Semi-Supervised Kernel Two-Sample Test

Statistics and Data Analysis
M.S.
Gyumin Lee

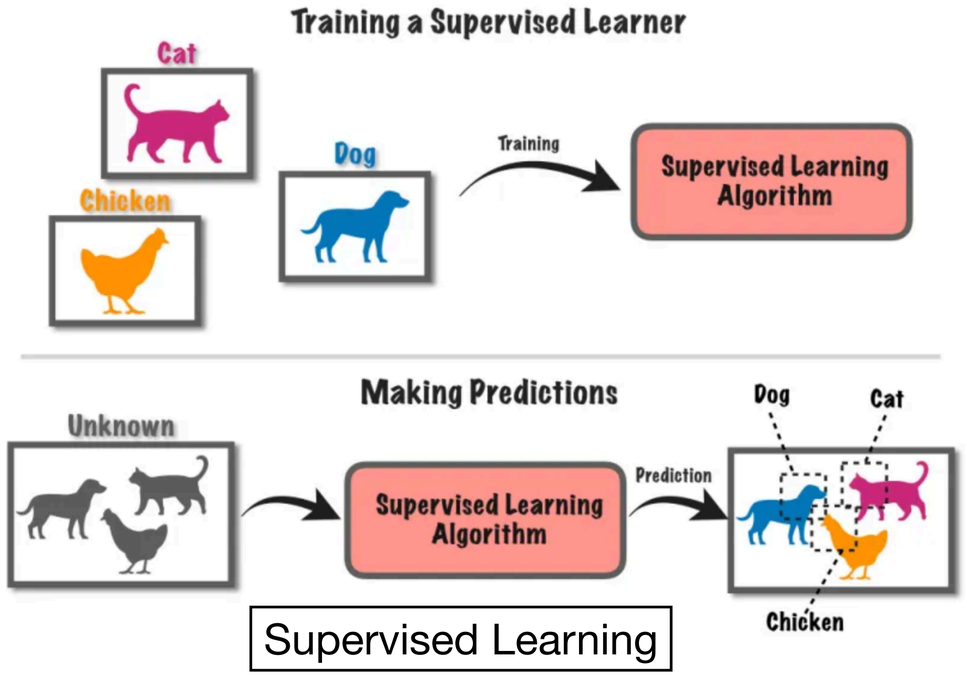
Contents

Semi-Supervised Inference

- Classic machine learning : supervised vs unsupervised

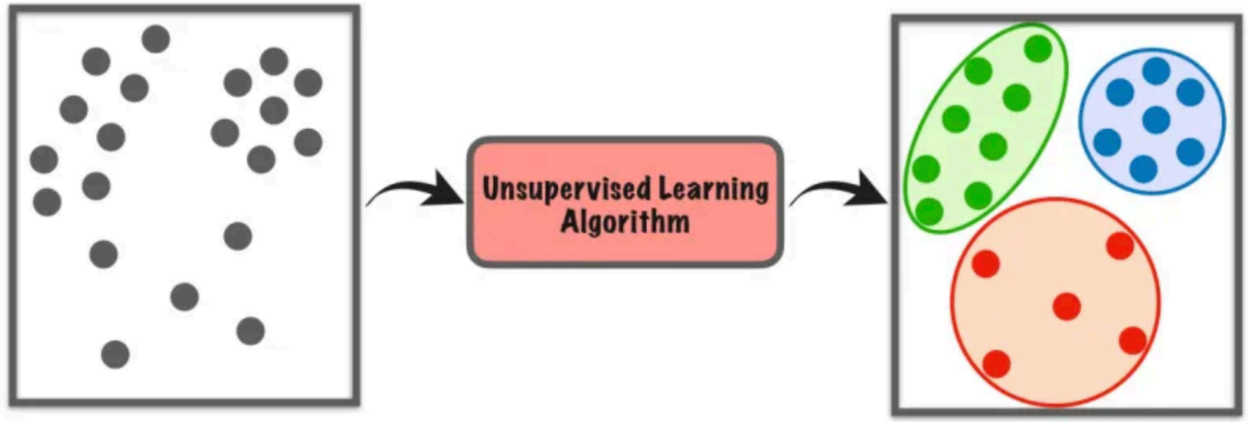
 - We know the outcome variables to guide the learning process?

- Or, we observe only the features and have no measurements of the outcome?



Supervised Learning

VS

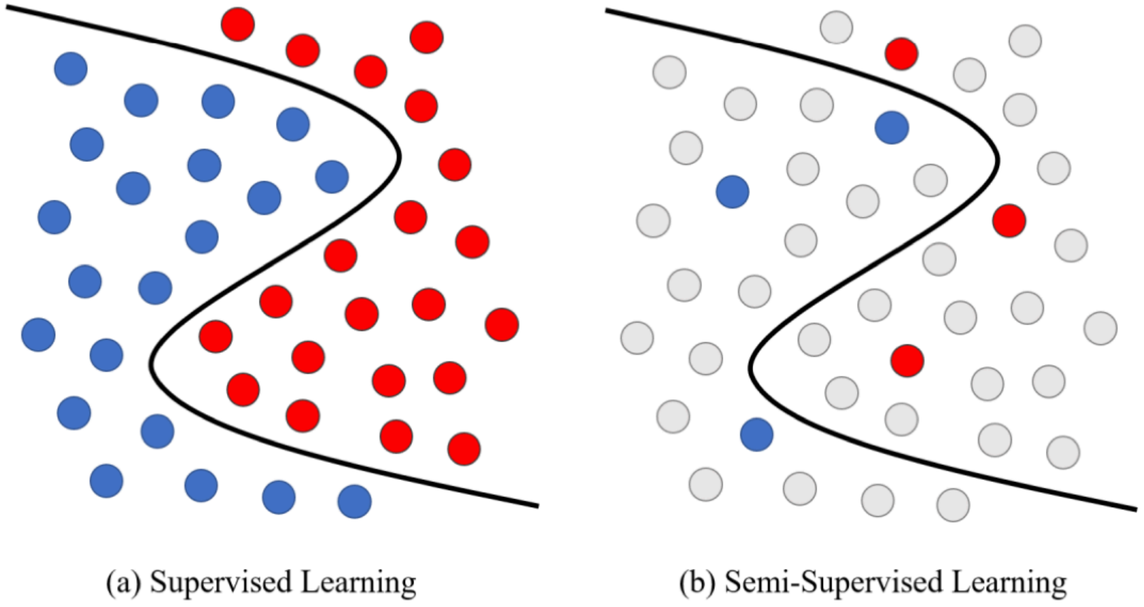


Unsupervised Learning

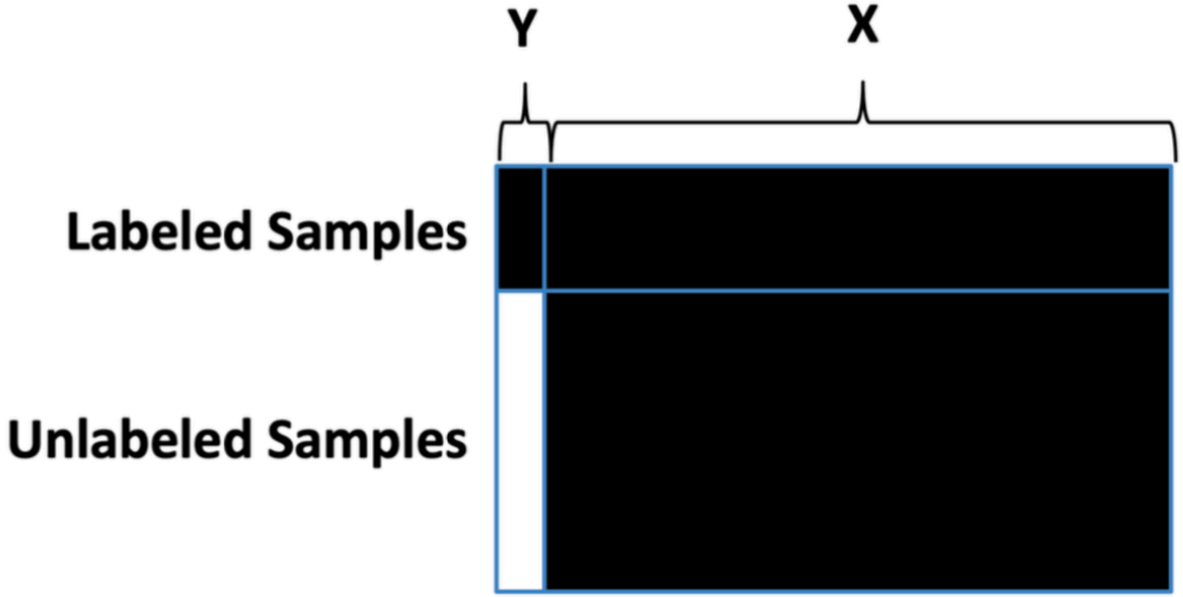
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242

Semi-Supervised Inference

Then, what is ‘Semi-supervised learning/setting’?
: problem of learning/setting based on a small labeled dataset together with a large unlabeled dataset.

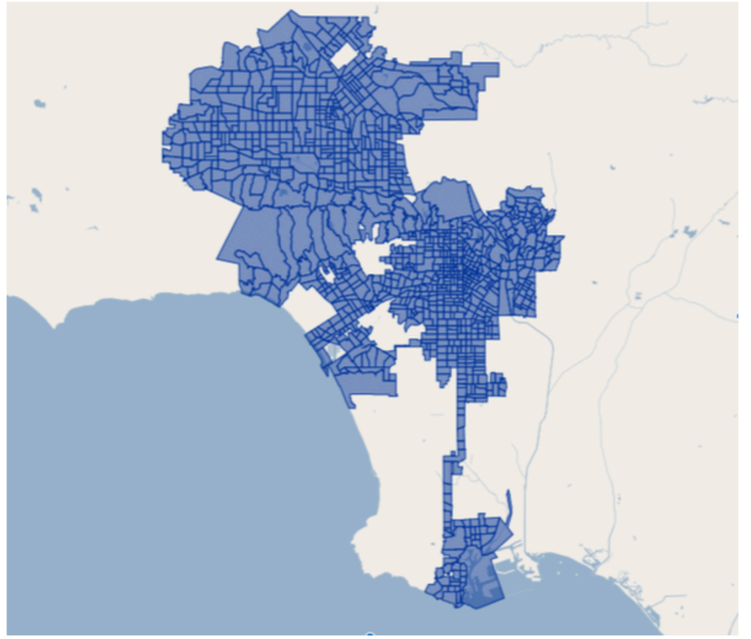


(a) Supervised Learning (b) Semi-Supervised Learning



- Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., & Jin, C. (2023). Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 107840.
- Zhang, A., Brown, L. D., & Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means.

Semi-Supervised Inference

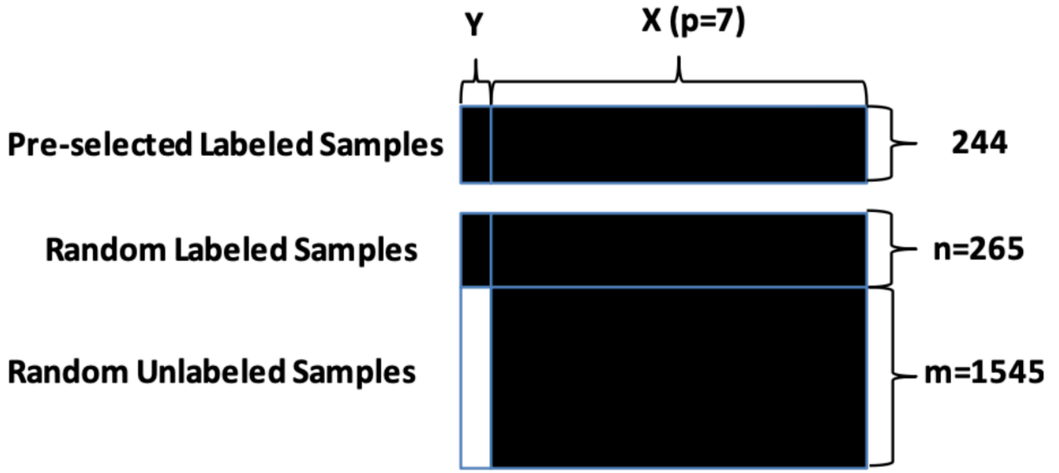


• Semi-supervised settings often appear across various fields handling regression / classification tasks.

• It plays an important role in many problems in statistics and machine learning, including model fine-tuning, model distillation, self-training, transfer learning and continual learning

Ex) Estimating Homeless in Los Angeles County

- inputs: 7 demographic covariates
- labels: the number of homeless people
- 265 labeled data set / 1545 unlabeled data set



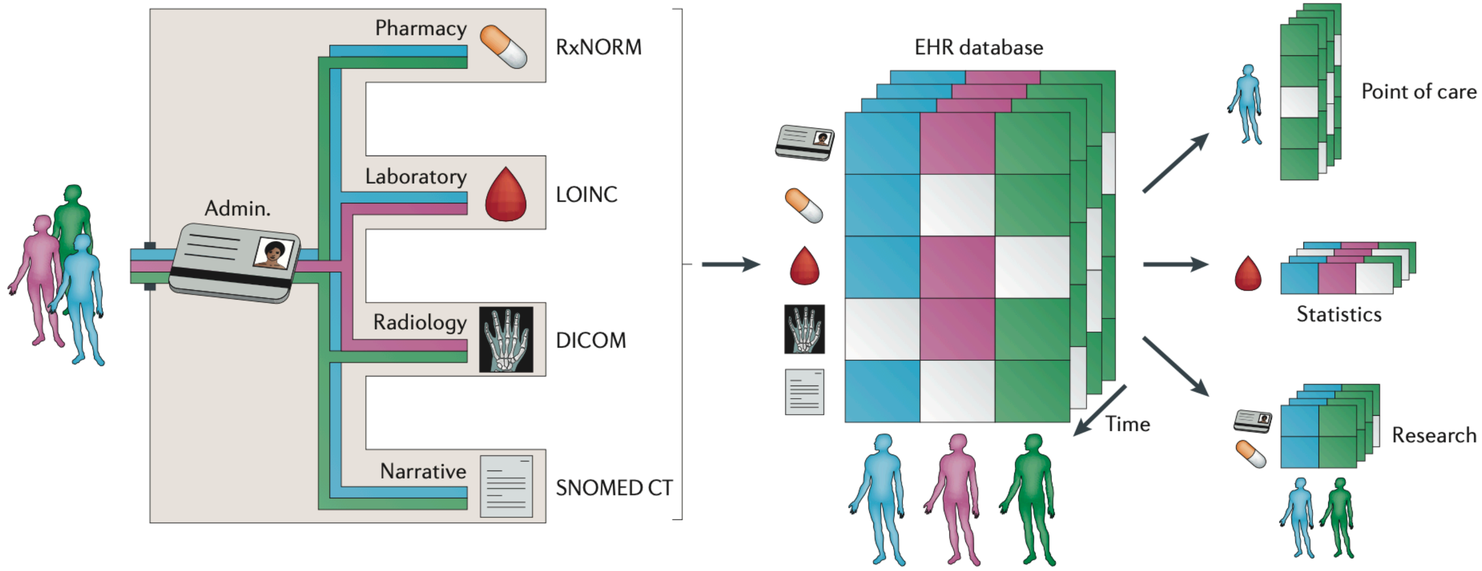
-Zhang, A., Brown, L. D., & Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means.

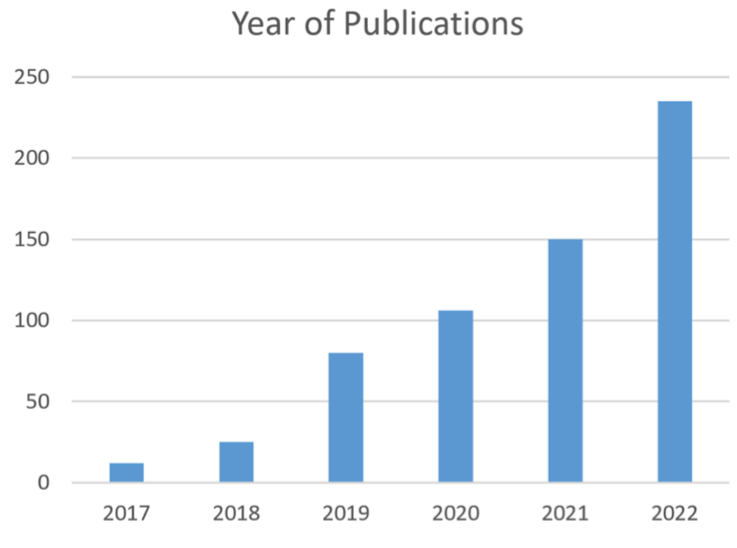
Semi-Supervised Inference

Ex) Object detection for autonomous driving
- 20K labeled images, 10M unlabeled images

Ex) Medical image segmentation

Ex) Electronic health records





- Han, J., Liang, X., Xu, H., Chen, K., Hong, L., Mao, J., ... & Xu, C. (2021). SODA10M: A large-scale 2D self/semi-supervised object detection dataset for autonomous driving. *arXiv preprint arXiv:2106.11118*.
- Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395-405.
- Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., & Jin, C. (2023). Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 107840.

Semi-Supervised Inference

- Basic Setting

$(Y, X_1, X_2, \cdots, X_p)$: $(p+1)$ dimensional random vector following an unkown joint distribution $P = P(dy, dx_1, \ldots, dx_p)$

P_X : marginal distribution of $X = (X_1, X_2, \cdots, X_p)$

Often, we suppose that there are n labeled samples from P,

$$[\mathbf{Y}, \mathbf{X}] = \{Y_k, X_{k1}, X_{k2}, \cdots, X_{kp}\}_{k=1}^{n}$$

and m unlabeled samples from P_X ,

$$\mathbf{X}_{\text{add}} = \{X_{k1}, X_{k2}, \cdots, X_{kp}\}_{k=n+1}^{n+m}$$

-Zhang, A., Brown, L. D., & Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means.

Semi-Supervised Inference

• Often, our goal is to estimate some estimand

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \mathbb{E}\left[\ell_\theta(X_1, Y_1)\right]$$

which could be applied for many different inference problems

- mean estimation $\theta^* = \underset{\theta \in \mathbb{R}}{\arg\min}\, \mathbb{E}[\ell_\theta(Y_1)] = \underset{\theta \in \mathbb{R}}{\arg\min}\, \mathbb{E}\left[\frac{1}{2}(Y_1 - \theta)^2\right]$

- quantile estimation $\theta^* = \min\left\{\theta : P\left(Y_1 \leq \theta\right) \geq q\right\}$

- inference for linear regression $\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}[\ell_\theta(X_1, Y_1)] = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}[(Y_1 - X_1^\top \theta)^2]$

and so on…

- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669-674.

Semi-Supervised Inference

• Simple example from ‘Mathematical Statistics 1’

Let our goal of estimation be $\theta^* = E[Y]$

We have several candidates of estimators (f is a prediction function)

- sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ vs semi-supervised mean $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i)) + \frac{1}{n+m}\sum_{i=1}^{n+m} f(X_i)$

- Both are unbiased, so we compare variance…

$$\mathrm{Var}(\bar{Y}) = \mathrm{Var}(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} Y_i) = \tfrac{1}{n}\mathrm{Var}(Y)$$

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i)) + \frac{1}{n+m}\sum_{i=1}^{n+m} f(X_i)\right)$$

$$= \frac{1}{n}\left(\mathrm{Var}(Y) + \frac{m-n}{m}\left(\mathrm{Var}(f(X)) - 2\mathrm{Cov}(Y, f(X))\right)\right)$$

If $\mathrm{Var}(f(X)) < 2\mathrm{Cov}(Y, f(X))$, then semi-supervised mean improves the original one!

-Ilmun Kim, (2023), Lecture notes “Selective Topics in Mathematical Statistics” for graduate mathematical statistics course

Semi-Supervised Inference

• Semi-supervised mean estimation in regression setting (Zhang, A. et al., 2019)

- For the ideal semi-supervised inference, where $m = \infty$

we have 'least square estimator' $\hat{\theta}_{LS} = \boldsymbol{\mu}'\hat{\beta} = \hat{\beta}_1 + \mu'\hat{\beta}_{(2)} = \bar{Y} - \hat{\beta}'_{(2)}(\bar{X} - \mu).$

- For the ordinary semi-supervise inference, where $m < \infty$

we have 'semi-supervised least squared estimator' $\hat{\theta}_{SSLS} = \hat{\boldsymbol{\mu}}'\hat{\beta} = \bar{Y} - \hat{\beta}'_{(2)}(\bar{X} - \hat{\mu}).$

- Both have asymptotic normality

$$\frac{\hat{\theta}_{\mathrm{LS}} - \theta}{\tau/\sqrt{n}} \xrightarrow{d} N(0,1) \quad \Big| \quad MSE/\tau^2 \xrightarrow{d} 1, \quad where \quad MSE := \frac{\sum_{i=1}^{n}(Y_i - \vec{X}_i^{\top}\hat{\beta})^2}{n-p-1}. \qquad + \ |F_n(x) - \Phi(x)| \le Cn^{-1/4}$$

$$\frac{\sqrt{n}(\hat{\theta}_{\mathrm{SSLS}} - \theta)}{\nu} \xrightarrow{d} N(0,1) \quad \Big| \quad \frac{\hat{\nu}^2}{\nu^2} \xrightarrow{d} 1 \quad where\ \hat{\nu}^2 = \tfrac{m}{m+n}MSE + \tfrac{n}{m+n}\hat{\sigma}_Y^2\ with\ MSE = \tfrac{1}{n-p-1}\textstyle\sum_{k=1}^{n}(Y_i - \vec{X}_k^{\top}\hat{\beta})^2\ and\ \hat{\sigma}_Y^2 = \tfrac{1}{n-1}\textstyle\sum_{k=1}^{n}(Y_i - \bar{\mathbf{Y}})^2.$$

-Zhang, A., Brown, L. D., & Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means.

Semi-Supervised Inference

• Semi-supervised mean estimation in regression setting (Zhang, A. et al., 2019)

- Comparison with respect to ℓ_2 Risk

$$n\mathbb{E}\left(\hat{\theta}^1_{\mathrm{LS}} - \theta\right)^2 = \tau_n^2 + s_n \text{ where } s_n = \frac{p^2}{n}A_{n,p} + \frac{p^2}{n^{5/4}}B_{n,p}, \quad \max(|A_{n,p}|, |B_{n,p}|) \le C$$

$$n\mathbb{E}\left(\hat{\theta}^1_{\mathrm{SSLS}} - \theta\right)^2 = \tau_n^2 + \frac{n}{n+m}\beta_{(2),n}^\top \Sigma_n \beta_{(2),n} + s_{n,m} \text{ where } |s_{n,m}| \le \frac{Cp^2}{n}.$$

$$\Rightarrow n\mathbb{E}\left(\hat{\theta}_{\mathrm{SSLS}} - \theta\right)^2 = \frac{n}{n+m}n\mathbb{E}(\bar{\mathbf{Y}} - \theta)^2 + \frac{m}{n+m}n\mathbb{E}(\hat{\theta}_{\mathrm{LS}} - \theta)^2 \quad \left(n\mathbb{E}\left(\bar{Y} - \theta\right)^2 = \tau^2 + \beta_{(2)}^\top \Sigma \beta_{(2)} \right)$$

• Other semi-supervised inference works are still ongoing…

- High-dimensional & mean(Zhang, Y. et al., 2022), high-dimensional & variance (Tony Cai, T. et al., 2020), robustness(Zhu, B. et al., 2024), general framework(Angelopoulos, A. N. et al., 2023), etc

- Zhang, Y., & Bradic, J. (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, *109*(2), 387-403.
- Zhu, B., Ding, M., Jacobson, P., Wu, M., Zhan, W., Jordan, M., & Jiao, J. (2024). Doubly-Robust Self-Training. *Advances in Neural Information Processing Systems*, *36*.
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669-674.
- Tony Cai, T., & Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(2), 391-419.

Kernel Two-Sample Test

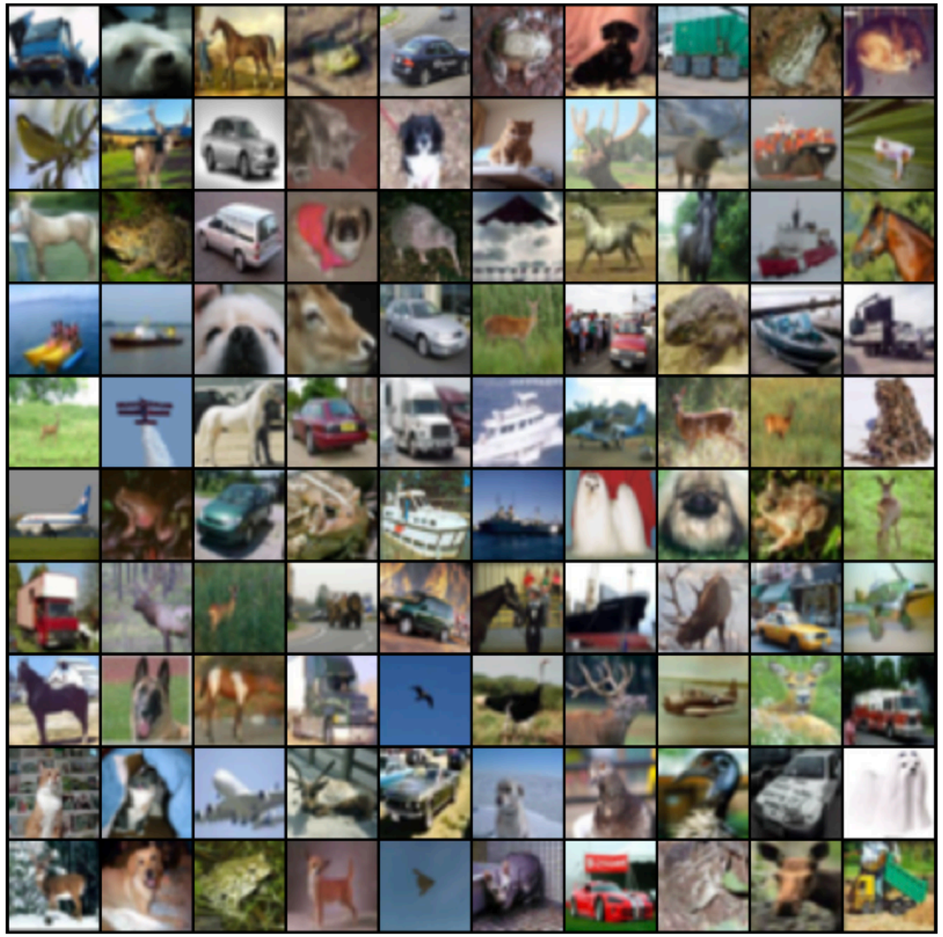
- Two-sample test: H_0: $\mu_1 - \mu_2 = d_0$, H_1: $\mu_1 - \mu_2 \neq d_0$,

$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

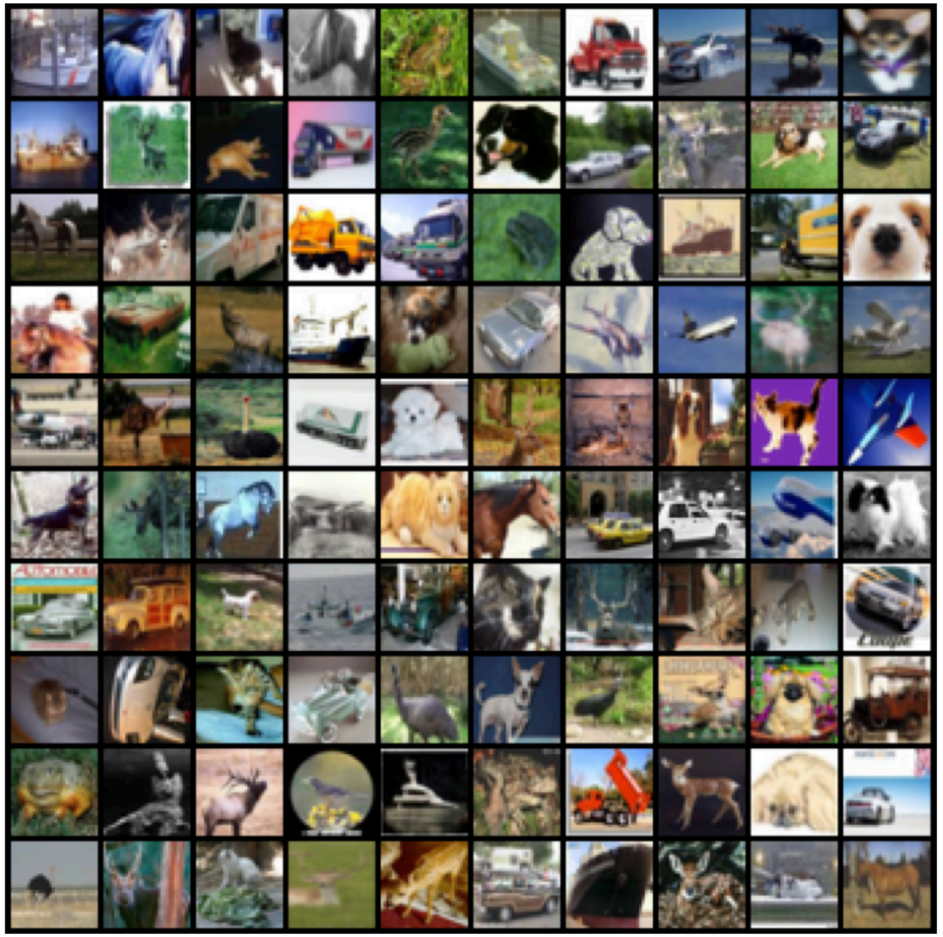
- Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). *Probability and statistics for engineers and scientists* (Vol. 5, pp. 326-332). New York: Macmillan.

Kernel Two-Sample Test

- Two-sample test: $H_0 : P_X = P_Y \quad \text{vs.} \quad H_1 : P_X \neq P_Y$



(a) *CIFAR-10* test set



(b) *CIFAR-10.1* test set

X_1: Now disturbing reports out of Newfoundland show that the fragile snow crab industry is in serious decline. First the west coast salmon, the east coast salmon and the cod, and now the snow crabs off Newfoundland.

X_2: To my pleasant surprise he responded that he had personally visited those wharves and that he had already announced money to fix them. What wharves did the minister visit in my riding and how much additional funding is he going to provide for Delaps Cove, Hampton, Port Lorne,

. . .

$$P_X \overset{?}{=} Q_Y$$

Y_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

Y_2: On the grain transportation system we have had the Estey report and the Kroeger report. We could go on and on. Recently programs have been announced over and over by the government such as money for the disaster in agriculture on the prairies and across Canada.

. . .

- Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., & Sutherland, D. J. (2020, November). Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning* (pp. 6316-6326). PMLR.
- http://www.gatsby.ucl.ac.uk/~gretton/papers/columbia23/columbia23_21

Kernel Two-Sample Test

- f-divergence

$$D_f(p||q) = \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}$$

Ex) $D_{\mathbb{KL}}(p \parallel q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$

$$D_\alpha^A(p||q) \triangleq \frac{4}{1-\alpha^2}\left(1 - \int p(\boldsymbol{x})^{(1+\alpha)/2} q(\boldsymbol{x})^{(1-\alpha)/2} d\boldsymbol{x}\right)$$

$$D_H^2(p||q) \triangleq \frac{1}{2}\int \left(p(\boldsymbol{x})^{\frac{1}{2}} - q(\boldsymbol{x})^{\frac{1}{2}}\right)^2 d\boldsymbol{x} = 1 - \int \sqrt{p(\boldsymbol{x})q(\boldsymbol{x})} d\boldsymbol{x}$$

$$\chi^2(p,q) \triangleq \frac{1}{2}\int \frac{(q(\boldsymbol{x}) - p(\boldsymbol{x}))^2}{q(\boldsymbol{x})} d\boldsymbol{x}$$

Computing divergence in terms of *ratio*

- Integral probability metrics

$$D_{\mathcal{F}}(P,Q) \triangleq \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{p(\boldsymbol{x})}\left[f(\boldsymbol{x})\right] - \mathbb{E}_{q(\boldsymbol{x}')}\left[f(\boldsymbol{x}')\right] \right|$$

Ex) $\mathcal{F} = \{||f||_L \leq 1\}, \quad ||f||_L = \sup_{\boldsymbol{x} \neq \boldsymbol{x}'} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{x}')|}{||\boldsymbol{x} - \boldsymbol{x}'||}$

⇒ Wasserstein-1 distance

$$W_1(P,Q) \triangleq \sup_{||f||_L \leq 1} \left| \mathbb{E}_{p(\boldsymbol{x})}\left[f(\boldsymbol{x})\right] - \mathbb{E}_{q(\boldsymbol{x}')}\left[f(\boldsymbol{x}')\right] \right|$$

Computing divergence in terms of *difference*

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

- Murphy, K. P. (2018). Machine learning: A probabilistic perspective (adaptive computation and machine learning series). *The MIT Press: London, UK*.

Kernel Two-Sample Test

• Reproducing Kernel Hilbert Space (RKHS)

: Hilbert space where the evaluation functional $L_x(f) = f(x)$ is bounded.

Riesz Representation Theorem

: For every bounded linear functional L on a Hilbert space $\mathcal{H}$, there exists a unique $\xi_L \in \mathcal{H}$ such that $L(f) = (\xi_L, f), \forall f \in \mathcal{H}$.

Or, we say there exists **kernel function** $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which satisfies reproducing property $\langle f(\cdot), \mathcal{K}(\cdot, \boldsymbol{x}') \rangle = f(\boldsymbol{x}')$ ($\langle \mathcal{K}(\boldsymbol{x}, \cdot), \mathcal{K}(\cdot, \boldsymbol{x}') \rangle = \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}')$)

• Some good properties

: Reproducing kernel is positive definite ($\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$ for every n, and every $x_1, \ldots, x_n \in \mathcal{X}$, and every $a_1, \ldots, a_n \in \mathbb{R}$)

: (Moore-Aronszajn) Suppose K is a symmetric p.d. kernel. There is a unique RKHS for which K is a reproducing kernel.

- Yongho, Jeon (2023), Lecture notes for nonparametric function estimation
- Murphy, K. P. (2018). Machine learning: A probabilistic perspective (adaptive computation and machine learning series). *The MIT Press: London, UK*.
- 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.

Kernel Two-Sample Test

Feature map $\varphi(x) \in \mathcal{F}$, $\varphi(x) = [\dots \varphi_i(x) \dots] \in \ell_2$

For positive definite k, $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$

• Expectiations of RKHS functions

Given P a Borel probability measure on $\mathcal{X}$, we define feature map of probability P, or mean embedding of P as

$$\mu_P = [\dots \mathbf{E}_P [\varphi_i(X)] \dots] \quad \text{which satisfies } \mathbf{E}_P(f(X)) = \langle f, \mu_P \rangle_{\mathcal{F}}$$

Then, for positive definite $k(x, x')$,

$$\langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathbf{E}_{P,Q} k(x, y)$$

Does it exist? Yes!

$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}}$

$$\text{If } \mathbf{E}_P \sqrt{k(x, x)} = \mathbf{E}_P \|\varphi(x)\|_{\mathcal{F}} < \infty \text{ then } \exists \mu_P \in \mathcal{F}.$$

Then we obtain its feature map as $\mu_P(t) = \langle \mu_P, \varphi(t) \rangle_{\mathcal{F}} = \langle \mu_P, k(\cdot, t) \rangle_{\mathcal{F}} = \mathbf{E}_{x \sim P} k(x, t)$

And MMD is the distance between feature maps

$$\begin{aligned} MMD^2(P, Q) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\ &= \underbrace{\mathbf{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(y, y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{P,Q} k(x, y)}_{\text{(b)}} \end{aligned}$$

- http://www.gatsby.ucl.ac.uk/~gretton/papers/columbia23/columbia23_21

Kernel Two-Sample Test

- Two-sample test using MMD

We obtain witness functions as

$$f^*(t) \quad \propto \quad \langle \phi(t), \mu_p - \mu_q \rangle_{\mathcal{H}} \quad = \quad \mathbf{E}_x [k(x,t)] - \mathbf{E}_y [k(y,t)],$$

So we take empirical witness function

$$\hat{f}^*(t) \quad \propto \quad \langle \phi(t), \mu_X - \mu_Y \rangle_{\mathcal{H}} \quad = \quad \frac{1}{m} \sum_{i=1}^{m} k(x_i,t) - \frac{1}{n} \sum_{i=1}^{n} k(y_i,t).$$

using empirical mean embedding as $\mu_X := \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$ and $\mu_Y := \frac{1}{n} \sum_{i=1}^{n} \phi(y_i)$

- 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.

Kernel Two-Sample Test

- Two-sample test using MMD

We use unbiased empirical estimator of MMD as

$$\mathrm{MMD}_u^2[\mathcal{F},X,Y] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}k(x_i,x_j)+\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j).$$

with asymptotic distribution under the null as

$$t\mathrm{MMD}_u^2[\mathcal{F},X,Y] \underset{D}{\to} \sum_{l=1}^{\infty}\lambda_l\left[(\rho_x^{-1/2}a_l-\rho_y^{-1/2}b_l)^2-(\rho_x\rho_y)^{-1}\right],\ a_l\sim\mathcal{N}(0,1)\ \textit{and}\ b_l\sim\mathcal{N}(0,1)\ \ \lambda_i\ \textit{are eigenvalues of}$$

$$\int_{\mathcal{X}}\tilde{k}(x,x')\psi_i(x)dp(x)=\lambda_i\psi_i(x').$$

and under the alternative as

$$m^{\frac{1}{2}}\left(\mathrm{MMD}_u^2-\mathrm{MMD}^2\left[\mathcal{F},p,q\right]\right)\xrightarrow{D}\mathcal{N}\left(0,\sigma_u^2\right)\ \text{ where }\sigma_u^2=4\left(\mathbf{E}_z\left[(\mathbf{E}_{z'}h(z,z'))^2\right]-\left[\mathbf{E}_{z,z'}(h(z,z'))\right]^2\right)$$

$\Rightarrow$ degenerate U-statistic under the null (intractable limiting distribution!)

So, we use $(1-\alpha)$ quantile using permutation test.

- 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.

Kernel Two-Sample Test

- Permutation-free kernel two-sample test

Instead, we use sample-splitting and studentization to handle this issue.

Original: $\widehat{\mathrm{MMD}}^2 := \frac{1}{n(n-1)m(m-1)} \sum_{1\le i\neq i'\le n} \sum_{1\le j\neq j'\le m} h(X_i, X_{i'}, Y_j, Y_{j'}), h(x, x', y, y') := k(x, x') - k(x, y') - k(y, x') + k(y, y')$

Permutation-free: $\widehat{\mathrm{xMMD}}^2 := \widehat{\mathrm{xMMD}}^2 / \widehat{\sigma}$

where $\widehat{\mathrm{xMMD}}^2 := \frac{1}{n_1 m_1 n_2 m_2} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_2} \sum_{j=1}^{m_1} \sum_{j'=1}^{m_2} h(X_i, X_{i'}, Y_j, Y_{j'}) = \frac{1}{n_1} \sum_{i=1}^{n_1} U_{X,i} - \frac{1}{m_1} \sum_{j=1}^{m_1} U_{Y,j}.$

$$\widehat{\sigma}_X^2 := \frac{1}{n_1} \sum_{i=1}^{n_1} \left(U_{X,i} - \bar{U}_X\right)^2, \ \widehat{\sigma}_Y^2 := \frac{1}{m_1} \sum_{j=1}^{m_1} \left(U_{Y,j} - \bar{U}_Y\right)^2 \text{ and } \widehat{\sigma}^2 := \frac{1}{n_1}\widehat{\sigma}_X^2 + \frac{1}{m_1}\widehat{\sigma}_Y^2.$$

$$U_{X,i} := \langle k(X_i, \cdot), \ \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k \quad U_{Y,j} := \langle k(Y_j, \cdot), \ \widehat{\mu}_2 - \widehat{\nu}_2 \rangle_k$$

- Shekhar, S., Kim, I., & Ramdas, A. (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, *35*, 18168-18180.

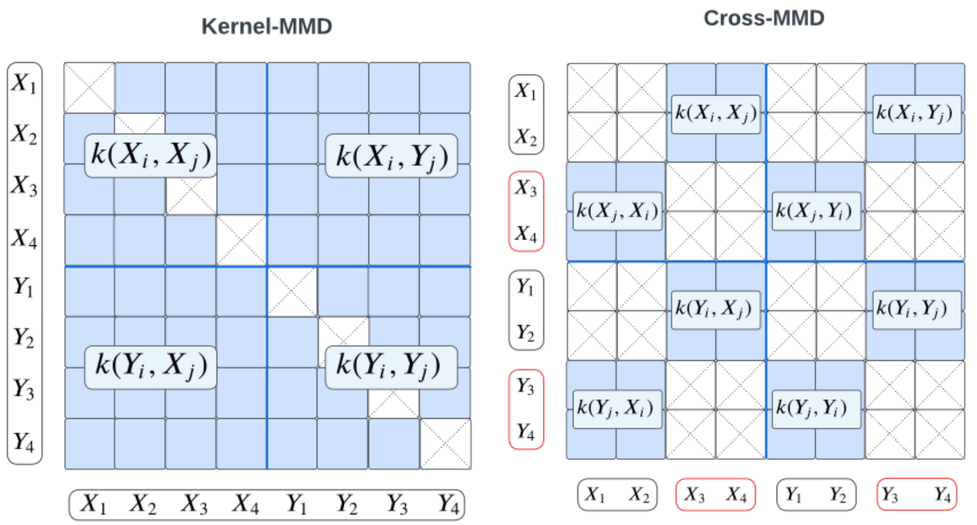
Kernel Two-Sample Test

• Permutation-free kernel two-sample test

- Asymptotic normality under the null

: If $\lim_{n\to\infty} \frac{\mathbb{E}_P[\bar{k}_n(X_1,X_2)^4]}{\mathbb{E}_P[\bar{k}_n(X_1,X_2)^2]^2}\left(\frac{1}{n}+\frac{1}{m_n}\right)=0, \quad and \quad \lim_{n\to\infty}\frac{\lambda_{1,n}^2}{\sum_{l=1}^{\infty}\lambda_{l,n}^2}\ exists,$

then $\bar{\mathrm{x}}\widehat{\mathrm{MMD}}^2 \xrightarrow{d} N(0,1).$



- Power consistency

: If $0 < \mathbb{E}_P[\bar{k}(X_1,X_2)^4] < \infty,\ and\ 0 < \mathbb{E}_Q[\bar{k}(Y_1,Y_2)^4] < \infty$, then it as asymptotic power of 1.

- Minimax rate optimality: $\lim_{n\to\infty}\sup_{P_n\in\mathcal{P}_n^{(0)}} \mathbb{E}_{P_n}[\Psi(\mathbb{X},\mathbb{Y})] \le \alpha \quad and \quad \lim_{n\to\infty}\inf_{(P_n,Q_n)\in\mathcal{P}_n^{(1)}} \mathbb{E}_{P_n,Q_n}[\Psi(\mathbb{X},\mathbb{Y})] = 1.$

with $s_n \asymp n^{4/(d+4\beta)}$, gaussian kernel and both densities lie in β-Sobolev ball.

- Shekhar, S., Kim, I., & Ramdas, A. (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, *35*, 18168-18180.

Semi-Supervised Kernel Two-Sample Test

- Setting

Now we have samples of both labeled and unlabeled data from two distributions

$$\left.\begin{aligned} &\mathcal{D}_{XV,1} := \{(X_i, V_i)\}_{i=1}^{n_1} := \{(X_i, V_i)\}_{i\in\mathcal{I}_{XV,1}} \overset{\text{iid}}{\sim} P_{XV}, \\ &\mathcal{D}_V := \{V_i\}_{i=n_1+1}^{n_1+m_1} := \{V_i\}_{i\in\mathcal{I}_V} \overset{\text{iid}}{\sim} P_V, \end{aligned}\right\} \ \mathcal{D}_{XV,1} \cup \mathcal{D}_V \text{ of size } n_1 + m_1$$

$$\left.\begin{aligned} &\mathcal{D}_{YW,1} := \{(Y_i, W_i)\}_{i=1}^{n_2} := \{(Y_i, W_i)\}_{i\in\mathcal{I}_{YW,1}} \overset{\text{iid}}{\sim} P_{YW}, \\ &\mathcal{D}_W := \{W_i\}_{i=n_2+1}^{n_2+m_2} := \{W_i\}_{i\in\mathcal{I}_W} \overset{\text{iid}}{\sim} P_W. \end{aligned}\right\} \ \mathcal{D}_{YW,1} \cup \mathcal{D}_W \text{ of size } n_2 + m_2$$

And our goal is to test

$$H_0 : P_X = P_Y \quad \text{vs.} \quad H_1 : P_X \neq P_Y$$

Semi-Supervised Kernel Two-Sample Test

- Idea

First, estimate the witness function

$$\widehat{f}(\cdot) = \frac{1}{n_1} \sum_{i \in \mathcal{I}_{XV,2}} k(X_i, \cdot) - \frac{1}{n_2} \sum_{i \in \mathcal{I}_{YW,2}} k(Y_i, \cdot)$$

From $\widehat{f}(X_1), \ldots, \widehat{f}(X_{n_1})$ and $\widehat{f}(Y_1), \ldots, \widehat{f}(Y_{n_2})$, we compute the oracle estimator

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

$$\widehat{\mu}_{Y,\widehat{f}} = \frac{1}{n_2} \sum_{i=1}^{n_2} \{\widehat{f}(Y_i) - \mathbb{E}[\widehat{f}(Y_i) \,|\, W_i, \widehat{f}]\} + \frac{1}{n_2 + m_2} \sum_{i=1}^{n_2+m_2} \mathbb{E}[\widehat{f}(Y_i) \,|\, W_i, \widehat{f}].$$

Then, we may estimate conditional expectations further.

Semi-Supervised Kernel Two-Sample Test

- Idea

- $\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}}$: unbiased estimator of MMD

- Since $\mathrm{Var}(\widehat{\mu}_{X,\widehat{f}} \mid \widehat{f}) := \sigma^2_{X,\widehat{f}} := \frac{1}{n_1} \underbrace{\mathbb{E}[\mathrm{Var}\{\widehat{f}(X) \mid V, \widehat{f}\} \mid \widehat{f}]}_{=\sigma^2_{1,X}} + \frac{1}{n_1+m_1} \underbrace{\mathrm{Var}[\mathbb{E}\{\widehat{f}(X) \mid V, \widehat{f}\} \mid \widehat{f}]}_{=\sigma^2_{2,X}}$

we estimate this as

$\widehat{\sigma}^2_{X,\widehat{f}} = \frac{1}{n_1}\widehat{\sigma}^2_{1,X} + \frac{1}{n_1+m_1}\widehat{\sigma}^2_{2,X}$ where $\widehat{\sigma}^2_{1,X} = \frac{1}{n_1}\sum_{i=1}^{n_1}\{\widehat{f}(X_i) - \mathbb{E}[\widehat{f}(X_i) \mid V_i, \widehat{f}]\}^2$ and

$$\widehat{\sigma}^2_{2,X} = \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\left\{\mathbb{E}[\widehat{f}(X_i) \mid V_i, \widehat{f}] - \frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\mathbb{E}[\widehat{f}(X_j) \mid V_j, \widehat{f}]\right\}^2$$

Then, our oracle test statistic becomes $T^\star = \frac{\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}}}{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}}$ where we reject the null if $T^\star > z_{1-\alpha} := \Phi^{-1}(1-\alpha)$

Semi-Supervised Kernel Two-Sample Test

- Asymptotic normality under the null

1) Test statistic with known variance $T_\sigma^\star = \dfrac{\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}}}{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}}$

Write numerator as $\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}} = \sum_{i=1}^{n_1+m_1} G_i - \sum_{i=1}^{n_2+m_2} H_i$ where

$$G_i = \begin{cases} \frac{1}{n_1}\{\widehat{f}(X_i) - \mathbb{E}[\widehat{f}(X_i) \mid \widehat{f}]\} - \frac{m_1}{n_1(n_1+m_1)}\{\mathbb{E}[\widehat{f}(X_i) \mid V_i, \widehat{f}] - \mathbb{E}[\widehat{f}(X_i) \mid \widehat{f}]\} & \text{if } 1 \le i \le n_1, \\ \frac{1}{n_1+m_1}\{\mathbb{E}[\widehat{f}(X_i) \mid V_i, \widehat{f}] - \mathbb{E}[\widehat{f}(X_i) \mid \widehat{f}]\} & \text{if } n_1+1 \le i \le n_1+m_1, \end{cases}$$

$$H_i = \begin{cases} \frac{1}{n_2}\{\widehat{f}(Y_i) - \mathbb{E}[\widehat{f}(Y_i) \mid \widehat{f}]\} - \frac{m_2}{n_2(n_2+m_2)}\{\mathbb{E}[\widehat{f}(Y_i) \mid W_i, \widehat{f}] - \mathbb{E}[\widehat{f}(Y_i) \mid \widehat{f}]\} & \text{if } 1 \le i \le n_2, \\ \frac{1}{n_2+m_2}\{\mathbb{E}[\widehat{f}(Y_i) \mid W_i, \widehat{f}] - \mathbb{E}[\widehat{f}(Y_i) \mid \widehat{f}]\} & \text{if } n_2+1 \le i \le n_2+m_2. \end{cases}$$

Define denominator as $s^2_{n+m} := \mathrm{Var}(\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}} \mid \widehat{f})$

$$= \sum_{i=1}^{n_1+m_1} \mathrm{Var}(G_i \mid \widehat{f}) + \sum_{i=1}^{n_2+m_2} \mathrm{Var}(H_i \mid \widehat{f})$$

$$= \frac{1}{n_1}\sigma^2_{1,X} + \frac{1}{n_1+m_1}\sigma^2_{2,X} + \frac{1}{n_2}\sigma^2_{1,Y} + \frac{1}{n_2+m_2}\sigma^2_{2,Y}$$

Semi-Supervised Kernel Two-Sample Test

- Asymptotic normality

To use Lyapunov's CLT, we prove $\sum_{i=1}^{n_1+m_1} \mathbb{E}\left[\left|\frac{G_i}{s_{n+m}}\right|^{2+\delta} \middle| \widehat{f}\right] + \sum_{i=1}^{n_2+m_2} \mathbb{E}\left[\left|\frac{H_i}{s_{n+m}}\right|^{2+\delta} \middle| \widehat{f}\right] \xrightarrow{p} 0$

with additional assumptions like

$$\frac{\mathbb{E}[\overline{k}^4(X_1,X_2)]\frac{n_1+n_2}{n_1n_2} + \mathbb{E}[\overline{k}^2(X_1,X_3)\overline{k}^2(X_2,X_3)]}{\left\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\right\}^2\left(\frac{n_1n_2}{n_1+n_2}\right)} \to 0 \qquad \inf_{\{\alpha_i\}_{i=1}^{\infty}} \left[(1-c)\sum_{i=1}^{\infty}\lambda_i^2\alpha_i^2 - \sum_{i,i'=1}^{\infty}\lambda_i\lambda_{i'}\alpha_i\alpha_{i'}\rho_{i,i'}\right] \geq 0$$

for centered kernel $\overline{k}(x,y) = k(x,y) - \mathbb{E}[k(X,Y)\,|\,X=x] - \mathbb{E}[k(X,Y)\,|\,Y=y] + \mathbb{E}[k(X,Y)]$

$$= \sum_{i=1}^{\infty}\lambda_i\phi_i(x)\phi_i(y)$$

and $\rho_{i,i'} = \mathrm{Cov}\{\mathbb{E}[\phi_i(X)\,|\,V], \mathbb{E}[\phi_{i'}(X)\,|\,V]\}$

Semi-Supervised Kernel Two-Sample Test

- Asymptotic normality

If $\inf_{\{\alpha_i\}_{i=1}^{\infty}} \left[(1-c) \sum_{i=1}^{\infty} \lambda_i^2 \alpha_i^2 - \sum_{i,i'=1}^{\infty} \lambda_i \lambda_{i'} \alpha_i \alpha_{i'} \rho_{i,i'} \right] \geq 0$

Then, $\sigma_{1,X}^2 = \mathbb{E}\left[\left\{ \sum_{i=1}^{\infty} \lambda_i \left(\underbrace{\frac{1}{n_1} \sum_{j=1}^{n_1} \phi_i(X_j)}_{=\overline{\phi}_{i,X}} - \underbrace{\frac{1}{n_2} \sum_{j=1}^{n_2} \phi_i(Y_j)}_{=\overline{\phi}_{i,Y}} \right) \left(\phi_i(X) - \mathbb{E}[\phi_i(X) \mid V] \right) \right\}^2 \mid \mathcal{D}_{XV,2}, \mathcal{D}_{YW,2} \right]$

$$\geq c \sum_{i=1}^{\infty} \lambda_i^2 (\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})^2,$$

which is used for proof.

Semi-Supervised Kernel Two-Sample Test

- Asymptotic normality

2) Test statistic with unknown variance

Using Slutsky's theorem, it suffices to show that $\frac{\widehat{\sigma}^2_{X,\widehat{f}} - \sigma^2_{X,\widehat{f}}}{\sigma^2_{X,\widehat{f}}} \xrightarrow{p} 0$ and $\frac{\widehat{\sigma}^2_{Y,\widehat{f}} - \sigma^2_{Y,\widehat{f}}}{\sigma^2_{Y,\widehat{f}}} \xrightarrow{p} 0$

which implies that $\frac{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}} \xrightarrow{p} 1 \iff \frac{\widehat{\sigma}^2_{X,\widehat{f}} - \sigma^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}} - \sigma^2_{Y,\widehat{f}}}{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}} \xrightarrow{p} 0$

We compute $\frac{\left|\widehat{\sigma}^2_{X,\widehat{f}} - \sigma^2_{X,\widehat{f}}\right|}{\sigma^2_{X,\widehat{f}}} \leq \underbrace{\frac{\left|\widehat{\sigma}^2_{1,X} - \sigma^2_{1,X}\right|}{\sigma^2_{1,X}}}_{\text{(I)}} + \underbrace{\frac{n_1}{n_1 + m_1} \frac{\left|\widehat{\sigma}^2_{2,X} - \sigma^2_{2,X}\right|}{\sigma^2_{1,X}}}_{\text{(II)}}$

and both terms are $o_P(1)$ which completes the proof.

Semi-Supervised Kernel Two-Sample Test

- Asymptotic normality

3) Test statistic using cross-fitting

We estimate $\mathbb{E}[\widehat{f}(X_i) \mid V_i, \widehat{f}]$ as $\widehat{\mathbb{E}}[\widehat{f}(X_i) \mid V_i, \widehat{f}]$ trained based on $\mathcal{D}_{XV,1,b}$ if $i \in \{1, 2, \ldots, n_1/2\} \cup \{n_1 + 1, \ldots, n_1 + m_1/2\}$ and on $\mathcal{D}_{XV,1,a}$ if $i \in \{n_1/2 + 1, n_1/2 + 2, \ldots, n_1\} \cup \{n_1 + m_1/2 + 1, \ldots, n_1 + m_1\}$.

Then we prove $T - T^\star = o_P(1)$ assuming $\sup_{i \geq 1} \mathbb{E}[\Delta_i^2] = o(1)$ for $\Delta_i = \mathbb{E}[\phi_i(X) \mid V] - \widehat{\mathbb{E}}[\phi_i(X) \mid V]$

$$
\begin{aligned}
T - T^\star &= \frac{\widehat{N}}{\widehat{D}} - \frac{\widehat{N}}{D} + \frac{\widehat{N}}{D} - \frac{N}{D} \\
&= \frac{\widehat{N} - N}{D}\left(\frac{D}{\widehat{D}} - 1\right) + \underbrace{\frac{N}{D}}_{= O_P(1)}\left(\frac{D}{\widehat{D}} - 1\right) + \frac{\widehat{N} - N}{D}.
\end{aligned}
$$

Semi-Supervised Kernel Two-Sample Test

- Power analysis

Define our test function $\phi^\star := \mathbb{1}_{\{T^\star > z_{1-\alpha}\}}$

Then we express its power as

$$\begin{aligned}
\mathbb{P}(T^\star > z_{1-\alpha}) &= \mathbb{P}\left(\frac{\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}}}{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}} > z_{1-\alpha}\right) \\
&= \mathbb{P}\left(\frac{(\widehat{\mu}_{X,\widehat{f}} - \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\mathbb{E}[\widehat{f}(X_i)|\widehat{f}]) - (\widehat{\mu}_{Y,\widehat{f}} - \frac{1}{n_2+m_2}\sum_{i=1}^{n_2+m_2}\mathbb{E}[\widehat{f}(Y_i)|\widehat{f}])}{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}}\right. \\
&\qquad \left. > z_{1-\alpha} - \frac{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\mathbb{E}[\widehat{f}(X_i)|\widehat{f}] - \frac{1}{n_2+m_2}\sum_{i=1}^{n_2+m_2}\mathbb{E}[\widehat{f}(Y_i)|\widehat{f}]}{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}}\right)
\end{aligned}$$

Semi-Supervised Kernel Two-Sample Test

- Asymptotic power expression with linear kernel

Assuming normality and denoting $\Sigma = n_1^{-1}\Sigma_{11} + n_2^{-1}\tilde{\Sigma}_{11}$, $\Lambda = \frac{1-\rho_{XV}^2}{n_1}\Sigma_{22} + \frac{\rho_{XV}^2}{n_1}\Sigma_{11}$, and

$$\tilde{\Lambda} = \frac{1-\rho_{YW}^2}{n_2}\tilde{\Sigma}_{22} + \frac{\rho_{YW}^2}{n_2}\tilde{\Sigma}_{11}$$

$$\begin{pmatrix} X \\ V \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_V \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

$$\begin{pmatrix} Y \\ W \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_W \end{pmatrix}, \begin{pmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{pmatrix}\right)$$

we guess the power function of

$$\mathbb{P}(T^\star > z_{1-\alpha}) = \Phi\Big(z_\alpha + \frac{(\mu_X - \mu_Y)^\top(\mu_X - \mu_Y)}{\sqrt{\mathrm{tr}((\Lambda + \tilde{\Lambda})\Sigma)}}\Big) + o_P(1)$$

$$= \Phi\Big(z_\alpha + \frac{(\mu_X - \mu_Y)^\top(\mu_X - \mu_Y)}{n_1^{-1}\sqrt{(1-\rho_{XV}^2)\mathrm{tr}(\Sigma_{22}\Sigma_{11}) + \rho_{XV}^2\mathrm{tr}(\Sigma_{11}^2)}}\Big) + o_P(1)$$

Semi-Supervised Kernel Two-Sample Test

- Power consistency (if we have asymptotic power of 1)

Prove $\mathbb{E}_{P,Q}[1-\phi^\star] = \mathbb{P}_{P,Q}(T^\star \le z_{1-\alpha}) = \mathbb{P}_{P,Q}(\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}} \le z_{1-\alpha}\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}})$ converges to 0 by showing

$$\frac{\mathbb{E}_{P,Q}[\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}]}{\gamma^4\delta_n} + \frac{\mathrm{Var}_{P,Q}(\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}})}{\gamma^4} \le \frac{\mathbb{E}_{P,Q}[\widehat{\sigma}^2]}{\gamma^4\delta_n} + \frac{\mathrm{Var}(\widehat{\mathrm{xMMD}})}{\gamma^4} \to 0$$

Also, using cross-fitted estimator, under the same assumption of $\sup_{i\ge 1}\mathbb{E}[\Delta_i^2] = o(1)$

for $\Delta_i = \mathbb{E}[\phi_i(X)\,|\,V] - \widehat{\mathbb{E}}[\phi_i(X)\,|\,V]$

$$\mathbb{E}_{P,Q}[\widehat{D}^2 - D^2] < \infty \quad \text{and} \quad \mathrm{Var}_{P,Q}(\widehat{N} - N) \to 0$$

Semi-Supervised Kernel Two-Sample Test

- Asymptotic normality under the alternative

We need additional assumptions like

$$\max\{\|p/q\|_\infty, \|q/p\|_\infty\} < C$$

$$\frac{\mathbb{E}[\overline{k}_X(X_1, X_2)^4] + n_1\mathbb{E}[\overline{k}_X(X_1, X_2)^2\overline{k}_X(X_1, X_3)^2]}{n_1^2\{\mathbb{E}[\overline{g}_X(X, X)]\}^2} \to 0$$

$$\frac{\mathrm{MMD}^4\mathbb{E}[\overline{k}_X(X, X)^2]}{\{n_1\mathbb{E}[\overline{g}_X(X, X)] + n_1^2\mathbb{E}[\overline{g}_X(Y_1, Y_2)]\}^2} \to 0,$$

Semi-Supervised Kernel Two-Sample Test

- Asymptotic normality under the alternative

Again, we use Lyapunov's CLT showing

$$\frac{1}{\sigma^4_{n_1,n_2}}\left[\frac{1}{n_1^4}\sum_{i=1}^{n_1}\mathbb{E}[\{\widehat{f}(X_i)-\mathbb{E}[\widehat{f}(X)\,|\,\widehat{f}]\}^4\,|\,\widehat{f}]+\frac{1}{n_2^4}\sum_{i=1}^{n_2}\mathbb{E}[\{\widehat{f}(Y_i)-\mathbb{E}[\widehat{f}(Y)\,|\,\widehat{f}]\}^4\,|\,\widehat{f}]\right]\xrightarrow{p}0$$

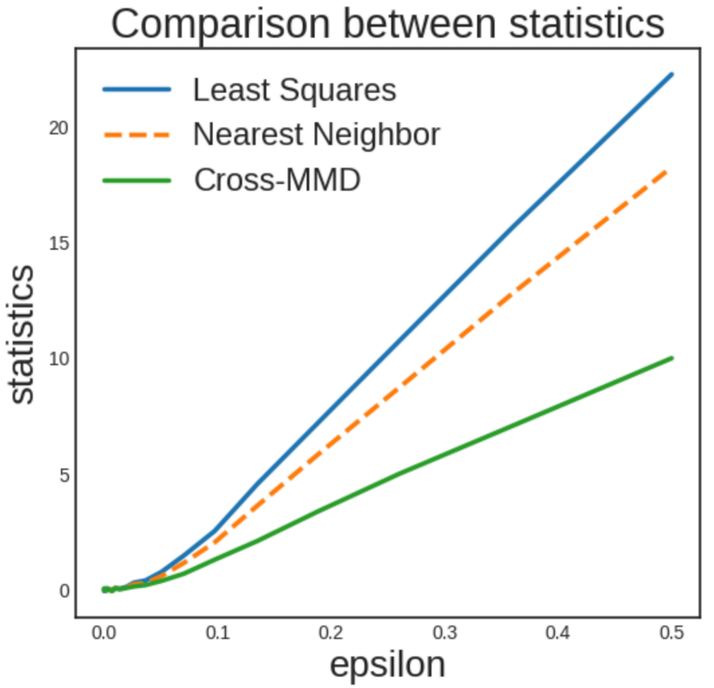
Since assuming the alternative,

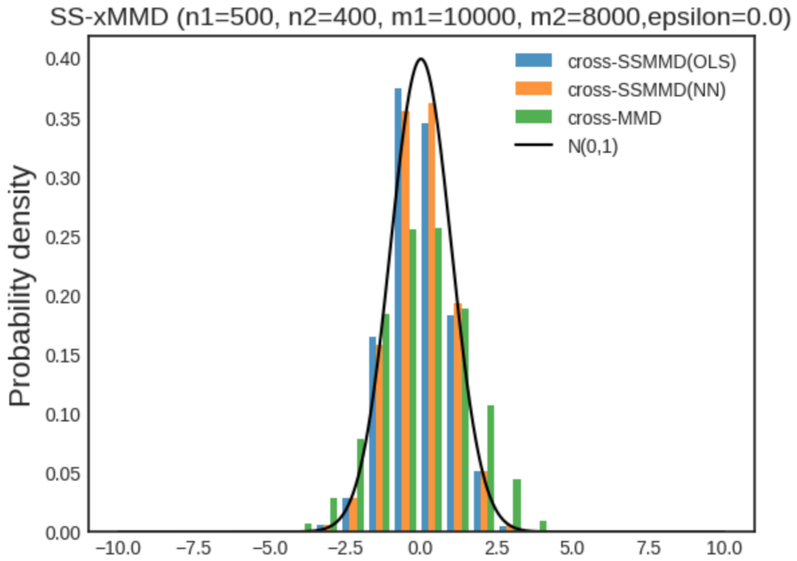
$$\begin{aligned}
\mathrm{Var}[\widehat{f}(X)\,|\,\widehat{f}] &= \mathbb{E}[\langle\bar{\psi}_X-\bar{\psi}_Y,\psi(X)-\mathbb{E}_P[\psi(X)]\rangle^2\,|\,\bar{\psi}_X,\bar{\psi}_Y]\\
&= \mathbb{E}[\langle\bar{\psi}_X-\mathbb{E}[\psi(X)]+\mathbb{E}[\psi(X)]-\bar{\psi}_Y,\psi(X)-\mathbb{E}_P[\psi(X)]\rangle^2\,|\,\bar{\psi}_X,\bar{\psi}_Y]\\
&= \mathbb{E}\left[\left\{\sum_{i=1}^{\infty}\lambda_i\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\phi_i(X_j')\right)\phi_i(X)-\sum_{i=1}^{\infty}\lambda_i\left(\frac{1}{n_2}\sum_{j=1}^{n_2}\phi_i(Y_j')\right)\phi_i(X)\right\}^2\,\middle|\,(X_j'),(Y_j')\right]\\
&= \mathbb{E}\left[\left\{\sum_{i=1}^{\infty}\lambda_i(\overline{\phi}_{i,X}-\overline{\phi}_{i,Y})\phi_i(X)\right\}^2\,\middle|\,(X_j'),(Y_j')\right]\\
&= \sum_{i=1}^{\infty}\lambda_i^2(\overline{\phi}_{i,X}-\overline{\phi}_{i,Y})^2.\ \text{and, similarly, } \mathrm{Var}[\widehat{f}(Y)\,|\,\widehat{f}]=\sum_{i=1}^{\infty}\mu_i^2(\overline{\varphi}_{i,X}-\overline{\varphi}_{i,Y})^2
\end{aligned}$$

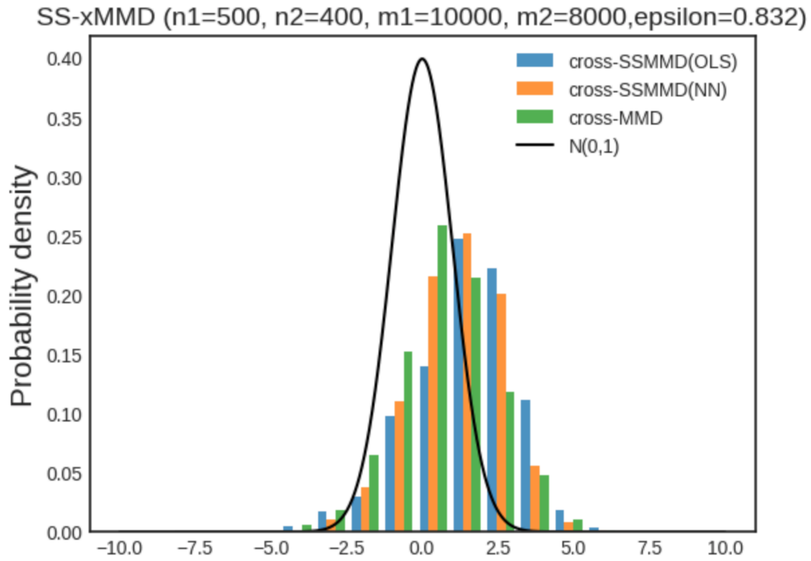
Semi-Supervised Kernel Two-Sample Test

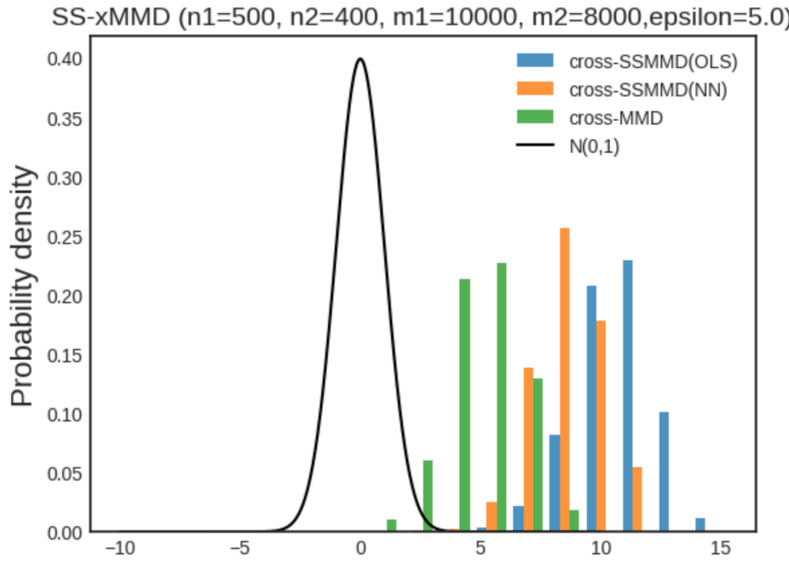
- Numerical Studies (d=4, RBF kernel)

$$X = V_{(1)} + V_{(2)} + 0.3Z$$
$$Y = W_{(1)} + W_{(2)} + 0.3Z$$
$$V \sim N_d(\mu_1, \Sigma_1), \quad \text{where} \quad \mu_1 = \mathbb{0}_d,\ \Sigma_1 = 0.3I_d + 0.7\mathbb{1}_d'\mathbb{1}_d$$
$$W \sim N_d(\mu_2, \Sigma_2), \quad \text{where} \quad \mu_2 = (\epsilon, ..., \epsilon)',\ \Sigma_2 = 0.3I_d + 0.7\mathbb{1}_d'\mathbb{1}_d$$





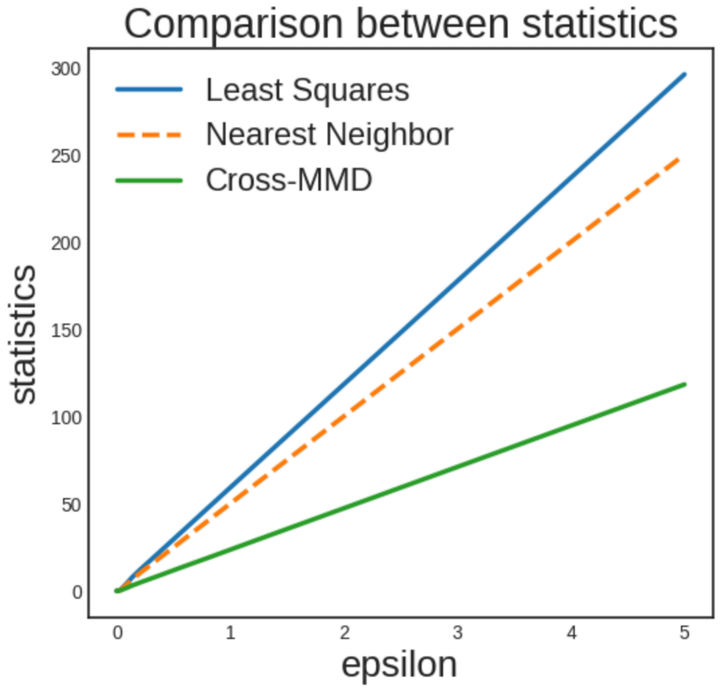


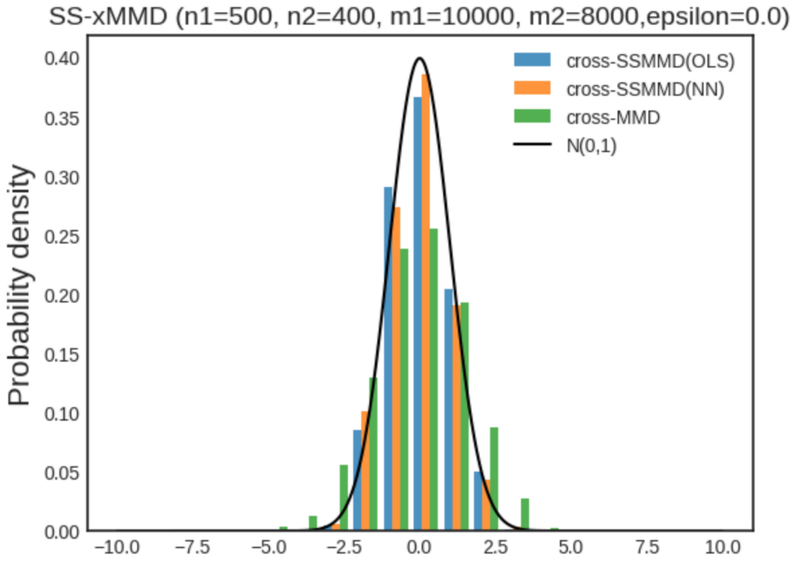


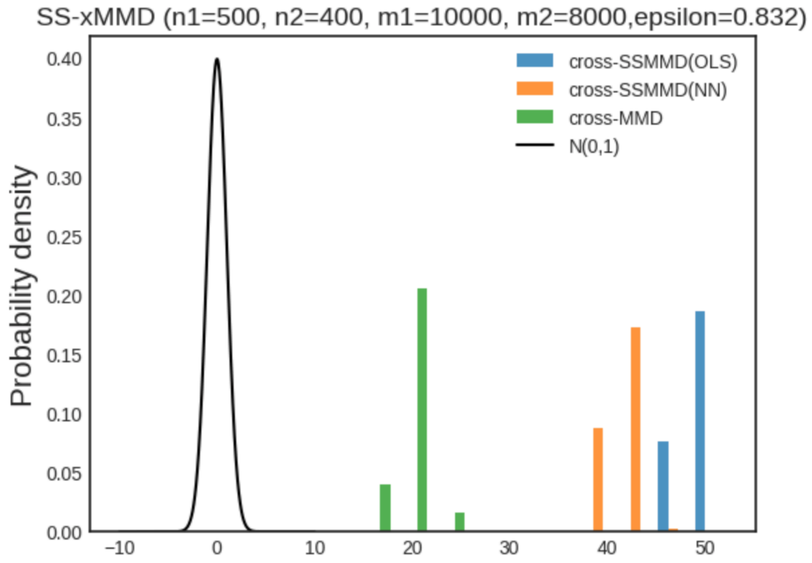
Semi-Supervised Kernel Two-Sample Test

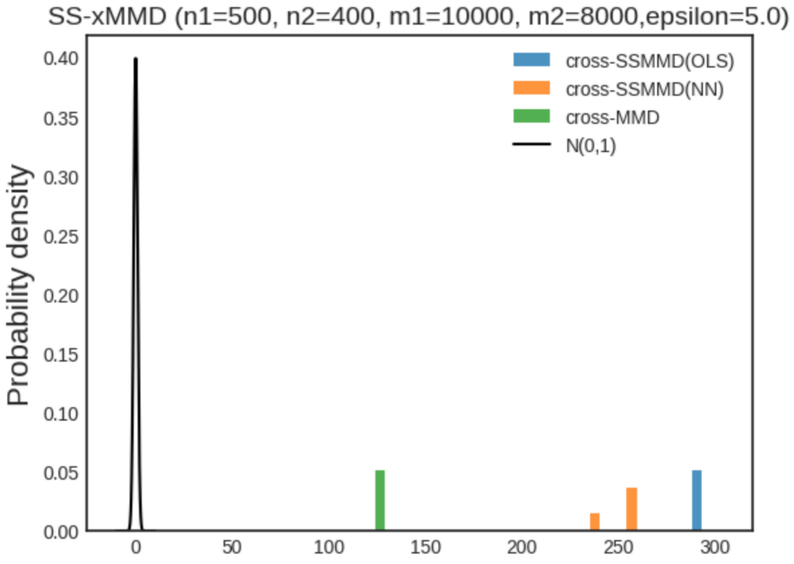
- Numerical Studies (d=4, linear kernel)

$$X = V_{(1)} + V_{(2)} + 0.3Z$$
$$Y = W_{(1)} + W_{(2)} + 0.3Z$$
$$V \sim N_d(\mu_1, \Sigma_1), \quad \text{where} \quad \mu_1 = \mathbb{0}_d,\ \Sigma_1 = 0.3I_d + 0.7\mathbb{1}_d'\mathbb{1}_d$$
$$W \sim N_d(\mu_2, \Sigma_2), \quad \text{where} \quad \mu_2 = (\epsilon, ..., \epsilon)',\ \Sigma_2 = 0.3I_d + 0.7\mathbb{1}_d'\mathbb{1}_d$$







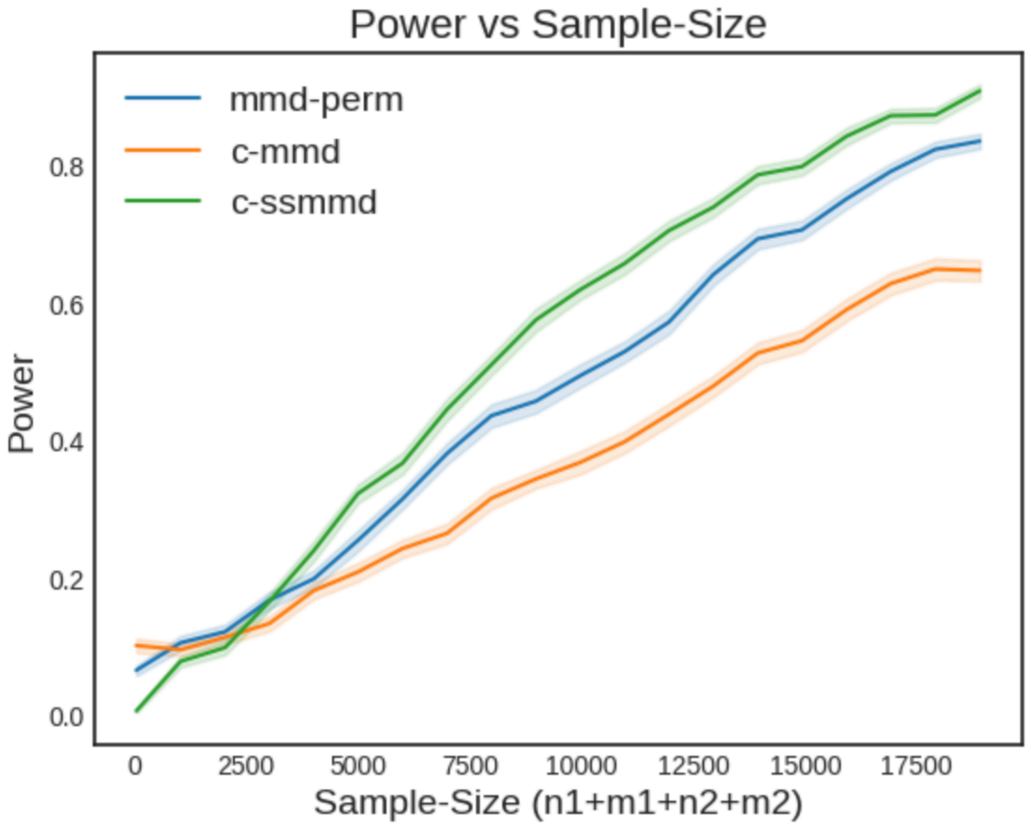


Semi-Supervised Kernel Two-Sample Test

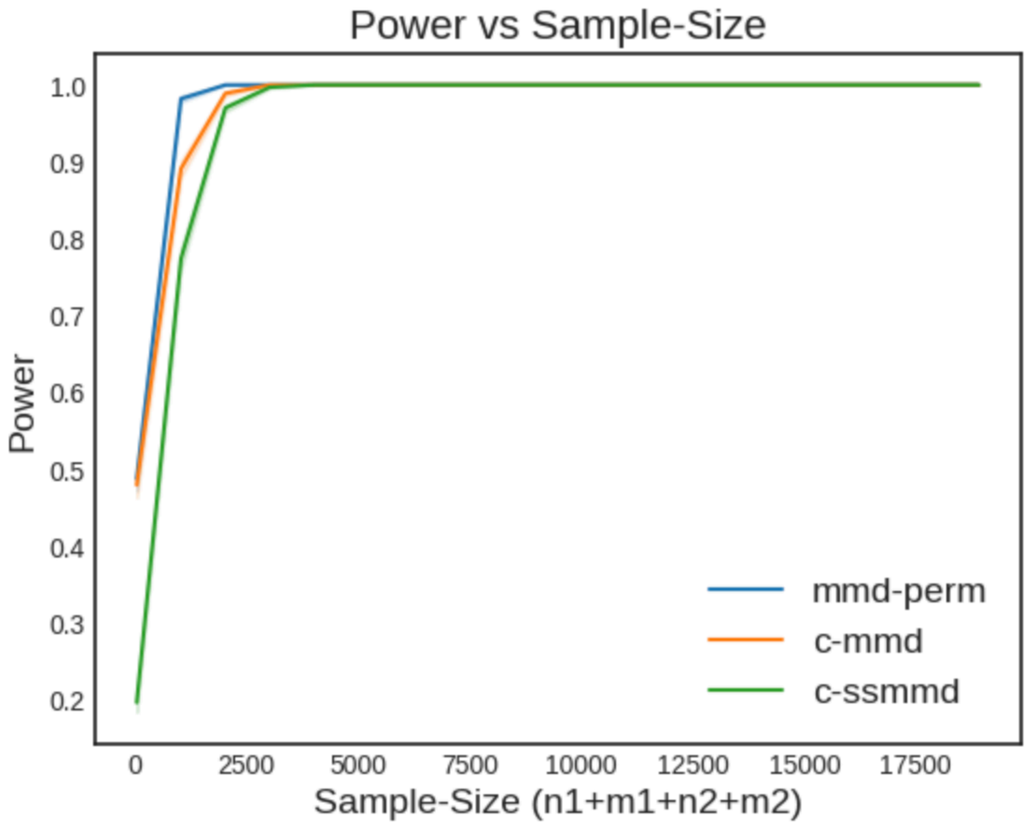
- Numerical Studies

$$X = V_{(1)} + V_{(2)} + V_{(d)} + 0.3Z \qquad V \sim N_d(\mu_1, \Sigma_1), \quad \text{where} \quad \mu_1 = \mathbb{0}_d,\ \Sigma_1 = I_d$$

$$Y = W_{(1)} + W_{(2)} + W_{(d)} + 0.3Z \qquad W \sim N_d(\mu_2, \Sigma_2), \quad \text{where} \quad \mu_2 = (\underbrace{\epsilon, ..., \epsilon}_{\text{p elements}}, 0, ..., 0)',\ \Sigma_2 = I_d$$



d=10,p=3,eps=0.3,n1=400,m1=8000,n2=500,m2=10000

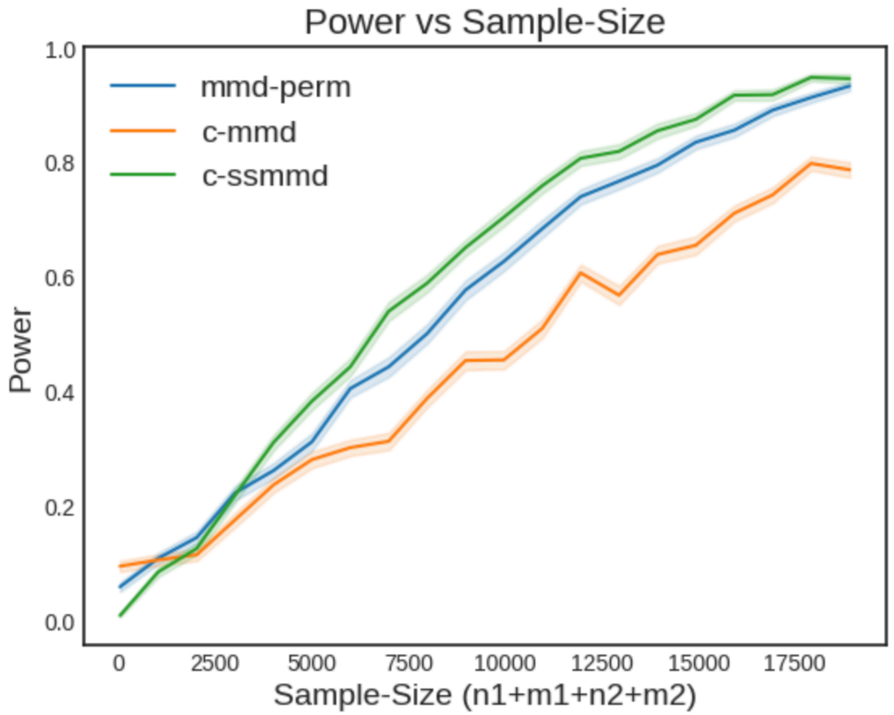


d=50,p=50,eps=1,n1=400,m1=8000,n2=500,m2=10000

Semi-Supervised Kernel Two-Sample Test

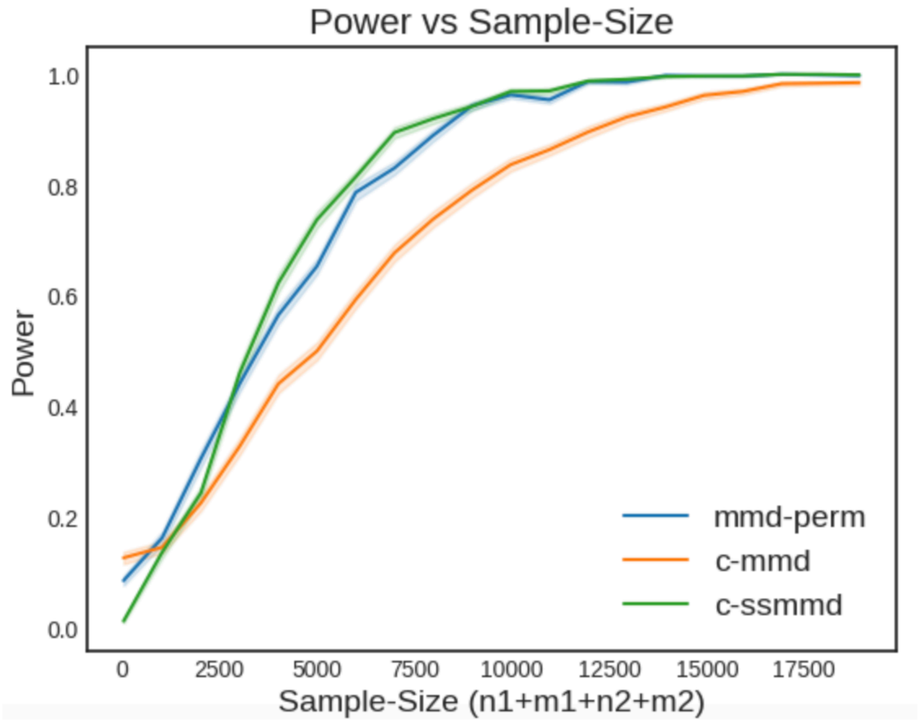
- Numerical Studies

$$X = V_{(1)} + V_{(2)} + V_{(d)} + 0.3Z$$
$$Y = W_{(1)} + W_{(2)} + W_{(d)} + 0.3Z$$
$$V \sim N_d(\mu_1, \Sigma_1), \quad \text{where} \quad \mu_1 = \mathbb{0}_d,\ \Sigma_1 = I_d$$
$$W \sim N_d(\mu_2, \Sigma_2), \quad \text{where} \quad \mu_2 = \mathbb{0}_d,\ \Sigma_2 = \text{diag}(\underbrace{\epsilon, ..., \epsilon}_{\text{p elements}}, 1, ..., 1)$$



d=10,p=3,eps=0.3,n1=400,m1=8000,n2=500,m2=10000

$$X = V_{(1)} + V_{(2)} + V_{(d)} + 0.3Z$$
$$Y = W_{(1)} + W_{(2)} + W_{(d)} + 0.3Z$$
$$V \sim N_d(\mu_1, \Sigma_1), \quad \text{where} \quad \mu_1 = \mathbb{0}_d,\ \Sigma_1 = I_d$$
$$W \sim N_d(\mu_2, \Sigma_2), \quad \text{where} \quad \mu_2 = (\underbrace{\epsilon, ..., \epsilon}_{\text{p elements}}, 0, ..., 0)',\ \Sigma_2 = \text{diag}(\underbrace{\epsilon, ..., \epsilon}_{\text{p elements}}, 1, ..., 1)$$



d=10,p=3,eps=0.3,n1=400,m1=8000,n2=500,m2=10000

Conclusion

- Proposed semi-supervised kernel two-sample test statistic using sample splitting, and cross-fitting.

- Proved asymptotic normality of the test statistic under the null and the alternative and figured out under which condition it shows the properties.

- Derived asymptotic power expression using linear kernel and showed power consistency with proper assumptions.

Future work

- What if not using cross-fitting? What if it is not linear?

- Are all the assumptions necessary?

- What if the distributions, so MMD are not fixed w.r.t. n?

Reference

- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669-674.
- 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.
- Han, J., Liang, X., Xu, H., Chen, K., Hong, L., Mao, J., ... & Xu, C. (2021). SODA10M: A large-scale 2D self/semi-supervised object detection dataset for autonomous driving. *arXiv preprint arXiv:2106.11118*.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
- Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395-405.
- Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., & Jin, C. (2023). Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 107840.
- Ilmun Kim, (2023), Lecture notes “Selective Topics in Mathematical Statistics” for graduate mathematical statistics course
- Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., & Sutherland, D. J. (2020, November). Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning* (pp. 6316-6326). PMLR.
- Murphy, K. P. (2018). Machine learning: A probabilistic perspective (adaptive computation and machine learning series). *The MIT Press: London, UK*.
- Shekhar, S., Kim, I., & Ramdas, A. (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, *35*, 18168-18180.
- Tony Cai, T., & Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(2), 391-419.
- Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). *Probability and statistics for engineers and scientists* (Vol. 5, pp. 326-332). New York: Macmillan.
- Yongho, Jeon (2023), Lecture notes for nonparametric function estimation
- Zhang, A., Brown, L. D., & Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means.
- Zhang, Y., & Bradic, J. (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, *109*(2), 387-403.
- Zhu, B., Ding, M., Jacobson, P., Wu, M., Zhan, W., Jordan, M., & Jiao, J. (2024). Doubly-Robust Self-Training. *Advances in Neural Information Processing Systems*, *36*.
- https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242
- http://www.gatsby.ucl.ac.uk/~gretton/papers/columbia23/columbia23_21