A Semi-Supervised Kernel Two-Sample Test

AISTATS 2026 Spotlight

Gyumin Lee[1], Shubhanshu Shekhar[2], and Ilmun Kim[3]

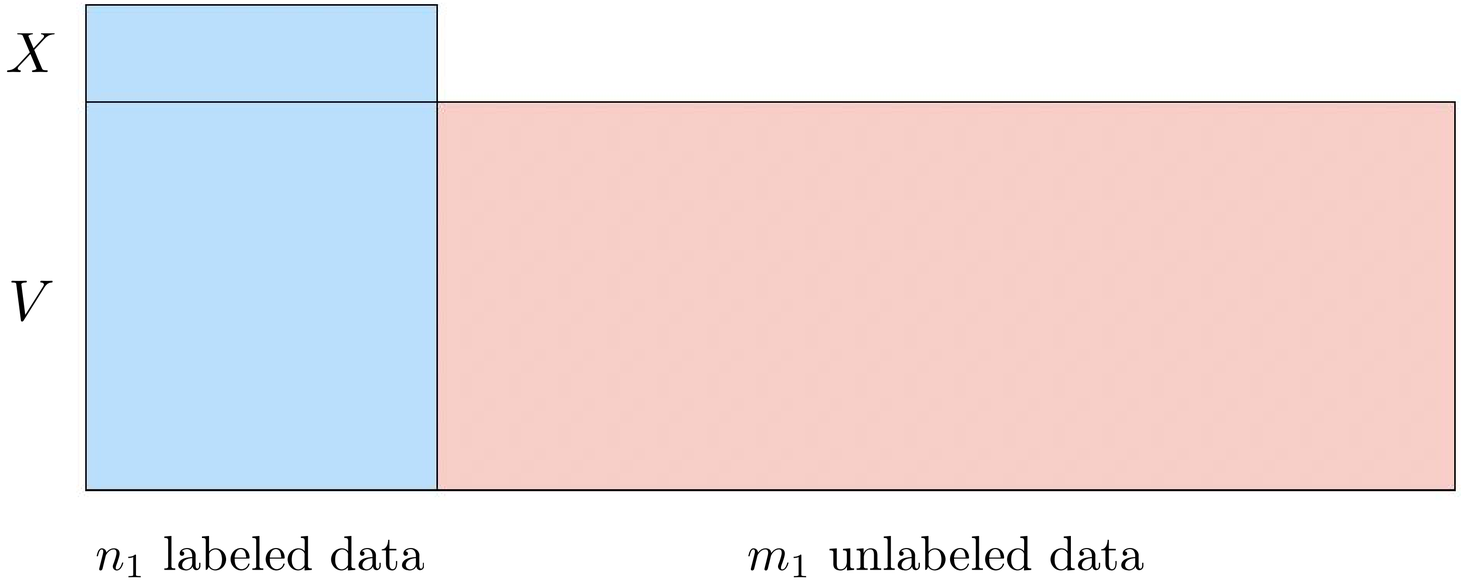
1) Pennsylvania State University 2) University of Michigan 3) Korea Advanced Institute of Science and Technology

May 4th, 2026

Motivation

Semi-Supervised Inference

: inference based on a (small) labeled dataset together with a (large) unlabeled dataset.



$$\widehat{\mu}_{X,f} := \frac{1}{n_1}\sum_{i=1}^{n_1}\{f(X_i) - \mathbb{E}[f(X_i)\,|\,V_i]\} + \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\mathbb{E}[f(X_i)\,|\,V_i]$$

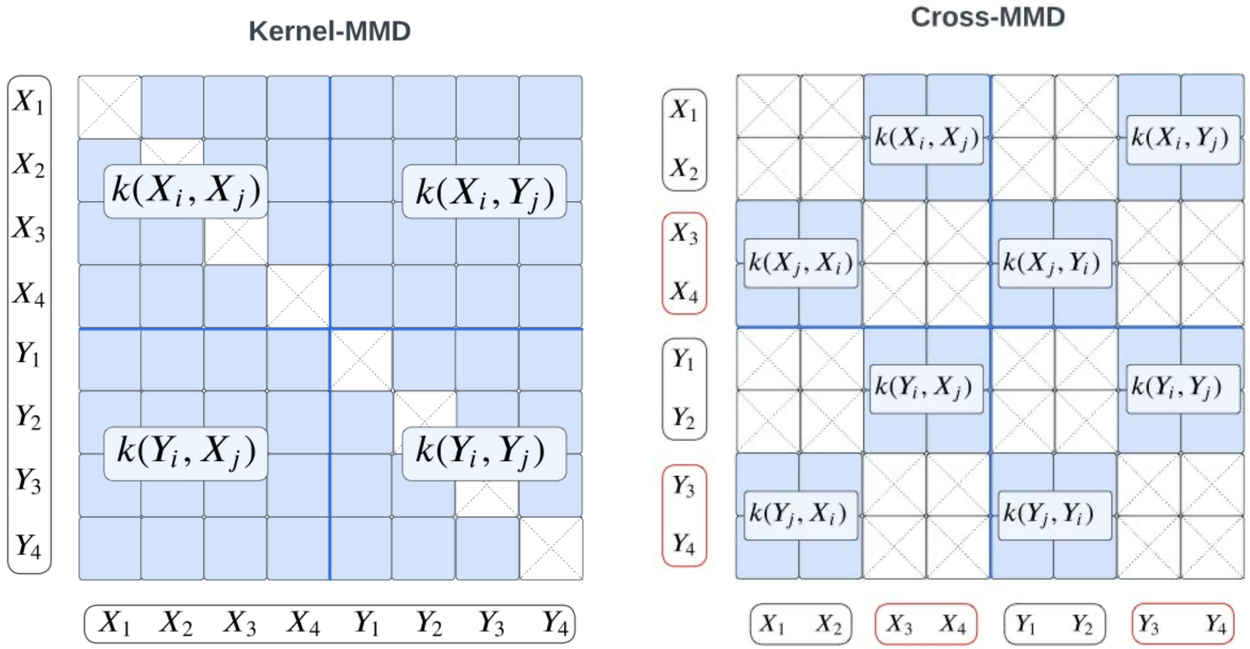
- Zhang, A., Brown, L. D., & Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means.
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669-674.

Motivation

Two-sample test: $H_0 : P_X = P_Y$ vs. $H_1 : P_X \neq P_Y$

Maximum Mean Discrepancy: $\mathrm{MMD}(P, Q; \mathcal{F}) = \sup_{\substack{f \in \mathcal{F} : \|f\| \leq 1 \\ \mathcal{F} \text{ is an RKHS}}} \left[\mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{x}')}[f(\boldsymbol{x}')] \right]$

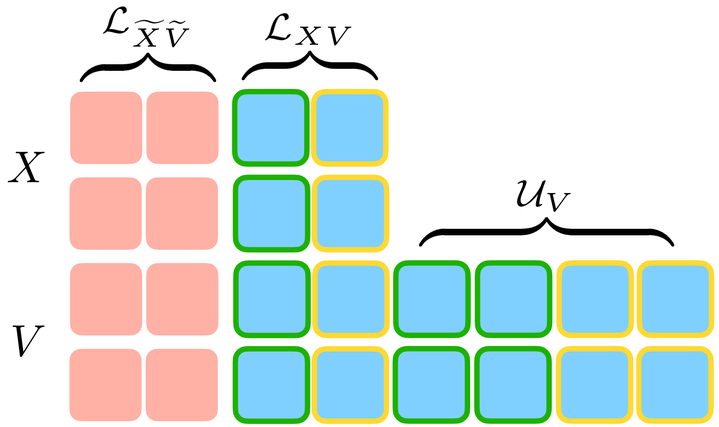
Permutation-free method:

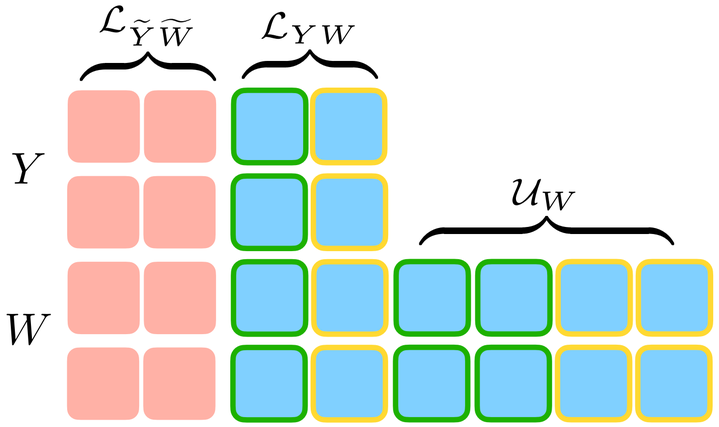


- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, *13*(1), 723-773.
- Kim, I., & Ramdas, A. (2020). Dimension-agnostic inference using cross U-statistics. *arXiv preprint arXiv:2011.05068*.
- Shekhar, S., Kim, I., & Ramdas, A. (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, *35*, 18168-18180.

xssMMD test: Semi-Supervised Kernel Two-Sample Test

Intuition: Use unlabeled data to refine the witness function and boost test power





① Construct $\widehat{f}(\cdot)$ from the first half

↓

② Use cross-fitting to estimate $\widehat{\mathbb{E}}[\widehat{f}(X) \mid V]$ and $\widehat{\mathbb{E}}[\widehat{f}(Y) \mid W]$

↓

③ Project the second half onto $\widehat{f}(\cdot)$

↓

④ Compute the test statistic

$$\widehat{\mathrm{xssMMD}}^2 = \frac{\widehat{\mu}^{\dagger}_{X,\widehat{f}} - \widehat{\mu}^{\dagger}_{Y,\widehat{f}}}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}}$$

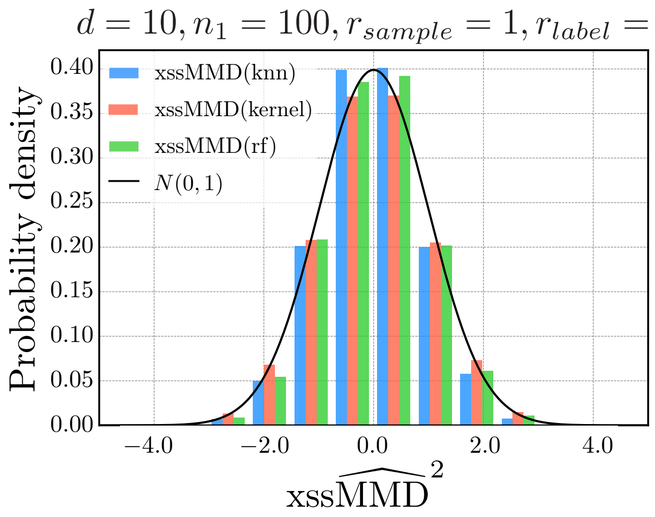
$$\widehat{f}(\cdot) = \frac{1}{n_1}\sum_{i=1}^{n_1} k(\widetilde{X}_i, \cdot) - \frac{1}{n_2}\sum_{i=1}^{n_2} k(\widetilde{Y}_i, \cdot)$$

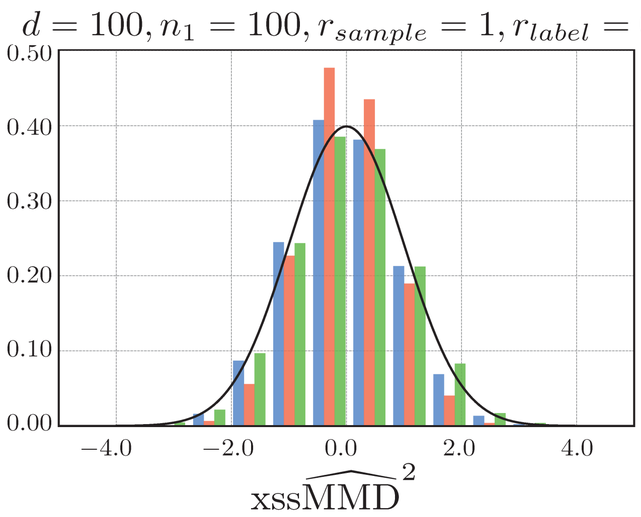
E.g. k-NN, random forest

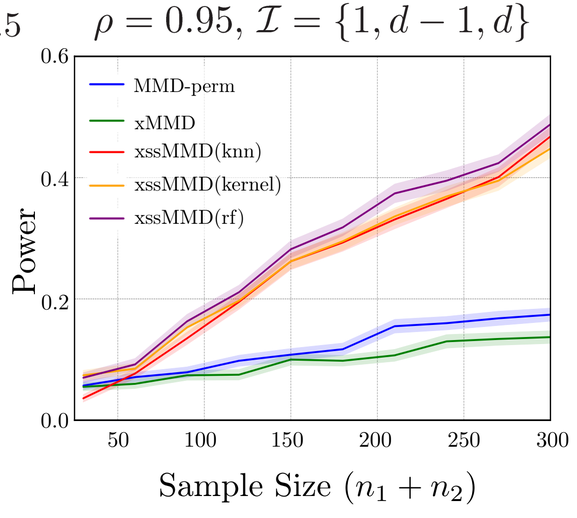
$$\hat{\mu}^{\dagger}_{X,\widehat{f}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\widehat{f}(X_i) - \widehat{\mathbb{E}}[\widehat{f}(X_i) \mid V_i]\right\} + \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\widehat{\mathbb{E}}[\widehat{f}(X_i) \mid V_i]$$

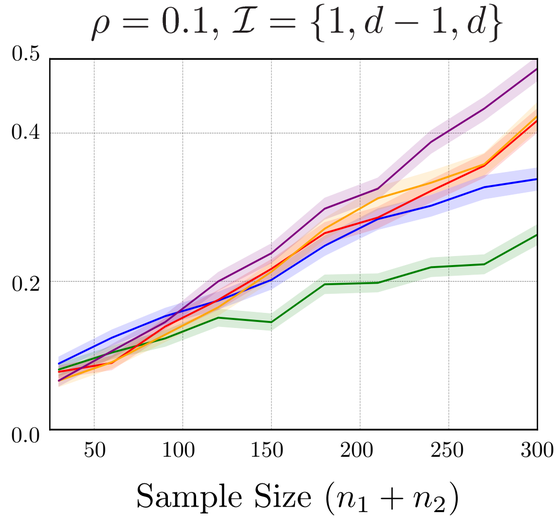
Theoretical & Numerical Results

- Asymptotic normality of xssMMD test under the null and alternative
- Asymptotic power guarantees of xssMMD compared to xMMD
- Power consistency of the xssMMD test against fixed and local alternatives



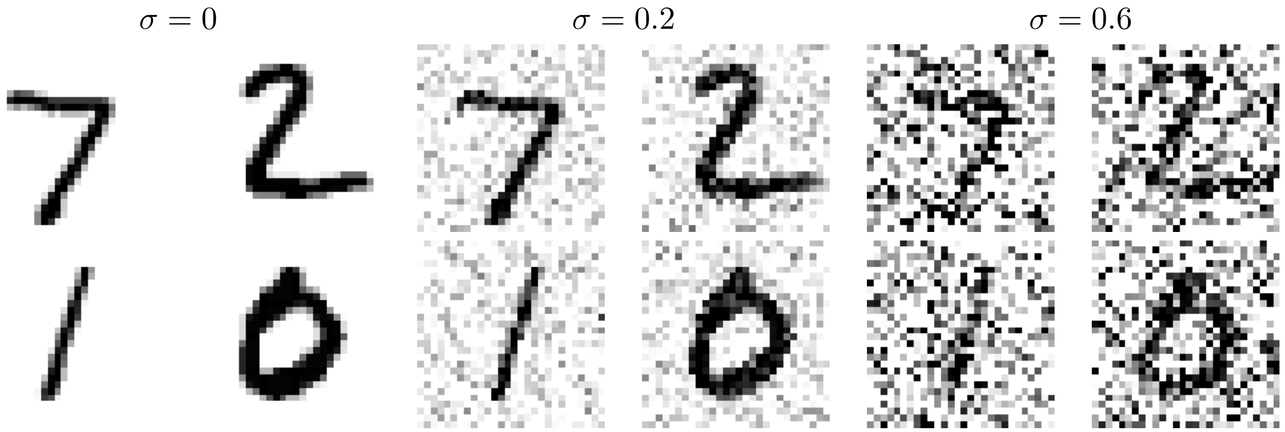




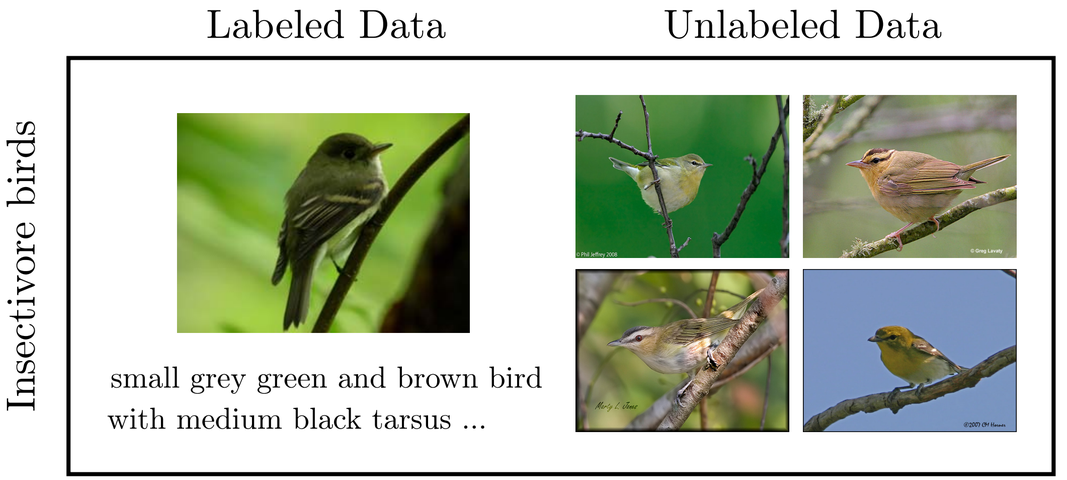


Experiments

- MNIST data



- CUB data



: xssMMD consistently outperforms standard tests!

Stop by Poster **178**

to continue the discussion!