MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting

Antonin Schrab
University College London
Centre for Artificial Intelligence
Gatsby Computational Neuroscience Unit
Inria London

MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting



Felix Biggs[⋆] [†§]

Antonin Schrab[⋆] [† ‡ §]

Arthur Gretton [‡]

† Centre for Artificial Intelligence, UCL
‡ Gatsby Computational Neuroscience Unit, UCL
§ Inria London Programme

⋆ *Joint first author*

Two-sample testing

- samples $\mathbb{X}_m := (X_1, \ldots, X_m)$, $X_i \overset{\text{iid}}{\sim} p$ in $\mathbb{R}^d$
- samples $\mathbb{Y}_n := (Y_1, \ldots, Y_n)$, $Y_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

$$\begin{array}{ccc} \mathcal{H}_0 \colon p = q & \text{against} & \mathcal{H}_1 \colon p \neq q \\ \Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1 & \Longleftrightarrow & \text{reject } \mathcal{H}_0 \\ \Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0 & \Longleftrightarrow & \text{fail to reject } \mathcal{H}_0 \end{array}$$

Test construction: need a distance between p, q (between $\mathbb{X}_m, \mathbb{Y}_n$)

Type I error: controlled by α by design

$$\mathbb{P}_{p \times p}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \ \leq \ \alpha$$

Type II error: find a condition on $\mathrm{dist}(p, q)$ to control by β

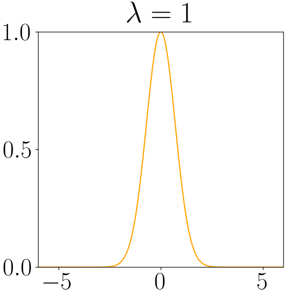
$$\mathbb{P}_{p \times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \ \leq \ \beta$$

Kernels and bandwidths

Kernel: $k_\lambda(x, y) := K\left(\frac{x - y}{\lambda}\right)$ **Bandwidth:** $\lambda > 0$

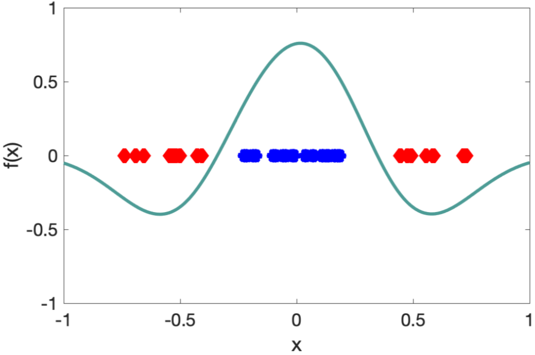
Gaussian kernel: $K(u) = \exp\left(-\|u\|^2\right), u \in \mathbb{R}^d$

$$k_\lambda(x, y) := \exp\left(-\|x - y\|^2/\lambda^2\right)$$



Kernel: $k_\lambda(x, y) := K\left(\frac{x - y}{\lambda}\right)$ **Bandwidth:** $\lambda > 0$

$$\mathrm{MMD}_\lambda(p, q) := \sup_{f \in \mathcal{H}_\lambda : \|f\|_{\mathcal{H}_\lambda} \leq 1} \left| \mathbb{E}_{X \sim p}[f(X)] - \mathbb{E}_{Y \sim q}[f(Y)] \right|$$



$p \neq q$

Kernel bandwidth intuition

- **Small sample sizes**: only global differences are detectable
  - **Small bandwidth**: wrongly detects artificial local differences under $\mathcal{H}_0$
  - **Large bandwidth**: well-suited to detect global differences under $\mathcal{H}_1$
- **Large sample sizes**: local differences are detectable
  - **Small bandwidth**: well-suited to detect local differences under $\mathcal{H}_1$
  - **Large bandwidth**: fails to detect local differences under $\mathcal{H}_1$

$\Longrightarrow$ **Bandwidths** should decrease as the **sample sizes** increase

- Choice of **kernel/bandwidth** is **crucial** for test power!
- **Kernel** selection: **median heuristic**, **data splitting**, & **aggregation**
- **Contribution:** propose a new method for **kernel** selection
  - no heuristic, no data splitting, no multiple testing

Maximum Mean Discrepancy estimator

$$\begin{aligned}\mathrm{MMD}_k^2(p,q) \;:=\;& \mathbb{E}_{p,p}[k(X,X')] \\ & - 2\,\mathbb{E}_{p,q}[k(X,Y)] \\ & + \mathbb{E}_{q,q}[k(Y,Y')]\end{aligned}$$

$$\begin{aligned}\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m,\mathbb{Y}_n) \;:=\;& \frac{1}{m(m-1)}\sum_{1\le i\neq i'\le m} k(X_i,X_{i'}) \\ & - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(X_i,Y_j) \\ & + \frac{1}{n(n-1)}\sum_{1\le j\neq j'\le n} k(Y_j,Y_{j'})\end{aligned}$$

Statistic construction

- For $k \in K$, compute $\widehat{\mathrm{MMD}}^2_k(\mathbb{X}_m, \mathbb{Y}_n)$ ... then what?
- Take the maximum:

$$\max_{k \in K} \widehat{\mathrm{MMD}}^2_k(\mathbb{X}_m, \mathbb{Y}_n)$$

  - **Issue:** kernels have different scales
  - **Issue:** kernels lead to difference variances $\mathrm{var}_{p \times q}\left(\widehat{\mathrm{MMD}}^2_k\right)$
- Normalise the MMD values:

$$\max_{k \in K} \widehat{\mathrm{MMD}}^2_k(\mathbb{X}_m, \mathbb{Y}_n) \ / \ \widehat{\mathrm{N}}_k(\mathbb{X}_m, \mathbb{Y}_n)$$

  - **Issue:** difficult to work with mathematically
- Relaxation via soft-maximum for $\eta > 0$:

$$\frac{1}{\eta} \log \left(\frac{1}{|K|} \sum_{k \in K} \exp \left(\eta \ \widehat{\mathrm{MMD}}^2_k(\mathbb{X}_m, \mathbb{Y}_n) \ / \ \widehat{\mathrm{N}}_k(\mathbb{X}_m, \mathbb{Y}_n) \right) \right)$$

Soft-maximum

$$
\begin{aligned}
&\max_{k\in K} \frac{\widehat{\mathrm{MMD}}^2_k}{\widehat{\mathrm{N}}_k} - \frac{\log(|K|)}{\eta} \\
\leq\ & \frac{1}{\eta} \log\left(\frac{1}{|K|}\sum_{k\in K} \exp\left(\eta\, \frac{\widehat{\mathrm{MMD}}^2_k}{\widehat{\mathrm{N}}_k}\right)\right) \\
\leq\ & \max_{k\in K} \frac{\widehat{\mathrm{MMD}}^2_k}{\widehat{\mathrm{N}}_k}
\end{aligned}
$$

Convergence to the maximum as $\eta \to \infty$

MMD-FUSE: Fusing U-Statistics by Exponentiation

Donsker-Varadhan equality

$$\frac{1}{\eta} \log \left(\mathbb{E}_{k\sim\pi} \exp \left(\eta \, \frac{\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m, \mathbb{Y}_n)}{\widehat{\mathrm{N}}_k(\mathbb{X}_m, \mathbb{Y}_n)} \right) \right)$$

$$= \sup_{\rho} \; \mathbb{E}_{k\sim\rho} \left[\frac{\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m, \mathbb{Y}_n)}{\widehat{\mathrm{N}}_k(\mathbb{X}_m, \mathbb{Y}_n)} \right] - \frac{\mathrm{KL}(\rho, \pi)}{\eta}$$

Un-normalised version: $\widehat{\mathrm{N}} = 1$

$$\sup_{\rho} \; \mathbb{E}_{k\sim\rho} \left[\widehat{\mathrm{MMD}}_k^2(\mathbb{X}_m, \mathbb{Y}_n) \right] - \frac{\mathrm{KL}(\rho, \pi)}{\eta}$$

$$= \sup_{\rho} \; \widehat{\mathrm{MMD}}_{\mathbb{E}_{k\sim\rho}[k]}^2(\mathbb{X}_m, \mathbb{Y}_n) - \frac{\mathrm{KL}(\rho, \pi)}{\eta}$$

Mean kernel: Gaussian kernel, Gamma prior $\Longrightarrow$ rational quadratic

Continuously optimizing the kernel bandwidth

Permutation test

$$\Delta_\alpha(\mathbb{X}_m, \mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{FUSE}}(\mathbb{X}_m, \mathbb{Y}_n) > \widehat{q}_{1-\alpha}\right)$$

Quantile: $\widehat{q}_{1-\alpha}$ is the $\lceil (B+1)(1-\alpha) \rceil$-th largest value of $\widehat{\mathrm{FUSE}}(\mathbb{X}_m, \mathbb{Y}_n)$ and B permuted test statistics

Permutations: $\widehat{\mathrm{FUSE}}(\mathbb{X}^\sigma_m, \mathbb{Y}^\sigma_n)$ where $(\mathbb{X}^\sigma_m, \mathbb{Y}^\sigma_n) = \sigma(\mathbb{X}_m \cup \mathbb{Y}_n)$

Non-asymptotic level: $\mathbb{P}_{p\times p}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \;\leq\; \alpha$

Hemerik and Goeman, 2018: control of type I error holds for any test parameters selected using all the data in a permutation-free way

Deep kernel: learn representation of ϕ via unsupervised methods without data splitting

$$k\big(\phi(x), \phi(y)\big)$$

Assumption: $\mathbb{E}_{p\times q}\left[1/\widehat{\mathrm{N}}_k(\mathbb{X}_m, \mathbb{Y}_n)\right]$ is bounded for all $k \in \mathrm{supp}(\pi)$

Satisfied for kernels that tend to zero only for data infinitely far apart

Power condition: The type II error is controlled

$$\mathbb{P}_{p\times q}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \;\leq\; \beta$$

if there exists a distribution ρ on kernels such that (WLOG $n \leq m$)

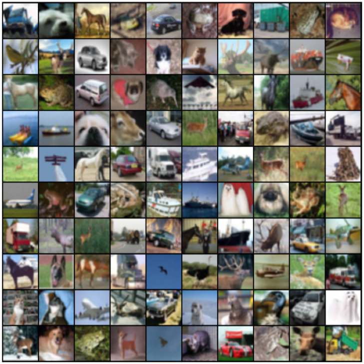
$$\mathrm{MMD}^2_{\mathbb{E}_{k\sim\rho}[k]}(p, q) \;\geq\; \frac{C}{n}\left(\frac{1}{\beta^2} + \log\left(\frac{1}{\alpha}\right) + \mathrm{KL}(\rho, \pi)\right)$$

Domingo-Enrich et al., 2023:

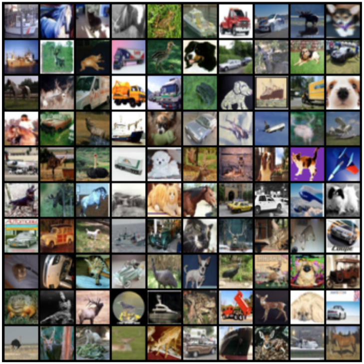
The MMD separation rate is optimal wrt sample size n.

Experiments

Experiment: CIFAR 10 vs CIFAR 10.1



(a) CIFAR-10 images



(b) CIFAR-10.1 images

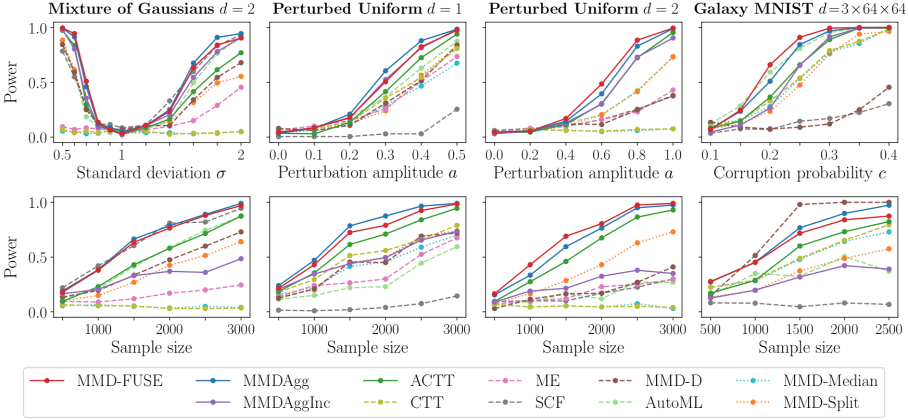
Figure 6: Images from the CIFAR-10 (Krizhevsky, 2009) and CIFAR-10.1 (Recht et al., 2019) test sets. This figure corresponds to Figure 5 of Liu et al. (2020).

Experiment: CIFAR 10 vs CIFAR 10.1

Table 1: Test power for detecting the difference between CIFAR-10 and CIFAR-10.1 images with test level $\alpha = 0.05$. The averaged numbers of rejections over 1000 repetitions are reported.

Tests	Power
MMD-FUSE	**0.937**
MMDAgg	0.883
MMD-D	0.744
CTT	0.711
MMD-Median	0.678
ACTT	0.652
ME	0.588
AutoML	0.544
C2ST-L	0.529
C2ST-S	0.452
MMD-O	0.316
MMDAggInc	0.281
SCF	0.171

More experiments



Thank you for your attention!
Any Questions?

SCAN ME

Arxiv

SCAN ME

Github