Gaussian:

Let \( \{x_n\}_{n=1}^N \) i.i.d. \( \mathcal{N}(m, \Sigma) \), \( \bar{x}_N = \frac{1}{N} \sum_{n=1}^N x_n \).

For any \( \eta \in (0, 1) \) with probability \( 1 - \eta \) [Hanson and Wright, 1971]

\[
\| \bar{x}_N - m \|_2 \leq \sqrt{\frac{\text{Tr}(\Sigma)}{N}} + \sqrt{\frac{2\lambda_{\text{max}}(\Sigma)\ln(1/\eta)}{N}}. 
\] (1)
Gaussian:

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

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

- Similar bound can be proved for sub-Gaussian variables.
Gaussian:
- Let \( \{x_n\}_{n=1}^N \) i.i.d. \( \mathcal{N}(\mathbf{m}, \Sigma) \), \( \bar{x}_N = \frac{1}{N} \sum_{n=1}^N x_n \).
- For any \( \eta \in (0, 1) \) with probability \( 1 - \eta \) [Hanson and Wright, 1971]
\[
\|\bar{x}_N - \mathbf{m}\|_2 \leq \sqrt{\frac{\operatorname{Tr}(\Sigma)}{N}} + \sqrt{\frac{2\lambda_{\max}(\Sigma) \ln(1/\eta)}{N}}.
\]

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 (\( \exists \Sigma \)).

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 contamination.

**MON:**

1. **Partition:**

\[
\begin{align*}
S_1 &= \{x_1, \ldots, x_{N/Q}\}, \\
S_Q &= \{x_{N-N/Q+1}, \ldots, x_N\}.
\end{align*}
\]

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** \(\mathbb{E}X\): \(\text{med}_{q \in [Q]} a_q\).
Recall:

\[ \text{MMD}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{B}} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle. \]

Replace the expectation with MON:

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

1. $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is continuous; $\mathcal{X}$: separable.

2. **Excessive outlier robustness** ($\delta$, median):
   
   Contaminated # of samples $< \frac{\# \text{ of blocks}}{2}$. 

---

Zoltán Szabó | The MONK
Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous; $\mathcal{X}$: separable.

2. Excessive outlier robustness ($\delta$, median):
   
   Contaminated # of samples $< \frac{\text{# of blocks}}{2}$.

   Formally:
   
   $$\{ (x_{nj}, y_{nj}) \}_{j=1}^{N_c}, \quad N_c \leq Q(1/2 - \delta), \quad \delta \in (0, 1/2].$$
Assumptions

1. $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is continuous; $\mathcal{X}$: separable.

2. **Excessive outlier robustness** ($\delta$, median):
   
   Contaminated # of samples $< \frac{\# \text{ of blocks}}{2}$.

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

   Clean data: $N_c = 0, \delta = \frac{1}{2}$. 
1. \( K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) is continuous; \( \mathcal{X} \): separable.

2. **Excessive outlier robustness** (\( \delta \), median):
   
   Contaminated \( \# \) of samples < \( \frac{\# \text{ of blocks}}{2} \).

3. **Minimal 2nd-order condition**:

   \[ \exists \ Tr(\Sigma_P), Tr(\Sigma_Q), \]
Assumptions

1. \( K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) is continuous; \( \mathcal{X} \): separable.

2. **Excessive outlier robustness** \((\delta, \text{median})\):
   \[
   \text{Contaminated \# of samples} < \frac{\# \text{ of blocks}}{2}.
   \]

3. **Minimal 2nd-order condition**:
   \[
   \exists \text{Tr}(\Sigma_P), \text{Tr}(\Sigma_Q), \quad \Sigma_P = \mathbb{E}_{x \sim P} [ (K(\cdot, x) - \mu_P) \otimes (K(\cdot, x) - \mu_P) ].
   \]
1. \( K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is continuous; \( \mathcal{X} \): separable.

2. **Excessive outlier robustness** (\( \delta \), median):
   
   Contaminated \( \# \) of samples < \( \frac{\# \text{ of blocks}}{2} \).

3. **Minimal 2nd-order condition**:
   
   \[
   \exists \ Tr(\Sigma_P), \ Tr(\Sigma_Q),
   \]
   
   \[
   \Sigma_P = \mathbb{E}_{x \sim P} \left[ (K(\cdot, x) - \mu_P) \otimes (K(\cdot, x) - \mu_P) \right].
   \]

   **Note:** \( \|A\| \leq \|A\|_{\text{HS}} \leq \|A\|_1 \).
Finite-sample guarantee

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

$$\left| \widehat{\text{MMD}}_Q(\mathbb{P}, \mathbb{Q}) - \text{MMD}(\mathbb{P}, \mathbb{Q}) \right| \leq f(N, \Sigma_\mathbb{P}, \Sigma_\mathbb{Q}, \eta, \delta).$$
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. $\geq 1 - \eta$

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

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

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

$$\left|\hat{\text{MMD}}_Q(P, Q) - \text{MMD}(P, Q)\right| \leq 12 \max \left( 2\sqrt{\frac{\text{Tr} (\Sigma_P) + \text{Tr} (\Sigma_Q)}{N}}, \frac{2}{\delta} \sqrt{\left(\frac{\|\Sigma_P\| + \|\Sigma_Q\|}{\delta N}\right)\ln(1/\eta)} \right).$$

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

Zoltán Szabó  The MONK
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. $\geq 1 - \eta$

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

- $\delta$-dependence: optimal?
Finite-sample guarantee

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

$$\left| \widehat{\text{MMD}}_Q(P, Q) - \text{MMD}(P, Q) \right|$$

$$\leq 12 \max \left( 2\sqrt{\frac{\text{Tr} (\Sigma_P) + \text{Tr} (\Sigma_Q)}{N}}, \sqrt{\frac{\|\Sigma_P\| + \|\Sigma_Q\| \ln(1/\eta)}{\delta N}} \right).$$

- Breakdown point can be 25% (asymptotic behavior).
No outliers / bounded kernel: MONK is a safe alternative.

Relevant case: outliers & unbounded kernel.

- $P := \mathcal{N}(\mu_1, \sigma_1^2) \neq \mathcal{Q} := \mathcal{N}(\mu_2, \sigma_2^2)$. $\mu_m, \sigma_m \sim U[0,1]$, fixed.
- $N \in \{200, 400, \ldots, 2000\}$.
- 5-5 corrupted samples: $(x_n)_{n=N-4}^N = 2000$, $(y_n)_{n=N-4}^N = 4000$.
- $(P, Q, K)$: $\text{MMD}(P, Q)$ is analytic.
- Performance:
  - 100 MC simulations,
  - median and quartiles.
Numerical demo: quadratic kernel, $N_c = 5$ outliers

![Graph]

- U-Stat
- MONK BCD Q=3
- MONK BCD Q=11
- MONK BCD-Fast Q=11

$\ln(10(|MMD - MMD|))$ vs $\ln($ time in s $)$
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 $\text{MMD} - \hat{q}_{1-\alpha}$.
- $\hat{q}_{1-\alpha}$: Using 150 bootstrap permutations.
Inter-class: EI-IE

DNA analysis: plots

Zoltán Szabó  The MONK
DNA analysis: plots

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).


Hanson, D. and Wright, F. (1971). 
A bound on tail probabilities for quadratic forms in independent random variables. 

Testing for homogeneity with kernel Fisher discriminant analysis. 

Retrospective multiple change-point estimation with kernels. 
*In IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772.

Convolution kernels on discrete structures.
Technical report, Department of Computer Science, University of California at Santa Cruz.


The Kendall and Mallows kernels for permutations.

Interpretable distribution features with maximum testing power.

An adaptive test of independence with analytic kernel embeddings.
In *International Conference on Machine Learning (ICML; PMLR)*, volume 70, pages 1742–1751. PMLR.

A linear-time kernel goodness-of-fit test.
(best paper award = in top 3 out of 3240 submissions).


Bayesian approaches to distribution regression.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Generative moment matching networks.

Automatic construction and natural-language description of nonparametric regression models.

Text classification using string kernels.


Kernel-based tests for joint independence.

*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.


Kernel belief propagation.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715.

Learning theory for distribution regression.

Testing for equal distributions in high dimension.
*InterStat*, 5.

A new test for multivariate normality.
Tolstikhin, I., Sriperumbudur, B. K., and Schölkopf, B. (2016).

Minimax estimation of maximal mean discrepancy with radial kernels.

Consistency of robust kernel density estimators.

Graph kernels.

Post selection inference with kernels.
Deep sets. 
In Advances in Neural Information Processing Systems (NIPS), 
pages 3394–3404.

Domain adaptation under target and conditional shift. 