A Witness Two-Sample Test

Jonas M. Kübler[1], Wittawat Jitkrittum[2], Bernhard Schölkopf[1], Krikamol Muandet[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen [2]Google Research

EMPIRICAL INFERENCE
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Overview

- Two-Sample Test:
  - X, Y are random variables with distributions P and Q on $\mathcal{X} \subseteq \mathbb{R}^d$.
  - Test the null hypothesis $P = Q$ against the alternative $P \neq Q$ based on samples $\mathbb{X} = \{x_1, \dots, x_n\}$ and $\mathbb{Y} = \{y_1, \dots, y_m\}$.
  - Control type-I error (probability of rejecting H_0 when true) at specified level α.
  - Test power: Probability of rejecting H_0 when it is false. Should be maximized.
- Notation: $\mathbb{X}_{\text{tr}}, \mathbb{X}_{\text{te}}$ and $\mathbb{Y}_{\text{tr}}, \mathbb{Y}_{\text{te}}$ denote disjoint training and test sets with $n = n_{\text{tr}} + n_{\text{te}}$, $m = m_{\text{tr}} + m_{\text{te}}$. $\mathbb{Z} = \{\mathbb{X}, \mathbb{Y}\}$, $\mathbb{Z}_{\text{tr}} = \{\mathbb{X}_{\text{tr}}, \mathbb{Y}_{\text{tr}}\}$ and $\mathbb{Z}_{\text{te}} = \{\mathbb{X}_{\text{te}}, \mathbb{Y}_{\text{te}}\}$.
- **Witness Two-Sample Test (WiTS Test)**:

$$\hat{\tau}(\mathbb{Z}_{\text{te}}|h) = \frac{1}{n_{\text{te}}} \sum_{x \in \mathbb{X}_{\text{te}}} h(x) - \frac{1}{m_{\text{te}}} \sum_{y \in \mathbb{Y}_{\text{te}}} h(y). \tag{1}$$

  - Stage I: Optimize (2) to find witness $h : \mathcal{X} \to \mathbb{R}$ on training data $\mathbb{Z}_{\text{tr}}$.
  - Stage II: Compute (1) and test its significance on $\mathbb{Z}_{\text{te}}$.
    - ∗ Thresholds via asymptotic normality or permutations.
- **Advantages:**
  - Easy and intuitive theory for a principled objective for Stage I.
  - Approximation, model selection, cross-validation techniques in Stage I.
  - Witness can be learned via any ML framework.
  - Fast computation of thresholds in Stage II.

Properties of WiTS tests

Theorem 1 (Asymptotic normality of WiTS test)**.** *For a witness function $h : \mathcal{X} \to \mathbb{R}$, let $\sigma_P^2 :=$ Var$[h(X)]$ and $\sigma_Q^2 :=$ Var$[h(Y)]$ such that $0 < \sigma_P^2, \sigma_Q^2 < \infty$. Let $\{X_i\}_{i \in [n]} \overset{\text{i.i.d.}}{\sim} P$, $\{Y_j\}_{j \in [m]} \overset{\text{i.i.d.}}{\sim} Q$, and $c := \frac{n}{n+m} \in (0,1)$ as $n + m \to \infty$. Denote by $\bar{h}_P := \mathbb{E}[h(X)]$ and $\bar{h}_Q := \mathbb{E}[h(Y)]$. We define the empirical means $\hat{h}_P^n := \frac{1}{n}\sum_{i \in [n]} h(X_i)$, $\hat{h}_Q^m := \frac{1}{m}\sum_{i \in [m]} h(Y_i)$ and denote the sample variance as $\hat{\sigma}_c^2(h) := \hat{\sigma}_P^2/c + \hat{\sigma}_Q^2/(1-c)$. Then*

$$\frac{\sqrt{n+m}}{\hat{\sigma}_c(h)} \left[\left(\hat{h}_P^n - \bar{h}_P \right) - \left(\hat{h}_Q^m - \bar{h}_Q \right) \right] \overset{d}{\to} \mathcal{N}(0, 1).$$

- Thm. 1 implies that the asympotic test power grows with the **signal-to-noise ratio**

$$\text{SNR}(h) = \frac{\bar{h}_P - \bar{h}_Q}{\sigma_c(h)}. \tag{2}$$

- Direct relation to MMD objective (3): $J(P, Q \mid k) = \frac{1}{\sqrt{2}} \text{SNR}(h_k^{P,Q})$.
- Stage I: Find witness that optimizes the SNR: $\hat{h}_\lambda = \underset{f \in \mathcal{F}}{\text{argmax}} \frac{\bar{f}_{\mathbb{X}_{\text{tr}}} - \bar{f}_{\mathbb{Y}_{\text{tr}}}}{\sigma_{c,\lambda}^{\mathbb{Z}_{\text{tr}}}(f)}$.
- Stage II leads to a consistent test when $\bar{h}_P > \bar{h}_Q$.

Witness via Kernel Fisher Discriminant Analysis

- $\mathbb{Z}_{\text{tr}} = \{x_1, \dots, x_{n_{\text{tr}}}, y_1, \dots, y_{m_{\text{tr}}}\}$.
- $K_{ij} = k(z_i, z_j)$ for $i, j \in [n_{\text{tr}} + m_{\text{tr}}]$.
- $\delta = (\frac{1}{n_{\text{tr}}}, \dots, \frac{1}{n_{\text{tr}}}, -\frac{1}{m_{\text{tr}}}, \dots, -\frac{1}{m_{\text{tr}}})^\top \in \mathbb{R}^{n_{\text{tr}} + m_{\text{tr}}}$.
- $P_l = I_l - l^{-1} \mathbf{1}_l \mathbf{1}_l^\top$ and $N_c = \begin{pmatrix} \frac{1}{c} P_{n_{\text{tr}}} & 0 \\ 0 & \frac{1}{1-c} P_{m_{\text{tr}}} \end{pmatrix}$.
- Empirical KFDA witness:

$$\hat{h}_\lambda(\cdot) = \sum_{i=1}^{n_{\text{tr}} + m_{\text{tr}}} \hat{\alpha}_i k(z_i, \cdot),$$

$$\hat{\alpha} = \left(\frac{K N_c K}{n_{\text{tr}} + m_{\text{tr}}} + \lambda K \right)^{-1} K \delta.$$

- Naive Scaling $O((n_{\text{tr}} + m_{\text{tr}})^3)$. We provide an adaption of FALKON [1], based on M Nyström centers and t conjugate gradient steps running in $O((n_{\text{tr}} + m_{\text{tr}}) M t + M^3)$.

From optimized MMD to WiTS tests

- Standard MMD with fixed kernel [2]:
  - MMD $:= \sup_{f \in \mathcal{H}, \|f\| \leq 1} \{\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\}$.
  - Kernel mean embedding: $\mu_P = \mathbb{E}[k(X, \cdot)]$.
  - MMD-Witness: $\underset{f \in \mathcal{H}, \|f\| \leq 1}{\text{argmax}} \{\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\} \propto \mu_P - \mu_Q =: h_k^{P,Q}$.

$$\begin{aligned} \text{MMD}^2 &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle = \langle \mu_P - \mu_Q, h_k^{P,Q} \rangle \\ &= \mathbb{E}\left[h_k^{P,Q}(X)\right] - \mathbb{E}\left[h_k^{P,Q}(Y)\right]. \end{aligned}$$

  - MMD test statistic with $h_k^{\mathbb{Z}} = \mu_{\mathbb{X}} - \mu_{\mathbb{Y}}$:

$$\widehat{\text{MMD}}^2_{\text{boot}}(\mathbb{Z}|k) = \frac{1}{n} \sum_{x \in \mathbb{X}} h_k^{\mathbb{Z}}(x) - \frac{1}{m} \sum_{y \in \mathbb{Y}} h_k^{\mathbb{Z}}(y)$$

- MMD with optimized (deep) kernels [3, 4] Based on data splitting: $\mathbb{X} \to \mathbb{X}_{\text{tr}}, \mathbb{X}_{\text{te}}$, $\mathbb{Y} \to \mathbb{Y}_{\text{tr}}, \mathbb{Y}_{\text{te}}$.
  - Optimize kernel on training data to maximize

$$J(P, Q|k) = \text{MMD}^2(P, Q|k) / \sigma_{H_1}(P, Q|k). \tag{3}$$

  - Perform standard MMD test on test data:

$$\widehat{\text{MMD}}^2_{\text{opt-boot}}(\mathbb{Z}_{\text{te}}|k_{\text{tr}}) = \frac{1}{n_{\text{te}}} \sum_{x \in \mathbb{X}_{\text{te}}} h_{k_{\text{tr}}}^{\mathbb{Z}_{\text{te}}}(x) - \frac{1}{m_{\text{te}}} \sum_{y \in \mathbb{Y}_{\text{te}}} h_{k_{\text{tr}}}^{\mathbb{Z}_{\text{te}}}(y).$$

  - "Problem": The witness is still defined via the test data.
- "Solution": Define the witness completely on the training data: $h_{k_{\text{tr}}}^{\mathbb{Z}_{\text{te}}} \to h_{k_{\text{tr}}}^{\mathbb{Z}_{\text{tr}}}$.

$$\widehat{\text{MMD}}^2_{\text{opt-witness}}\left(\mathbb{Z}_{\text{te}} \middle| h_{k_{\text{tr}}}^{\mathbb{Z}_{\text{tr}}}\right) = \frac{1}{n_{\text{te}}} \sum_{x \in \mathbb{X}_{\text{te}}} h_{k_{\text{tr}}}^{\mathbb{Z}_{\text{tr}}}(x) - \frac{1}{m_{\text{te}}} \sum_{y \in \mathbb{Y}_{\text{te}}} h_{k_{\text{tr}}}^{\mathbb{Z}_{\text{tr}}}(y).$$

Algorithm 1 WiTS test with kfda-witness

- **Input:** $\mathbb{X}, \mathbb{Y}, \alpha$, paramGrid, r
- $\mathbb{X}_{\text{tr}}, \mathbb{X}_{\text{te}}, \mathbb{Y}_{\text{tr}}, \mathbb{Y}_{\text{te}} \leftarrow$ RandomSplit$(\mathbb{X}, \mathbb{Y}, r)$
- # Optionally perform model selection
- $k, \lambda \leftarrow$ GridSearchCV(paramGrid, $\mathbb{Z}_{\text{tr}}$)
- # Stage I - Optimize Witness
- $h \leftarrow$ kfdaWitness$(\mathbb{Z}_{\text{tr}}, k, \lambda)$
- # Stage II - Test
- **return:** witnessTest$(\mathbb{Z}_{\text{te}}, h, \alpha)$
- **function** witnessTest$(\mathbb{Z}_{\text{te}}, h(\cdot), \alpha, B = 200)$
  - $h_{\mathbb{Z}_{\text{te}}} \leftarrow [h(z)$ for z in $\mathbb{Z}_{\text{te}}]$
  - $\tau \leftarrow$ mean$(h_{\mathbb{Z}_{\text{te}}}[: n_{\text{te}}])$ − mean$(h_{\mathbb{Z}_{\text{te}}}[n_{\text{te}} :])$
  - $p \leftarrow 0$ ▷ simulate p-value via permutations
  - **for** i in $[B]$ **do**
    - $h_{\mathbb{Z}_{\text{te}}} \leftarrow$ Permute$(h_{\mathbb{Z}_{\text{te}}})$
    - **if** mean$(h_{\mathbb{Z}_{\text{te}}}[: n_{\text{te}}])$ − mean$(h_{\mathbb{Z}_{\text{te}}}[n_{\text{te}} :]) \geq \tau$ **then**
      - $p \leftarrow p + 1/B$
  - **if** $p \leq \alpha$ **then** return: 1 **else** return: 0

Experiments

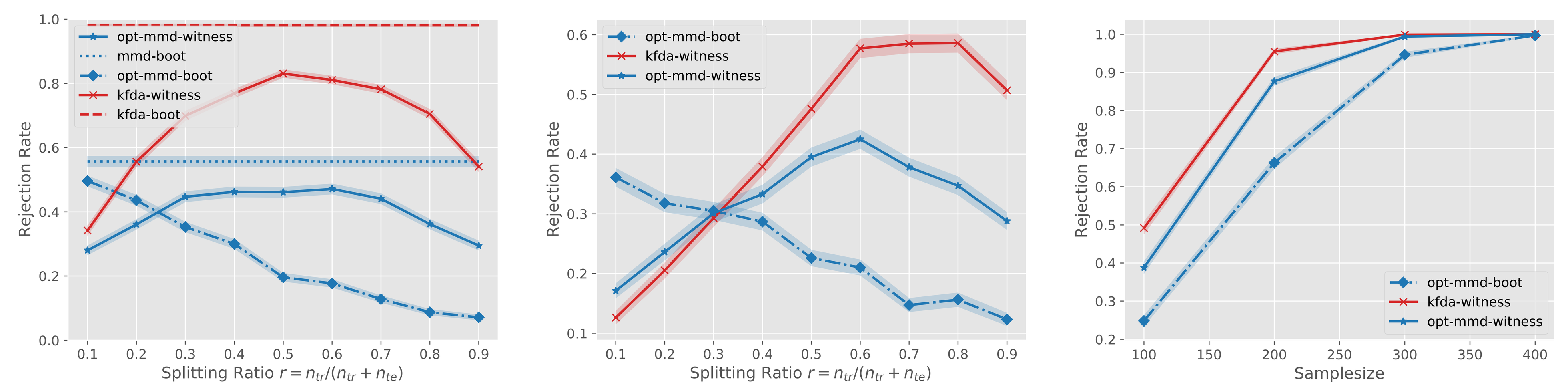

Fig. 1: Instructive experiments on "Blobs" dataset. **Left:** Fixed kernel and fixed regularization for sample size $n = m = 100$. **Middle:** For multiple candidate kernels ($\mathcal{K}_{10}$) kernel optimization becomes more important and the difference of kfda-witness and opt-mmd-witness becomes smaller. Further, opt-mmd-witness already outperforms opt-mmd-boot. **Right:** Same kernels as in the middle figure and $r = 1/2$. All the tests are consistent, i.e., converge to power equal 1.

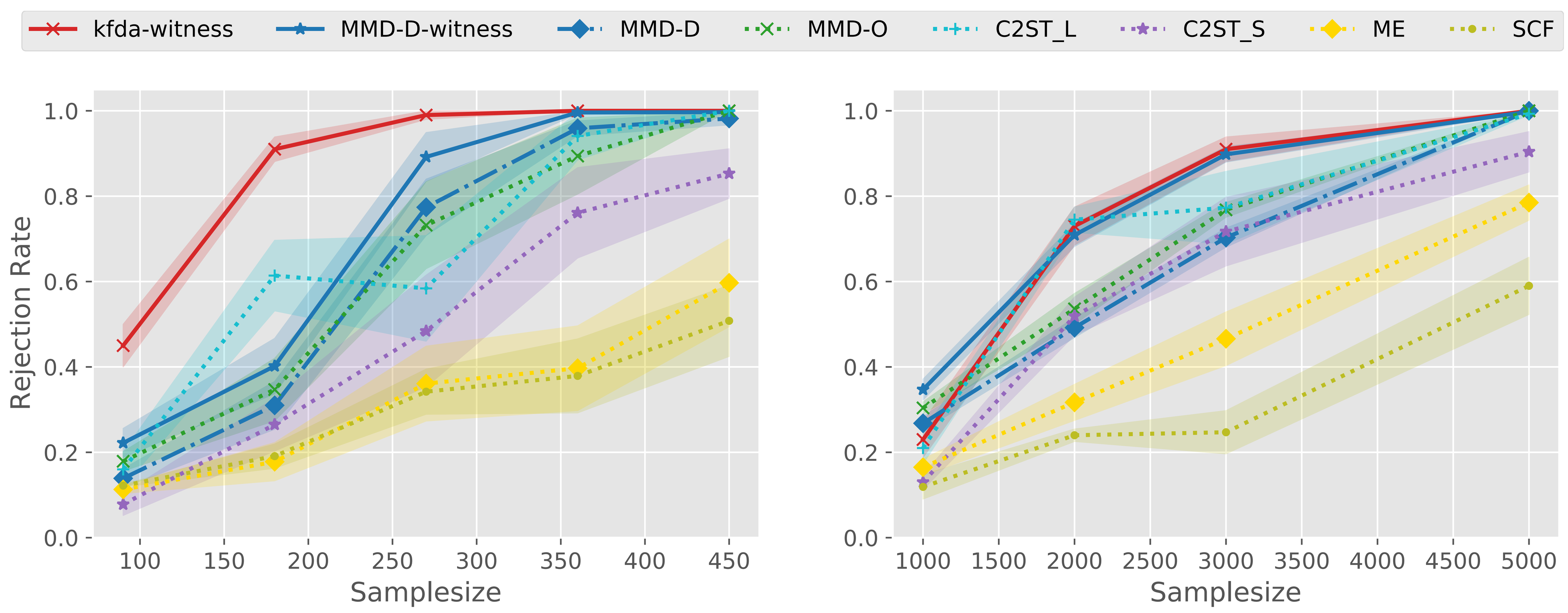

Fig. 2: Benchmark experiments adapted from et al. [4] **Left:** Blobs, **Right:** HIGGS. Computing the MMD witness after kernel optimization and performing a witness test (mmd-d-witness) improves the test power over mmd-d.

References

[1] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *NeurIPS*, 2017.
[2] Arthur Gretton et al. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.
[3] Danica J. Sutherland et al. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*, 2017.
[4] Feng Liu et al. Learning deep kernels for non-parametric two-sample tests. In *ICML*, 2020.