Maximum Mean Discrepancy with Unequal Sample Sizes via Generalized U-Statistics

Aaron Wei[1] Milad Jilali[2] Danica J. Sutherland[1,3]

[1]University of British Columbia [2]Independent Researcher [3]Alberta Machine Intelligence Institute

Published at TMLR (May 2026) openreview.net/forum?id=KjXW75GHHF

Two-sample Testing with MMD

We have i.i.d. samples $(x_i)_{i=1}^{n_X} \sim P$ and $(y_j)_{j=1}^{n_Y} \sim Q$.
We want to know if P and Q are the same distribution:

$$H_0 : P = Q \quad \text{vs.} \quad H_1 : P \neq Q.$$

The kernel **Maximum Mean Discrepancy (MMD)** [1]

$$\mathrm{MMD}(P, Q) := \sqrt{\mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')] - 2\,\mathbb{E}[k(X, Y)]}$$

is a metric for many k (*characteristic* kernels), leading to

$$H_0 : \mathrm{MMD}(P, Q) = 0 \quad \text{vs.} \quad H_1 : \mathrm{MMD}(P, Q) \neq 0.$$

We should use permutation to set test thresholds, but knowing asymptotics still helps! Especially, if $\sqrt{n}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2)/\sigma \xrightarrow{d} \mathcal{N}(0, 1)$, then

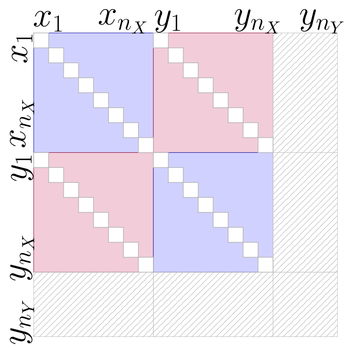
$$\Pr\left(n\,\widehat{\mathrm{MMD}}^2 > c_\alpha\right) \sim \Phi\left(\sqrt{n}\frac{\mathrm{MMD}^2}{\sigma} - \frac{c_\alpha}{\sqrt{n}\,\sigma}\right).$$

RHS is approximately monotone in MMD^2/σ, so choosing k to maximize this signal-to-noise ratio gives better test power [2].

Shortcomings in MMD Estimator Theory

Existing asymptotics to get σ above (e.g. in [2]) are for a *U*-statistic estimator:
letting $n_{\min} = \min\{n_X, n_Y\}$,

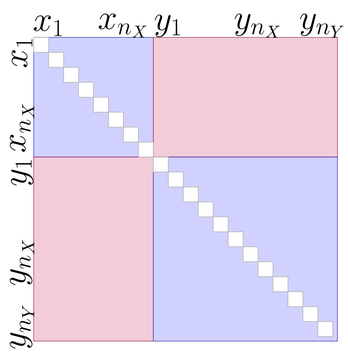
$$\widehat{\mathrm{MMD}_{\mathrm{U}}}^2 := \binom{n_{\min}}{2}^{-1} \sum_{1 \le i < j \le n_{\min}} \Big[k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i) \Big].$$



Lets us use simple textbook results to get the asymptotics.
But if $n_X \neq n_Y$, we either waste data or use theory for the wrong estimator!

In practice, we estimate with all the data, which we can write as a generalized U-statistic:

$$\widehat{\mathrm{MMD}}^2 := \binom{n_X}{2}^{-1}\binom{n_Y}{2}^{-1} \sum_{1 \le i < j \le n_X} \sum_{1 \le a < b \le n_Y} \Big[k(x_i, x_j) + k(y_a, y_b) - \frac{1}{2}[k(x_i, y_a) + k(x_j, y_a) + k(x_i, y_b) + k(x_j, y_b)] \Big].$$



This incorporates all data, so it has lower variance than $\widehat{\mathrm{MMD}_{\mathrm{U}}}$.
But generalized U-statistics are understudied.
So, we study them! Our general results specialize readily to MMD:

Asymptotic distribution of $\widehat{\mathrm{MMD}}^2$

Suppose $n_X, n_Y \to \infty$ with $n_X/(n_X + n_Y) \to r_X \in [0, 1]$.
Take suitable regularity conditions; let $n_{\mathrm{harm}} = \frac{2}{1/n_X + 1/n_Y}$. Under H_0,

$$n_{\mathrm{harm}}\,\widehat{\mathrm{MMD}}^2 \xrightarrow{d} 2\sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1)$$

where the Z_i are i.i.d. standard normal and λ_i are the eigenvalues of a particular trace-class operator depending on P and k. Under H_1,

$$\sqrt{n_{\mathrm{harm}}}(\widehat{\mathrm{MMD}}^2 - \widehat{\mathrm{MMD}}) \xrightarrow{d} \mathcal{N}\left(0, 8(1 - r_X)\zeta_P + 8 r_X \zeta_Q\right)$$

for $\zeta_P = \langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle$, $\zeta_Q = \langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle$.

Worth noting: [1] had (a more complicated form of) the H_0 result when $r_X \in (0, 1)$. Not possible to extend their result to cases where n_X, n_Y are of different orders!

Variance characterization

We also get nice exact, finite-sample variance expressions. Using

$$\mu_P = \mathop{\mathbb{E}}_{X \sim P}[k(X, \cdot)] \qquad C_P = \mathop{\mathbb{E}}_{X \sim P}[k(X, \cdot) \otimes k(X, \cdot)] - \mu_P \otimes \mu_P,$$

the kernel mean embedding and covariance operator, $\widehat{\mathrm{MMD}_{\mathrm{U}}}^2$'s variance is

$$\frac{4}{n}\langle \mu_P - \mu_Q, (C_P + C_Q)(\mu_P - \mu_Q)\rangle + \frac{2}{n(n-1)}\|C_P + C_Q\|_{\mathrm{HS}}^2$$

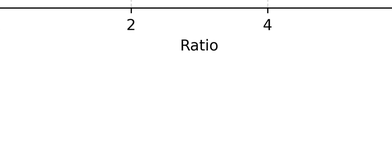
and $\widehat{\mathrm{MMD}}^2$'s is

$$\begin{aligned}
&\frac{4}{n_X}\left(1 - \frac{4}{n_Y}\cdot\frac{n_X - 2}{n_X - 1}\cdot\frac{n_Y - 2}{n_Y - 1}\right)\langle \mu_P - \mu_Q, C_P(\mu_P - \mu_Q)\rangle \\
&+ \frac{4}{n_Y}\left(1 - \frac{4}{n_X}\cdot\frac{n_X - 2}{n_X - 1}\cdot\frac{n_Y - 2}{n_Y - 1}\right)\langle \mu_P - \mu_Q, C_Q(\mu_P - \mu_Q)\rangle \\
&+ \frac{2}{n_X(n_X - 1)}\|C_P\|_{\mathrm{HS}}^2 + \frac{2}{n_Y(n_Y - 1)}\|C_Q\|_{\mathrm{HS}}^2 + \frac{4}{n_X n_Y}\langle C_P, C_Q\rangle_{\mathrm{HS}} \\
&+ \frac{16}{n_X n_Y}\left(\frac{n_X - 2}{n_X - 1}\cdot\frac{n_Y - 2}{n_Y - 1}\right)\left[\langle \mu_P, C_P\mu_P\rangle + \langle \mu_Q, C_Q\mu_Q\rangle\right].
\end{aligned}$$

Each of these terms is straightforward to estimate from samples.

Degeneracy (surprise!)

In an early draft of this paper, Danica wrote that the estimator is *degenerate* (mixture of chi-squareds asymptotics) iff $\mathrm{MMD}(P, Q) = 0$.
This is not true! Trivially: take P, Q distinct point masses.
Nontrivially: MMD > 0 but still mixture-of-chi-squareds!



$k(X, Y) = 0$, but $k(X, X')$ $k(Y, Y')$ have positive variance

But we prove this can't happen in "reasonable" situations.

CIFAR 10/10.1 Image Data Experiment

Horizontal axis: ratio of CIFAR-10 to CIFAR-10.1 data points.
Vertical: empirical power of a test based on $\widehat{\mathrm{MMD}}^2$ with the given ratio, and using a learned kernel based on a deep net.

RR: train to maximize the power of $\widehat{\mathrm{MMD}}^2$, using our variance estimator for the given ratio.
RE: train to maximize the power of $\widehat{\mathrm{MMD}_{\mathrm{U}}}^2$ on subsampled minibatches of equal size (optimizing for the wrong test) – an improved version of [2].

References

[1] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[2] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *ICML*, 2020.