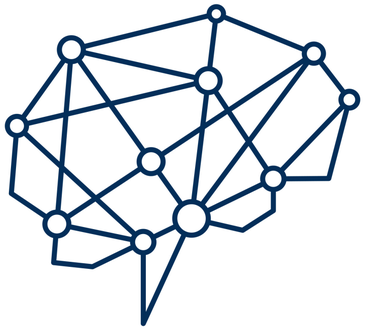
Optimal Kernel Hypothesis Testing

Antonin Schrab
UCL PhD 2020-2024
CDT in Foundational AI
Centre for Artificial Intelligence
Gatsby Computational Neuroscience Unit



Research Papers

MMD Aggregated Two-Sample Test
Schrab, Kim, Albert, Laurent, Guedj & Gretton
JMLR 2022

KSD Aggregated Goodness-of-fit Test
Schrab, Guedj & Gretton
NeurIPS 2022

Efficient Aggregated Kernel Tests using Incomplete U-statistics
Schrab, Kim, Guedj & Gretton
NeurIPS 2022

MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting
Biggs*, Schrab* & Gretton
NeurIPS 2023 (Spotlight)

Differentially Private Permutation Tests: Applications to Kernel Methods
Kim* & Schrab*
Under journal review

Robust Kernel Hypothesis Testing under Data Corruption
Schrab* & Kim*
Under conference review

Outline

Part I: Kernel Discrepancies & Estimators

- Maximum Mean Discrepancy (MMD)
- Hilbert-Schmidt Independence Criterion (HSIC)
- Kernel Stein Discrepancy (KSD)
- Efficient estimators as incomplete U-statistics
- Adaptive estimators via kernel pooling

Part II: Optimal Kernel Hypothesis Testing

- Hypothesis testing
- Adaptive testing via kernel aggregation
- Testing constraints: efficiency, privacy & robustness
- Two-sample MMD testing
- Independence HSIC testing
- Goodness-of-fit KSD testing
- Open problems

Part I
Kernel Discrepancies & Estimators

Maximum Mean Discrepancy

Kernel Methods

Kernel: function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. $k(x, y)$ measures the similarity between x and y

Inner product of features: feature map ϕ

$$k(x, y) = \langle \phi(x), \phi(y) \rangle$$

Positive definite: for all $x_1, \ldots, x_n \in \mathcal{X}$ and all $c_1, \ldots, c_n \in \mathbb{R}$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, k(x_i, x_j) \;\geq\; 0$$

Reproducing Kernel Hilbert Space: $\mathcal{H}_k$ space of real-valued functions on $\mathcal{X}$

$$k(x, \cdot\,) \in \mathcal{H}_k \qquad\qquad \langle f, k(x, \cdot\,) \rangle_{\mathcal{H}_k} = f(x)$$

MMD: Maximum Mean Discrepancy

Integral probability metric:

$$\mathrm{MMD}_k(P, Q) = \sup_{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq 1} \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]$$

Difference in mean embeddings: $\langle f, \mu_P \rangle_{\mathscr{H}_k} = \mathbb{E}_P[f(X)]$

$$\mathrm{MMD}_k = \sup_{f \in \mathscr{H}_k : \|f\|_{\mathscr{H}_k} \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathscr{H}_k} = \|\mu_P - \mu_Q\|_{\mathscr{H}_k}$$

Expected kernel expression:

$$\mathrm{MMD}_k^2 = \|\mu_P - \mu_Q\|_{\mathscr{H}_k}^2 = \mathbb{E}_{P,P}[k(X, X')] - 2\mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_{Q,Q}[k(Y, Y')]$$

MMD Estimators

Biased V-statistic estimator:

$$V_{\mathrm{MMD}_k^2} = \frac{1}{m^2} \sum_{1 \le i,i' \le m} k(X_i, X_{i'}) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j) + \frac{1}{n^2} \sum_{1 \le j,j' \le n} k(Y_j, Y_{j'})$$

Unbiased U-statistic estimator:

$$U_{\mathrm{MMD}_k^2} = \frac{1}{m(m-1)} \sum_{1 \le i \ne i' \le m} k(X_i, X_{i'}) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j) + \frac{1}{n(n-1)} \sum_{1 \le j \ne j' \le n} k(Y_j, Y_{j'})$$

Core function: $h(x, x'; y, y') = k(x, x') - k(x', y) - k(x, y') + k(y, y')$

$$V_{\mathrm{MMD}_k^2} = \frac{1}{m^2 n^2} \sum_{1 \le i,i' \le m} \sum_{1 \le j,j' \le n} h(X_i, X_{i'}; Y_j, Y_{j'}) \qquad U_{\mathrm{MMD}_k^2} = \frac{1}{m(m-1)n(n-1)} \sum_{1 \le i \ne i' \le m} \sum_{1 \le j \ne j' \le n} h(X_i, X_{i'}; Y_j, Y_{j'})$$

$$V_{\mathrm{MMD}_k^2} = \frac{1}{n^2} \sum_{1 \le i,i' \le n} h(X_i, X_{i'}; Y_i, Y_{i'}) \qquad \tilde{U}_{\mathrm{MMD}_k^2} = \frac{1}{n(n-1)} \sum_{1 \le i \ne i' \le n} h(X_i, X_{i'}; Y_i, Y_{i'})$$

MMD Kernel Impact

Radial kernel with bandwidth λ:

$$k_\lambda(x, y) = f\left(\frac{\|x - y\|_r}{\lambda}\right)$$

MMD behaviour:

$$\mathrm{MMD}^2_\lambda \to 0 \text{ when } \lambda \to 0 \text{ or } \lambda \to \infty$$

$$V_{\mathrm{MMD}^2_\lambda} \to \frac{1}{m} + \frac{1}{n} \text{ when } \lambda \to 0$$

$$U_{\mathrm{MMD}^2_\lambda} \to 0 \text{ and } \tilde{U}_{\mathrm{MMD}^2_\lambda} \to 0 \text{ when } \lambda \to 0 \text{ or } \lambda \to \infty$$

$\Longrightarrow$ **Kernel bandwidth choice is crucial for having a meaningful metric**

Hilbert-Schmidt Independence Criterion

HSIC: Hilbert-Schmidt Independence Criterion

Cross-covariance operator $\mathbf{C}_{P_{XY}}$:

$$\langle f, \mathbf{C}_{P_{XY}} g \rangle_{\mathscr{H}_k} = \mathbb{E}_{P_{XY}}\Big[\big(f(X) - \mathbb{E}_{P_X}[f(X')]\big)\big(g(Y) - \mathbb{E}_{P_Y}[g(Y')]\big)\Big]$$

$$\Big\langle f \otimes g, \mathbf{C}_{P_{XY}} \Big\rangle_{\mathrm{HS}} = \mathbb{E}_{P_{XY}}\left[\Big\langle f \otimes g, \big(k(X, \cdot) - \mu_{P_X}\big) \otimes \big(\ell(Y, \cdot) - \mu_{P_Y}\big) \Big\rangle_{\mathrm{HS}}\right]$$

$$\mathbf{C}_{P_{XY}} = \mathbb{E}_{P_{XY}}\Big[\big(k(X, \cdot) - \mu_{P_X}\big) \otimes \big(\ell(Y, \cdot) - \mu_{P_Y}\big)\Big]$$

HSIC:

$$\mathrm{HSIC}^2_{k,\ell}(P_{XY}) = \|\mathbf{C}_{P_{XY}}\|^2_{\mathrm{HS}}$$

$$= \mathbb{E}_{P_{XY},P_{XY}}\left[\Big\langle \big(k(X, \cdot) - \mu_{P_X}\big) \otimes \big(\ell(Y, \cdot) - \mu_{P_Y}\big), \big(k(X', \cdot) - \mu_{P_X}\big) \otimes \big(\ell(Y', \cdot) - \mu_{P_Y}\big) \Big\rangle_{\mathrm{HS}}\right]$$

$$= \mathbb{E}_{P_{XY},P_{XY}}\Big[k(X, X')\ell(Y, Y')\Big] - 2\mathbb{E}_{P_{XY}}\Big[\mathbb{E}_{P_X}[k(X, X')]\mathbb{E}_{P_Y}[\ell(Y, Y')]\Big] + \mathbb{E}_{P_X,P_X}\Big[k(X, X')\Big]\mathbb{E}_{P_Y,P_Y}\Big[\ell(Y, Y')\Big]$$

HSIC Estimators

Notation: $Z_i = X_i, \quad Z_{n+j} = Y_j, \quad K_{ij} = k(X_i, X_j), \quad L_{ij} = \ell(Y_i, Y_j)$

Biased V-statistic estimator:

$$V_{\mathrm{HSIC}^2_{k,\ell}} = \frac{1}{N^2} \sum_{1 \le i,j \le N} K_{ij} L_{ij} - \frac{2}{N} \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{j=1}^{N} K_{ij} \right) \left(\frac{1}{N} \sum_{r=1}^{N} L_{ir} \right) + \left(\frac{1}{N^2} \sum_{1 \le i,j \le N} K_{ij} \right) \left(\frac{1}{N^2} \sum_{1 \le r,s \le N} L_{rs} \right)$$

Unbiased U-statistic estimator:

$$U_{\mathrm{HSIC}^2_{k,\ell}} = \frac{1}{|\,\mathbf{i}_2^N\,|} \sum_{(i,j) \in \mathbf{i}_2^N} K_{ij} L_{ij} - \frac{2}{|\,\mathbf{i}_3^N\,|} \sum_{(i,j,r) \in \mathbf{i}_3^N} K_{ij} L_{ir} + \frac{1}{|\,\mathbf{i}_4^N\,|} \sum_{(i,j,r,s) \in \mathbf{i}_4^N} K_{ij} L_{rs}$$

Core function: $h(Z_i, Z_j, Z_r, Z_s) = K_{ij}\left(L_{ij} - L_{is} - L_{rj} + L_{rs}\right)$

$$V_{\mathrm{HSIC}^2_{k,\ell}} = \frac{1}{N^4} \sum_{1 \le i,j,r,s \le N} h(Z_i, Z_j, Z_r, Z_s)$$

$$U_{\mathrm{HSIC}^2_{k,\ell}} = \frac{1}{|\,\mathbf{i}_4^N\,|} \sum_{(i,j,r,s) \in \mathbf{i}_4^N} h(Z_i, Z_j, Z_r, Z_s)$$

$$\tilde{V}_{\mathrm{HSIC}^2_{k,\ell}} = \frac{1}{N^2} \sum_{1 \le i,j \le N} h(Z_i, Z_j, Z_{i+N/2}, Z_{j+N/2})$$

$$\tilde{U}_{\mathrm{HSIC}^2_{k,\ell}} = \frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} h(Z_i, Z_j, Z_{i+N/2}, Z_{j+N/2})$$

HSIC Kernel Impact

Radial kernels with bandwidths λ, μ:

$$k_\lambda(x,y) = f\left(\frac{\|x-y\|_r}{\lambda}\right) \qquad \ell_\mu(x,y) = g\left(\frac{\|x-y\|_s}{\mu}\right)$$

HSIC behaviour:

$$\mathrm{HSIC}^2_{\lambda,\mu} \to 0 \text{ when } \lambda \to 0 \text{ or } \mu \to 0 \text{ or } \lambda \to \infty \text{ or } \mu \to \infty$$

$$V_{\mathrm{HSIC}^2_{\lambda,\mu}} \to \frac{1}{N} - \frac{1}{N^2} \text{ when } \lambda \to 0 \text{ and } \mu \to 0$$

$$V_{\mathrm{HSIC}^2_{\lambda,\mu}} \to 0 \text{ when } \lambda \to \infty \text{ or } \mu \to \infty$$

$$U_{\mathrm{HSIC}^2_{\lambda,\mu}} \to 0 \text{ when } \lambda \to 0 \text{ or } \mu \to 0 \text{ or } \lambda \to \infty \text{ or } \mu \to \infty$$

$\Longrightarrow$ **Kernel bandwidth choice is crucial for having a meaningful metric**

HSIC as an MMD

Product kernel:

$$\big(k \times \ell\big)\big((X,Y),(X',Y')\big) \;=\; k(X,X')\ell(Y,Y')$$

HSIC as an MMD:

$$\begin{aligned}
&\mathrm{MMD}^2_{k\times\ell}(P_{XY}, P_X \otimes P_Y) \\
&= \mathbb{E}_{P_{XY},P_{XY}}[k(X,X')\ell(Y,Y')] - 2\mathbb{E}_{P_{XY},P_XP_Y}[k(X,X')\ell(Y,Y')] + \mathbb{E}_{P_XP_Y,P_XP_Y}[k(X,X')\ell(Y,Y')] \\
&= \mathbb{E}_{P_{XY},P_{XY}}\Big[k(X,X')\ell(Y,Y')\Big] - 2\,\mathbb{E}_{P_{XY}}\Big[\mathbb{E}_{P_X}[k(X,X')]\mathbb{E}_{P_Y}[\ell(Y,Y')]\Big] + \mathbb{E}_{P_X,P_X}\Big[k(X,X')\Big]\mathbb{E}_{P_Y,P_Y}\Big[\ell(Y,Y')\Big] \\
&= \mathrm{HSIC}^2_{k,\ell}(P_{XY})
\end{aligned}$$

Why use HSIC? MMD estimator requires data splitting (independent samples)
HSIC estimator uses all paired samples to estimate all three terms (more accurate)

MMD as an HSIC

Setting: Given distributions P and Q, construct joint P_{XY} such that:

$$P_X = w_P P + w_Q Q \qquad Y = 1 \text{ if } X \sim P \text{ and } Y = -1 \text{ if } X \sim Q$$

HSIC as an MMD:

$$\mathrm{HSIC}^2_{k,\mathbf{1}}(P_{XY}) = 2w_P^2 w_Q^2 \,\mathrm{MMD}^2_k(P, Q)$$

$$\mathrm{HSIC}^2_{k,\ell_\mu}(P_{XY}) \to 2w_P^2 w_Q^2 \,\mathrm{MMD}^2_k(P, Q) \text{ when } \mu \to 0$$

Estimators: $w_P = m/(m+n),\ w_Q = n/(m+n)$

$$V_{\mathrm{HSIC}^2_{k,\mathbf{1}}} = \frac{2m^2n^2}{(m+n)^4} V_{\mathrm{MMD}^2_k} \qquad V_{\mathrm{HSIC}^2_{k,\ell_\mu}} \to \frac{2m^2n^2}{(m+n)^4} V_{\mathrm{MMD}^2_k} \text{ when } \mu \to 0$$

Kernel Stein Discrepancy

Stein Discrepancy

Stein operator: $\mathsf{A}_P \colon \mathrm{Func}(\mathbb{R}^d \to \mathbb{R}^d) \to \mathrm{Func}(\mathbb{R}^d \to \mathbb{R})$

$$P = Q \quad \iff \quad \mathbb{E}_Q\big[(\mathsf{A}_P \mathbf{f})(X)\big] = 0 \;\text{ for all }\; \mathbf{f} \in \mathscr{F}$$

Stein discrepancy:

$$\mathrm{SD}_{\mathsf{A}_P}(P, Q) = \sup_{\mathbf{f} \in \mathscr{F}} \; \mathbb{E}_Q\big[(\mathsf{A}_P \mathbf{f})(X)\big] - \mathbb{E}_P\big[(\mathsf{A}_P \mathbf{f})(X)\big] = \sup_{\mathbf{f} \in \mathscr{F}} \; \mathbb{E}_Q\big[(\mathsf{A}_P \mathbf{f})(X)\big]$$

Langevin Stein operator:

$$(\mathsf{A}_P \mathbf{f})(x) = \mathbf{f}(x)^\top \nabla \log p(x) + \nabla^\top \mathbf{f}(x) \quad \text{where} \quad \nabla^\top \mathbf{f}(x) = \sum_{i=1}^{d} \frac{\partial}{\partial x_i} f_i(x)$$

Diffusion Stein operator:

$$(\mathsf{A}_P \mathbf{f})(x) = \mathbf{f}(x)^\top \nabla \log p(x) + \nabla^\top \mathbf{f}(x) = \mathbf{f}(x)^\top \left(\frac{\nabla p(x)}{p(x)} \right) + \nabla^\top \mathbf{f}(x) = \frac{1}{p(x)} \Big(\mathbf{f}(x)^\top \nabla p(x) + \big(\nabla^\top \mathbf{f}(x) \big) p(x) \Big) = \frac{1}{p(x)} \Big(\nabla^\top \big(\mathbf{f}(x) p(x) \big) \Big)$$

Stein's identity satisfied:

$$\mathbb{E}_P\big[(\mathsf{A}_P \mathbf{f})(X)\big] = \int_{\mathbb{R}^d} (\mathsf{A}_P \mathbf{f})(x) p(x)\, dx = \int_{\mathbb{R}^d} \nabla^\top \big(\mathbf{f}(x) p(x) \big)\, dx = \sum_{i=1}^{d} \int_{\mathbb{R}^d} \frac{\partial}{\partial x_i} \big(f_i(x) p(x) \big)\, dx = 0$$

KSD: Kernel Stein Discrepancy

Kernelise the Stein operator:

$$(\mathsf{A}_P \mathbf{f})(x) = \mathbf{f}(x)^\top \nabla \log p(x) + \nabla^\top \mathbf{f}(x) = \langle \mathbf{f}, \xi_P(x) \rangle_{\mathscr{H}_k^d} \qquad \xi_P(x) = \nabla \log p(x) k(x, \cdot) + \nabla k(x, \cdot)$$

Kernel Stein Discrepancy:

$$\mathrm{KSD}_P(Q) = \sup_{\mathbf{f} \in \mathscr{F}_k} \mathbb{E}_Q\big[(\mathsf{A}_P \mathbf{f})(X)\big] = \sup_{\mathbf{f} \in \mathscr{F}_k} \big\langle \mathbf{f}, \mathbb{E}_Q\left[\xi_P(X)\right] \big\rangle_{\mathscr{H}^d} = \left\| \mathbb{E}_Q\left[\xi_P(X)\right] \right\|_{\mathscr{H}_k^d}$$

$$\mathrm{KSD}_P^2(Q) = \left\| \mathbb{E}_Q\big[\xi_P(X)\big] \right\|_{\mathscr{H}_k^d}^2 = \big\langle \mathbb{E}_Q\big[\xi_P(X)\big], \mathbb{E}_Q[\xi_P(Y)] \big\rangle_{\mathscr{H}_k^d} = \mathbb{E}_{Q,Q}\Big[\langle \xi_P(X), \xi_P(Y) \rangle_{\mathscr{H}_k^d} \Big] = \mathbb{E}_{Q,Q}[h_P(X, Y)]$$

Stein kernel / core function:

$$\begin{aligned} h_P(x, y) &= \langle \xi_P(x), \xi_P(y) \rangle_{\mathscr{H}_k^d} \\ &= \big(\nabla \log p(x)^\top \nabla \log p(y) \big) k(x, y) + \nabla \log p(x)^\top \nabla_y k(x, y) + \nabla \log p(y)^\top \nabla_x k(x, y) + \sum_{i=1}^{d} \frac{\partial}{\partial x_i y_i} k(x, y) \\ &= \mathsf{A}_{P,2} \mathbf{A}_{P,1} \big(k(x, y) I_{d \times d} \big) \end{aligned}$$

Stein's identity:

$$\mathbb{E}_P[h_P(X, \cdot)] = \mathbb{E}_P[\xi_P(X)] = 0$$

KSD Estimators

Kernel Stein Discrepancy:

$$\mathrm{KSD}_P^2(Q) = \mathbb{E}_{Q,Q}[h_P(X, Y)]$$

Stein kernel / core function:

$$h_P(x, y) = \left(\nabla \log p(x)^\top \nabla \log p(y)\right) k(x, y) + \nabla \log p(x)^\top \nabla_y k(x, y) + \nabla \log p(y)^\top \nabla_x k(x, y) + \sum_{i=1}^{d} \frac{\partial}{\partial x_i y_i} k(x, y)$$

Biased V-statistic estimator:

$$V_{\mathrm{KSD}_k^2} = \frac{1}{n^2} \sum_{1 \le i,j \le n} h_P(X_i, X_j)$$

Unbiased U-statistic estimator:

$$U_{\mathrm{KSD}_k^2} = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h_P(X_i, X_j)$$

KSD Kernel Impact

Notation:

$$\mathbf{s}_P(x) = \nabla \log p(x), \quad \delta(x) = \mathbf{s}_P(x) - \mathbf{s}_Q(x), \quad (\mathsf{A}_P\mathbf{f})(x) = \mathbf{f}(x)^\top \mathbf{s}_P(x) + \nabla^\top \mathbf{f}(x), \quad k_\lambda(x,y) = f\big(\|x-y\|_r / \lambda\big)$$

Useful property:

$$\mathbb{E}_Q\big[(\mathsf{A}_P\mathbf{f})(X)\big] = \mathbb{E}_Q\big[(\mathsf{A}_P\mathbf{f})(X) - (\mathsf{A}_Q\mathbf{f})(X)\big] = \mathbb{E}_Q\big[\mathbf{f}(X)^\top\big(\mathbf{s}_P(X) - \mathbf{s}_Q(X)\big)\big] = \mathbb{E}_Q\big[\mathbf{f}(X)^\top \delta(X)\big]$$

Alternative KSD expression:

$$\mathbb{E}_{Q,Q}\left[k(X,Y)\delta(Y)^\top\delta(X)\right] = \mathbb{E}_{Q,Q}\Big[\Big(k(X,Y)\mathbf{s}_P(X) + \nabla_X k(X,Y)\Big)^\top \delta(Y)\Big] = \mathbb{E}_{Q,Q}\left[h_P(X,Y)\right] = \mathrm{KSD}_P^2(Q)$$

KSD behaviour:

$$\mathrm{KSD}_\lambda^2 = \mathbb{E}_{Q,Q}\left[k_\lambda(X,Y)\delta(X)^\top\delta(Y)\right] \to \mathbb{E}_Q\left[\delta(X)^\top\delta(X)\right] = \mathrm{Fisher}(P,Q) \quad \text{when} \quad \lambda \to 0$$

$$\mathrm{KSD}_\lambda^2 = \mathbb{E}_{Q,Q}\big[k_\lambda(X,Y)\delta(X)^\top\delta(Y)\big] \to \mathbb{E}_{Q,Q}\big[\delta(X)^\top\delta(Y)\big] = \Big\|\mathbb{E}_Q\big[\delta(X)\big]\Big\|_2^2 \quad \text{when} \quad \lambda \to \infty$$

Link to Fisher divergence:

$$\mathrm{KSD}_P^2(Q) = \mathbb{E}_{Q,Q}\left[k(X,Y)\delta(X)^\top\delta(Y)\right] \le \sqrt{\mathbb{E}_{Q,Q}\left[k(X,Y)^2\right]\mathbb{E}_{Q,Q}\left[\left(\delta(X)^\top\delta(Y)\right)^2\right]} \le \sqrt{\mathbb{E}_{Q,Q}\big[k(X,Y)^2\big]}\,\mathrm{Fisher}(P,Q)$$

KSD as an MMD

Stein's identity:

$$\mathbb{E}_P[h_P(X, \cdot)] = 0$$

KSD as an MMD:

$$\begin{aligned}
&\mathrm{MMD}^2_{h_P}(P, Q) \\
&= \mathbb{E}_{Q,Q}[h_P(X, X')] - 2\mathbb{E}_{Q,P}[h_P(X, Y)] + \mathbb{E}_{P,P'}[h_P(Y, Y')] \\
&= \mathbb{E}_{Q,Q}[h_P(X, X')] \\
&= \mathrm{KSD}^2_P(Q)
\end{aligned}$$

Why use KSD?

MMD requires samples from the model P while KSD does not

MMD as a KSD

Simple Stein operator:

$$(\tilde{\mathsf{A}}_P \mathbf{f})(x) = f_1(x) - \mathbb{E}_P[f_1(X)]$$

Stein's identity:

$$\mathbb{E}_P[(\tilde{\mathsf{A}}_P \mathbf{f})(X)] = 0$$

KSD as an MMD:

$$\mathrm{KSD}_{\tilde{\mathsf{A}}_P}(Q) = \sup_{\mathbf{f} \in \mathscr{H}_k^d} \mathbb{E}_Q\left[(\tilde{\mathsf{A}}_P \mathbf{f})(X)\right] = \sup_{f_1 \in \mathscr{H}_k} \mathbb{E}_Q\left[f_1\right] - \mathbb{E}_P\left[f_1\right] = \mathrm{MMD}_k(P, Q)$$

One-sample MMD:

$$\frac{1}{N^2} \sum_{1 \le i,j \le N} k(X_i, X_j) - \frac{2}{N} \sum_{i=1}^{N} \mathbb{E}_P\left[k(X_i, Y)\right] + \mathbb{E}_{P,P}\left[k(Y, Y')\right]$$

Efficient Kernel Estimators

Efficient Kernel Estimators

Expectation:

$$\mathbb{E}[h(X, X')]$$

Statistic with design $\mathscr{D}$:

$$\frac{1}{|\mathscr{D}|} \sum_{(i,j)\in\mathscr{D}} h(X_i, X_j)$$

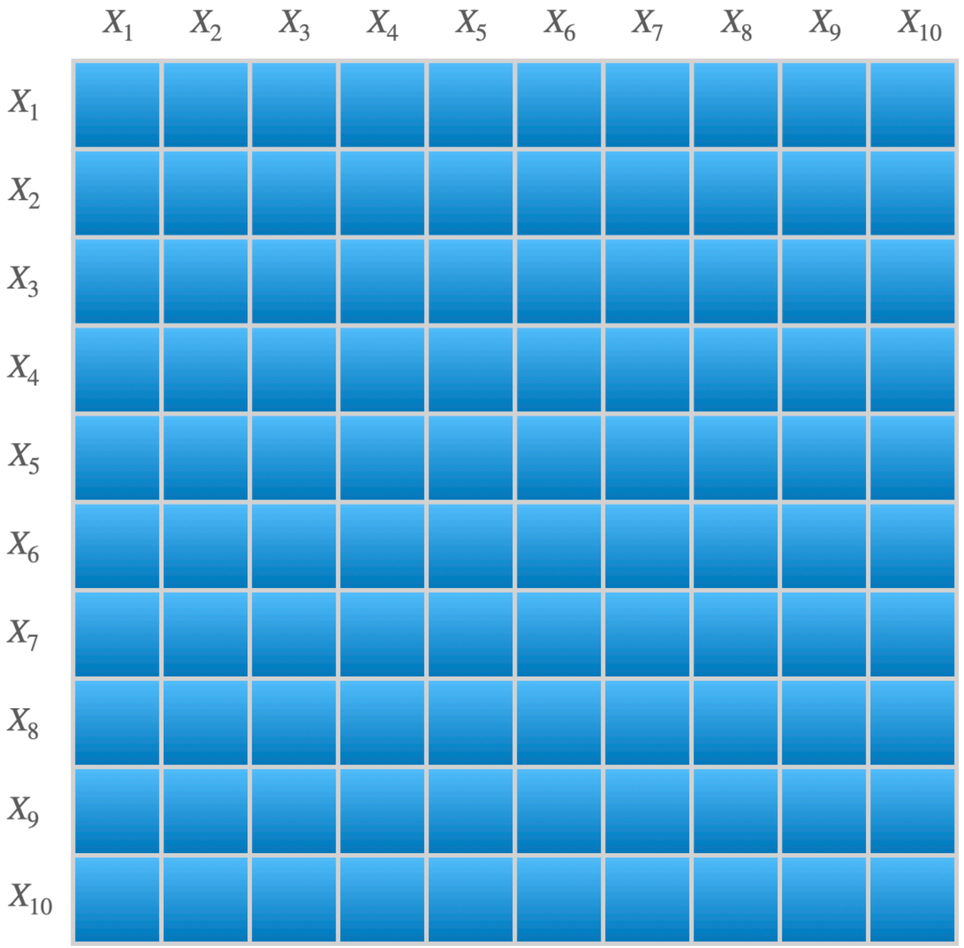
Examples:

$$V_{\mathrm{MMD}_k^2} = \frac{1}{n^2} \sum_{1\le i,j\le n} h(X_i, X_j; Y_i, Y_j), \qquad h(x, x'; y, y') = k(x, x') - k(x', y) - k(x, y') + k(y, y')$$

$$\tilde{V}_{\mathrm{HSIC}_{k,\ell}^2} = \frac{1}{N^2} \sum_{1\le i,j\le N} h(Z_i, Z_j, Z_{i+N/2}, Z_{j+N/2}), \qquad h(Z_i, Z_j, Z_r, Z_s) = K_{ij}\left(L_{ij} - L_{is} - L_{rj} + L_{rs}\right)$$

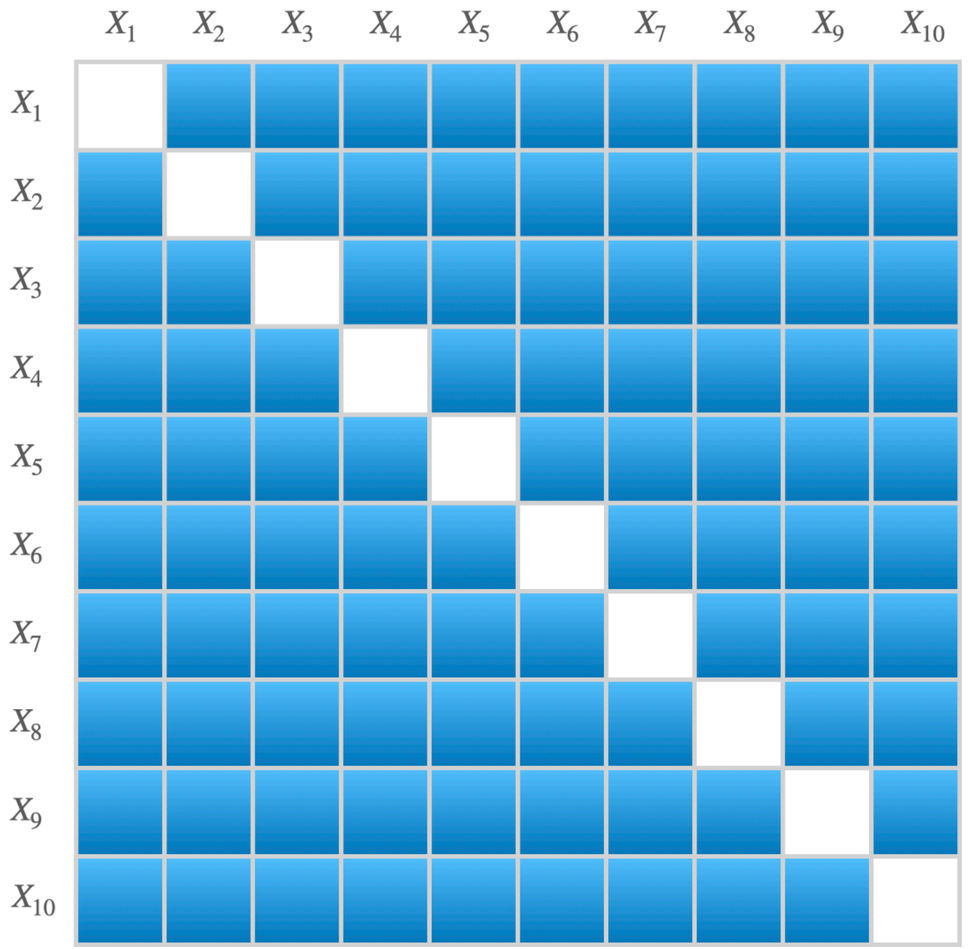
$$V_{\mathrm{KSD}_k^2} = \frac{1}{n^2} \sum_{1\le i,j\le n} h_P(X_i, X_j), \quad h_P(x, y) = \left(\mathbf{s}_P(x)^\top \mathbf{s}_P(y)\right) k(x, y) + \mathbf{s}_P(x)^\top \nabla_y k(x, y) + \mathbf{s}_P(y)^\top \nabla_x k(x, y) + \sum_{i=1}^{d} \frac{\partial}{\partial x_i y_i} k(x, y)$$

V-statistic (von Mises)



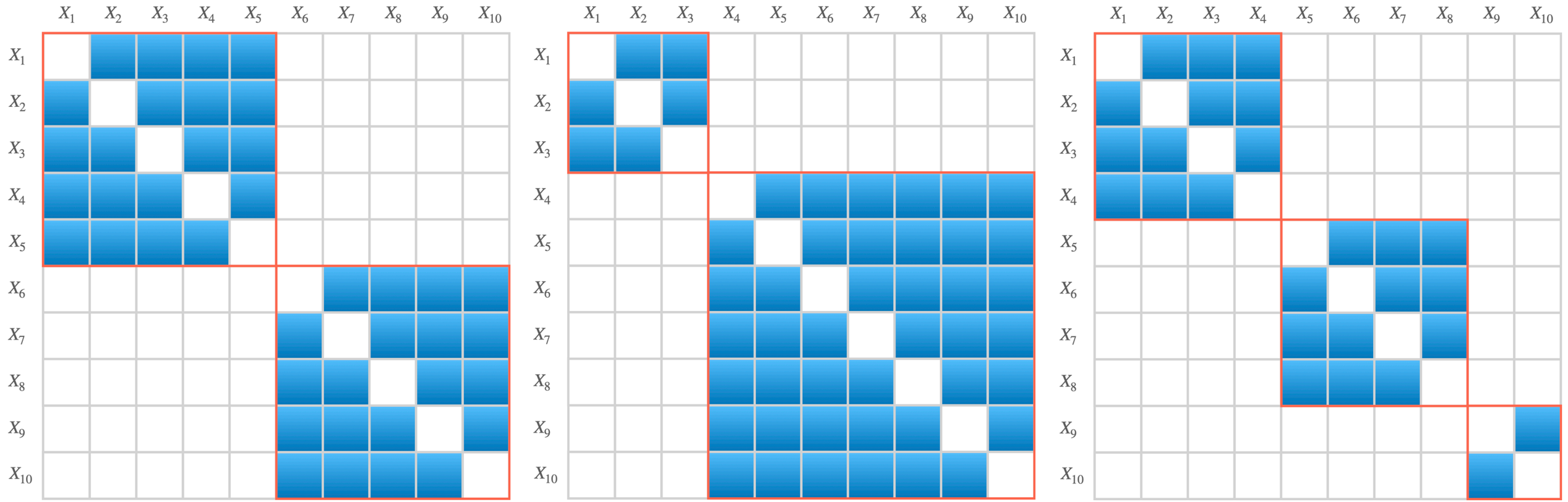
$$V = \frac{1}{n^2} \sum_{1 \le i,j \le n} h(X_i, X_j)$$

U-statistic (unbiased)



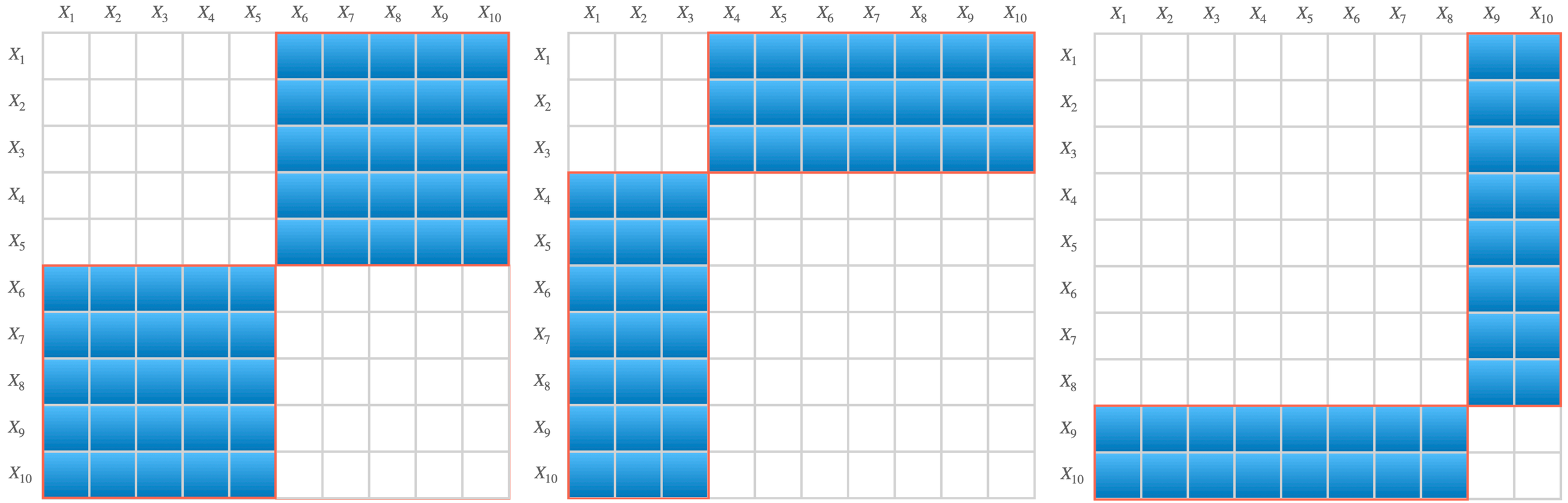
$$U = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j)$$

B-statistic (block)



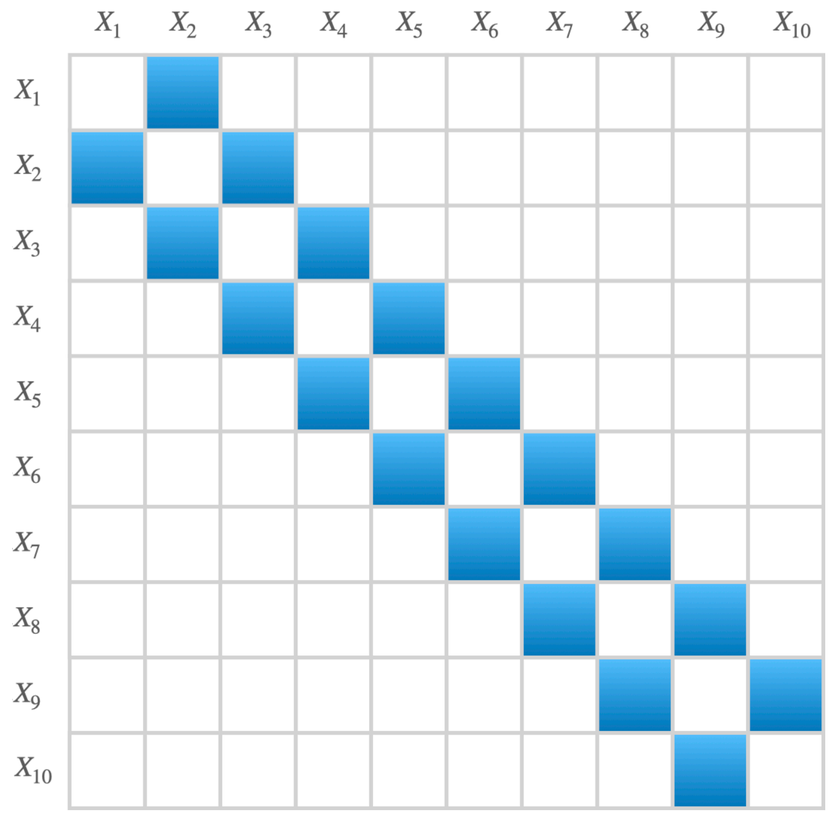
$$B = \sum_{s=1}^{b} U\left(X_{1+\sum_{t=0}^{s-1} n_t}, \ldots, X_{\sum_{t=0}^{s} n_t}\right)$$

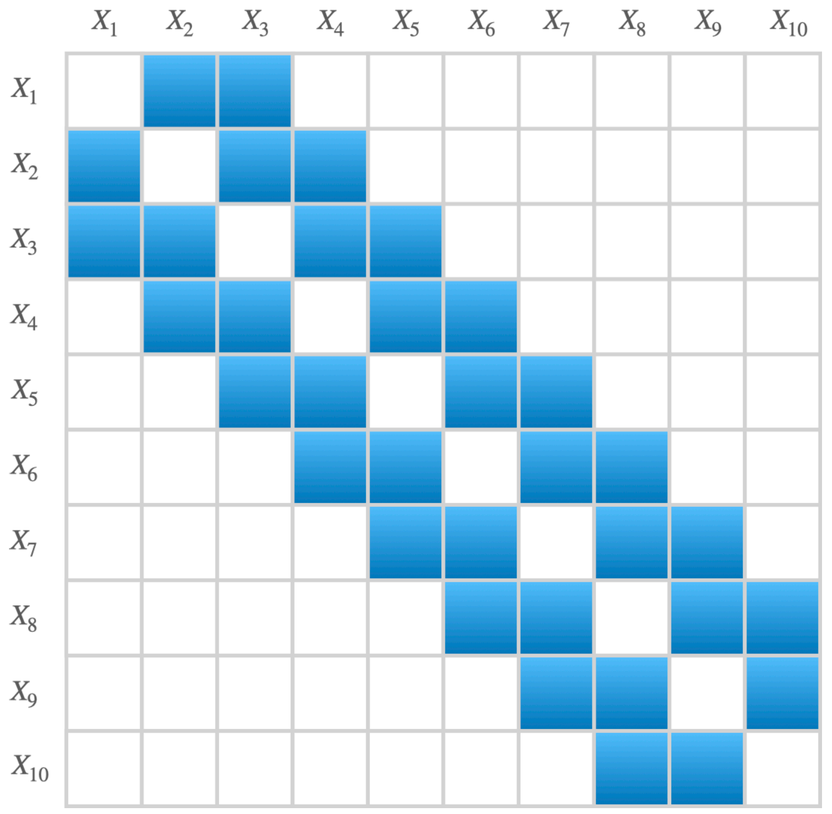
X-statistic (cross)

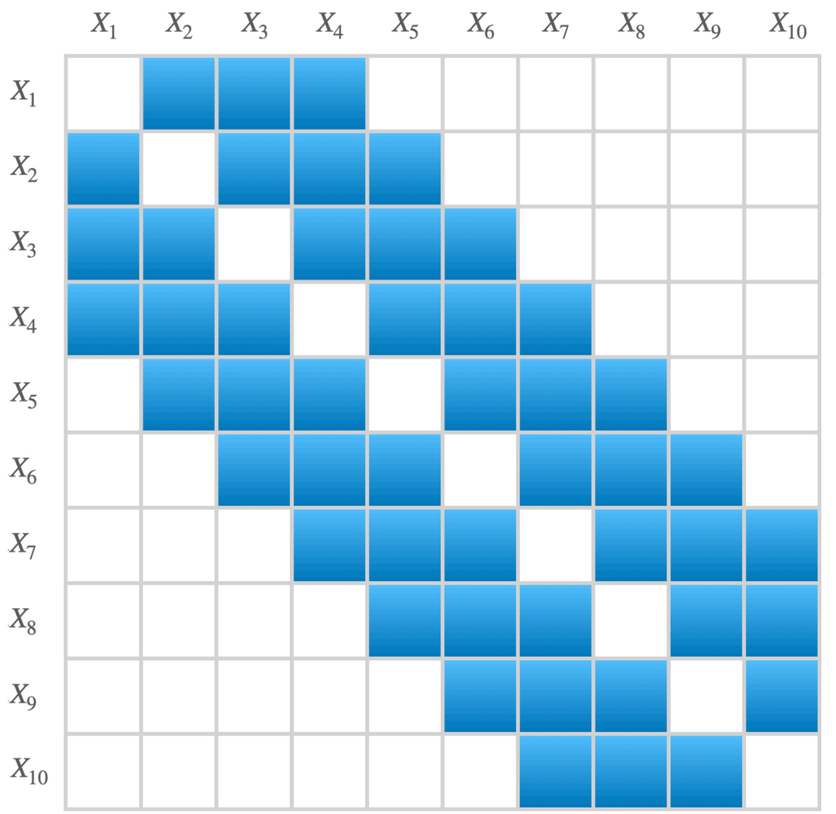


$$X = \frac{1}{n_1(n-n_1)} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} h(X_i, X_j)$$

D-statistic (diagonals)

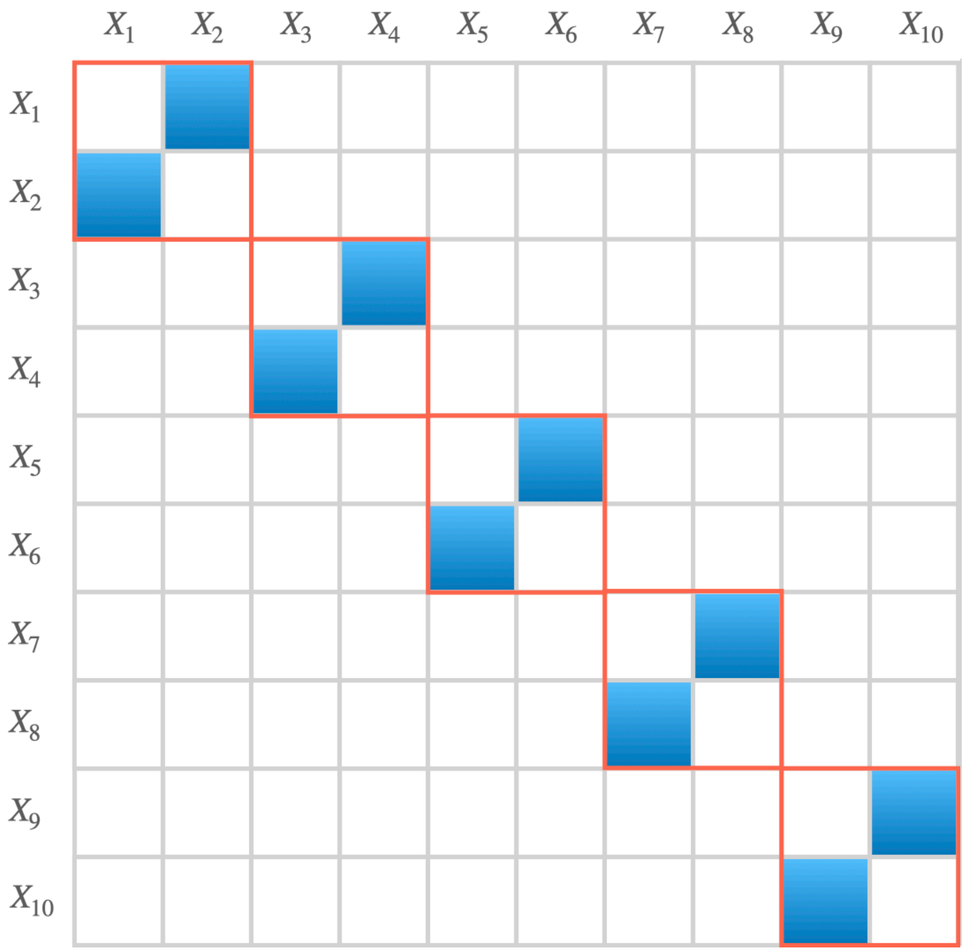






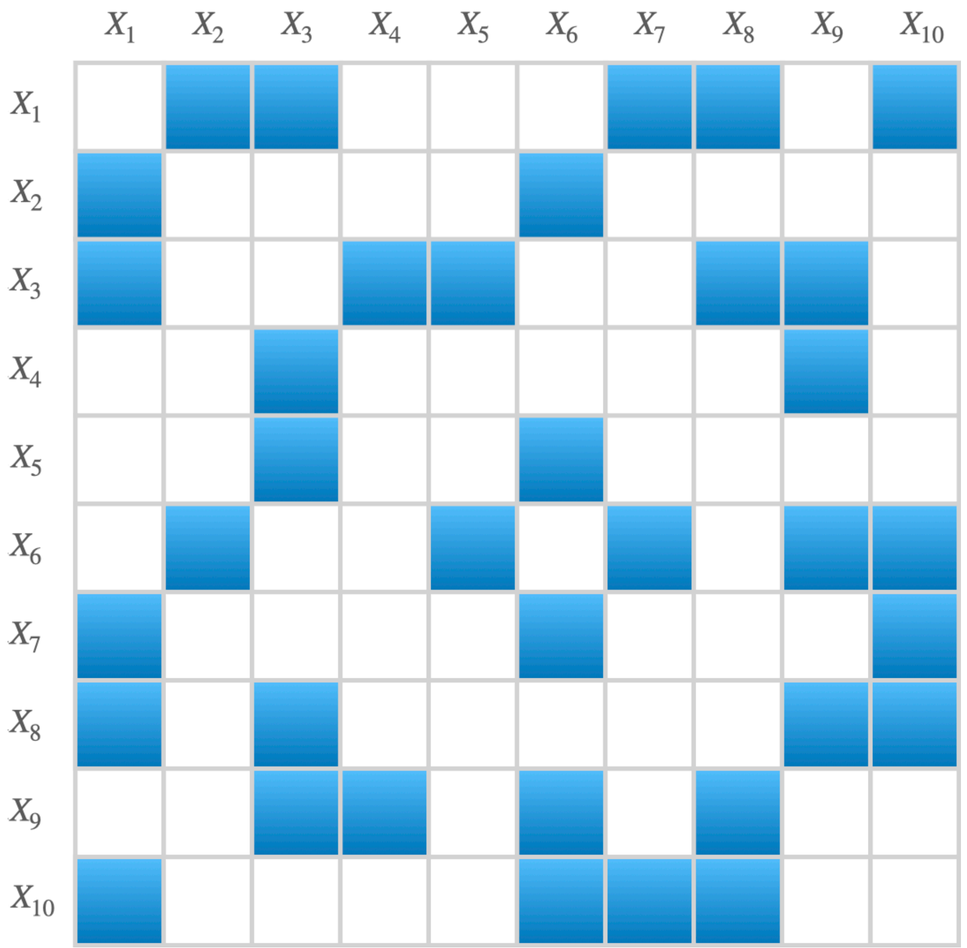
$$D = \frac{2}{r(2n - r - 1)} \sum_{j=1}^{r} \sum_{i=1}^{n-j} h(X_i, X_{i+j})$$

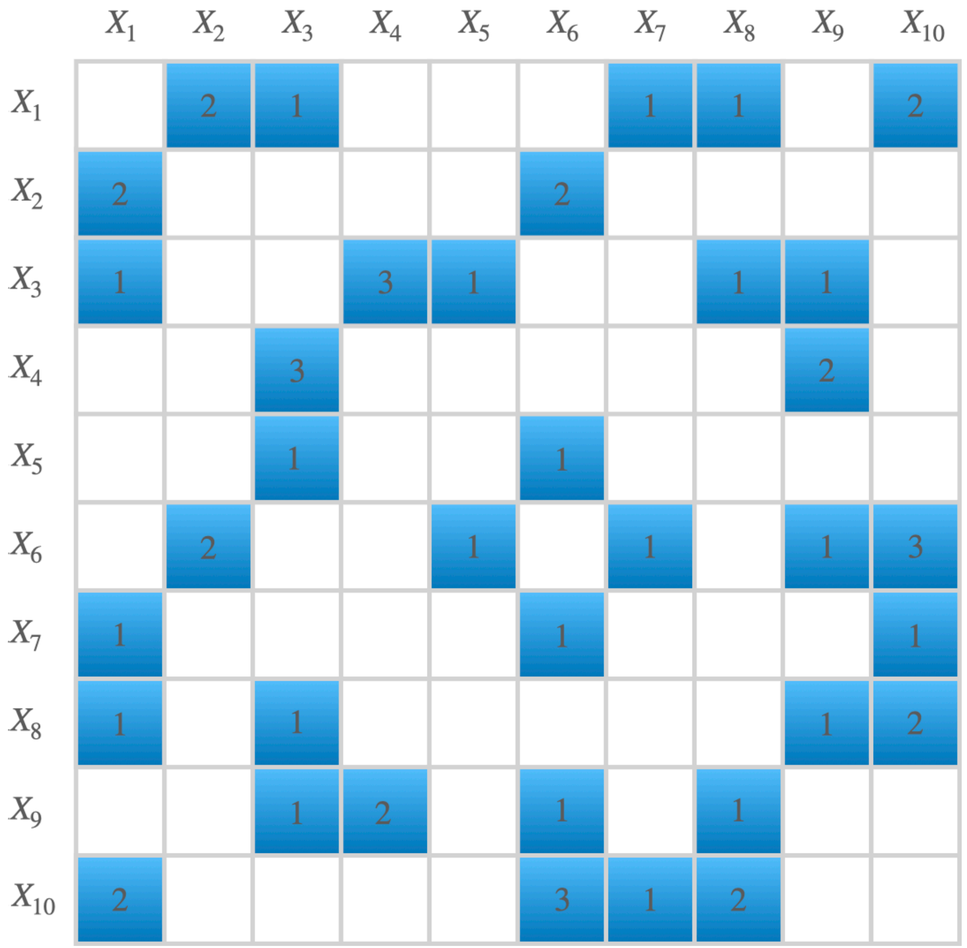
L-statistic (linear)



$$L = \frac{1}{\lfloor n/2 \rfloor} \sum_{1 \le i \le \lfloor n/2 \rfloor} h(X_{2i}, X_{2i-1})$$

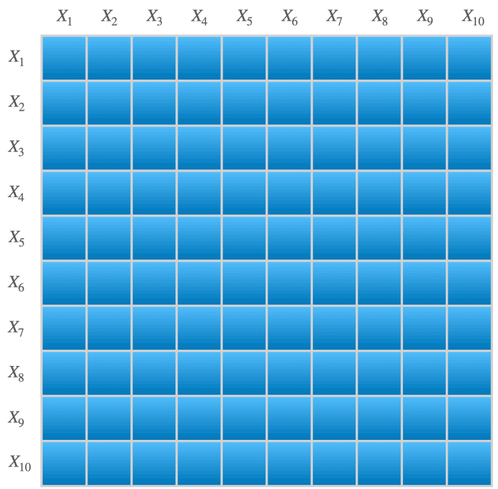
R-statistic (random)

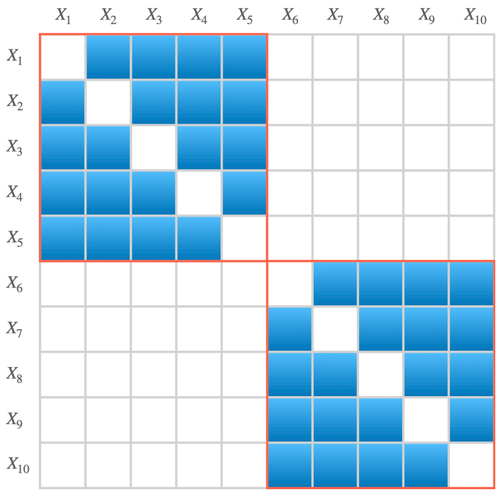


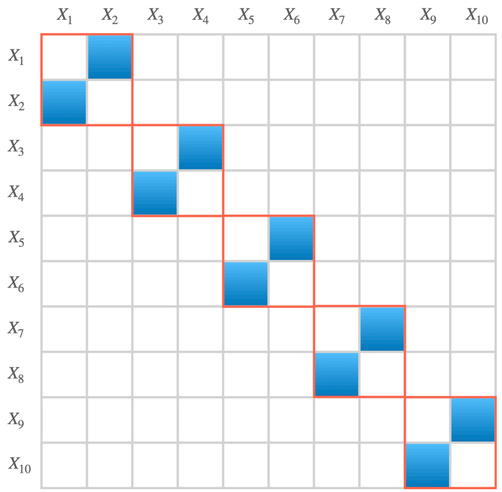


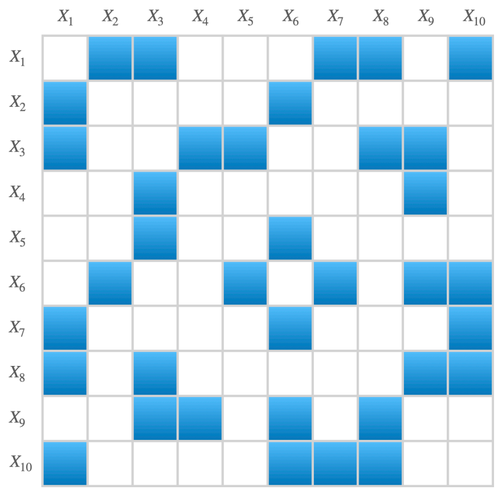
$$R = \frac{1}{|\mathscr{D}_r|} \sum_{(i,j)\in\mathscr{D}_r} h(X_i, X_j)$$

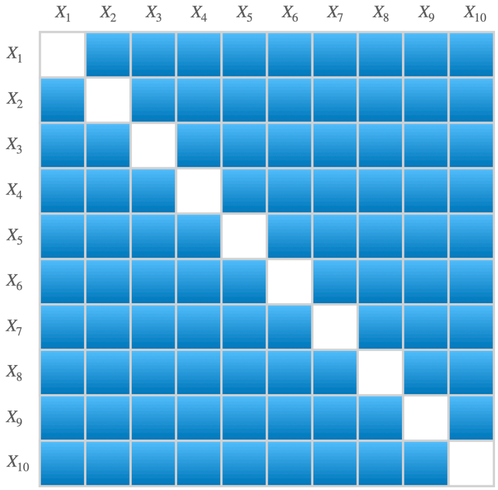
Efficient Kernel Estimators

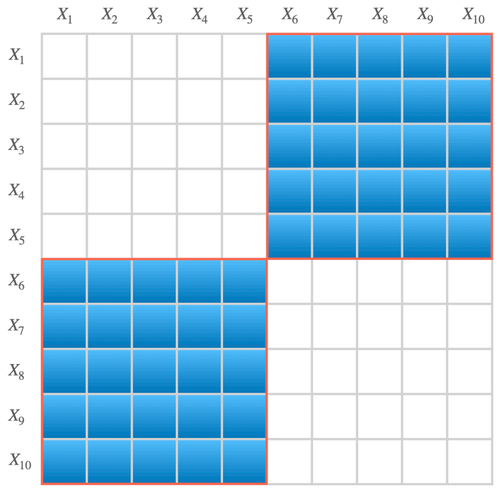


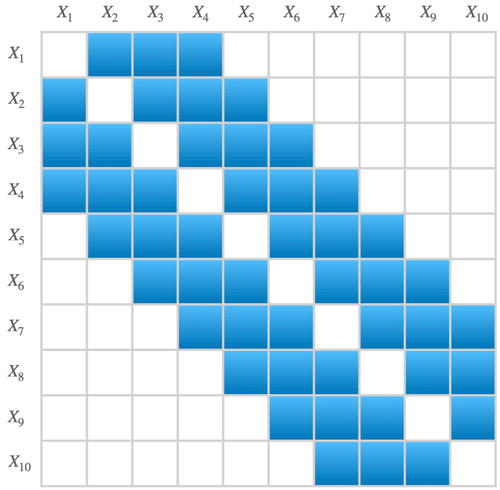


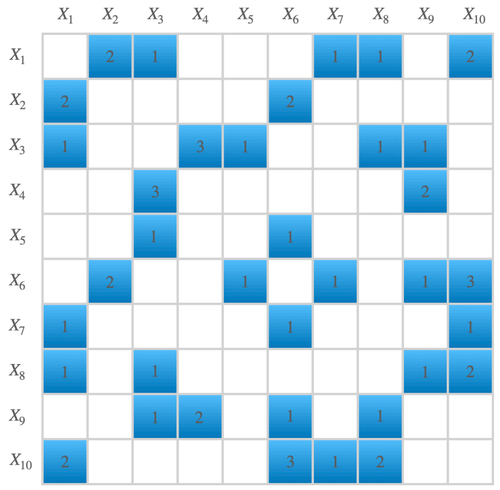












Adaptive Kernel Estimators

Kernel Pooling for Adaptivity

Collection of statistics / kernels:

$$S_k = \frac{1}{|\mathscr{D}_k|} \sum_{(i,j)\in\mathscr{D}_k} h_k(X_i, X_j), \qquad k \in K$$

Normalisation (needed to compare statistics):

$$\frac{S_k}{\sigma_k} \quad \text{where} \quad \sigma_k^2 := \frac{4}{|\mathcal{D}_k^1|} \sum_{i\in\mathcal{D}_k^1} \left(\frac{1}{|\mathcal{D}_k^{2,i}|} \sum_{j\in\mathcal{D}_k^{2,i}} h_k(X_i, X_j) \right)^2 - \left(\frac{2}{|\mathcal{D}_k|} \sum_{(i,j)\in\mathcal{D}_k} h_k(X_i, X_j) \right)^2$$

Kernel pooling:

$$\operatorname*{pool}_{k\in K} \frac{S_k}{\sigma_k}$$

Kernel Pooling for Adaptivity

Mean kernel pooling:

$$\underset{k\in K}{\text{mean}}\ \frac{S_k}{\sigma_k} = \frac{1}{|K|}\sum_{k\in K}\frac{S_k}{\sigma_k} \qquad \frac{1}{|K|}\sum_{k\in K} S_k = S_{\frac{1}{|K|}\sum_{k\in K} k}$$

Max kernel pooling:

$$\underset{k\in K}{\max}\ \frac{S_k}{\sigma_k}$$

Fuse kernel pooling:

$$\underset{k\in K}{\text{fuse}}\ \frac{S_k}{\sigma_k} = \frac{1}{\nu}\log\left(\frac{1}{|K|}\sum_{k\in K}\exp\left(\nu\frac{S_k}{\sigma_k}\right)\right)$$

$$\underset{k\in K}{\max}\frac{S_k}{\sigma_k} - \frac{\log(|K|)}{\nu} \leq \underset{k\in K}{\text{fuse}}\ \frac{S_k}{\sigma_k} \leq \underset{k\in K}{\max}\frac{S_k}{\sigma_k}$$

Kernel Collection

Radial kernel with bandwidth λ:

$$k_\lambda(x, y) = f\left(\frac{\|x - y\|_r}{\lambda}\right)$$

Inter-sample distances:

$$\mathsf{D} = \left\{\|x - x'\|_r : x, x' \in \{X_1, \ldots, X_n\}\right\} \setminus \{0\}$$

Bandwidth collection:

$$\Lambda(k, M) = \left\{q^{\mathsf{D}}_{5\%} + i(q^{\mathsf{D}}_{95\%} - q^{\mathsf{D}}_{5\%})/M : i = 0, \ldots, M\right\}$$

Kernel collection:

$$K = \left\{k_\lambda : k \in \{\text{Gaussian, Laplace}\}, \lambda \in \Lambda(k, 10)\right\}$$

Part II
Optimal Kernel Hypothesis Testing

Hypothesis Testing

Hypothesis Testing

Partitioned space of distributions: $\mathscr{P} = \mathscr{P}_0 \cup \mathscr{P}_1$

Hypothesis testing: Given i.i.d. samples from $P \in \mathscr{P}$ we test whether

$$\mathscr{H}_0 \colon P \in \mathscr{P}_0 \qquad \text{vs} \qquad \mathscr{H}_1 \colon P \in \mathscr{P}_1$$

Testing with a discrepancy:

$$\mathscr{H}_0 \colon \mathrm{Disc}(P) = 0 \qquad \text{vs} \qquad \mathscr{H}_1 \colon \mathrm{Disc}(P) > 0$$

Statistic: estimator using samples from P

$$\widehat{\mathrm{Disc}}$$

Test output: reject $\mathscr{H}_0$ if $\widehat{\mathrm{Disc}}$ is larger than some threshold

Hypothesis Testing: Level

Test output: reject $\mathscr{H}_0$ if $\widehat{\mathrm{Disc}}$ is larger than some threshold

Level: type I error

$$\sup_{P_0 \in \mathscr{P}_0} \mathbb{P}_{P_0}(\text{reject } \mathscr{H}_0) \;\leq\; \alpha$$

Threshold: $(1-\alpha)$-quantile of the simulated distribution of $\widehat{\mathrm{Disc}}$ under the null

Permutations: when $\mathscr{H}_0$ holds if and only if exchangeability holds

$$\widehat{\mathrm{Disc}}\big(X_{\pi(1)}, \ldots, X_{\pi(n)}\big)$$

Exchangeability: $\mathrm{joint}\big(X_1, \ldots, X_n\big) = \mathrm{joint}\big(X_{\pi(1)}, \ldots, X_{\pi(n)}\big)$ for all permutation π

Wild bootstrap: $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher variables (± 1 with probability 0.5)

$$\frac{1}{|\mathscr{D}|} \sum_{(i,j) \in \mathscr{D}} \varepsilon_i \, \varepsilon_j \, h(X_i, X_j)$$

Hypothesis Testing: Power

Type II error:

$$\mathbb{P}_{P_1}\big(\text{fail to reject } \mathscr{H}_0\big),\ P_1 \in \mathscr{P}_1$$

Power:

$$\mathbb{P}_{P_1}\big(\text{reject } \mathscr{H}_0\big),\ P_1 \in \mathscr{P}_1$$

Impossible to control both types of errors uniformly over $\mathscr{P}_0$ and $\mathscr{P}_1$:

$$\sup_{P_0 \in \mathscr{P}_0} \mathbb{P}_{P_0}\big(\text{reject } \mathscr{H}_0\big) \ \leq\ \alpha \qquad \sup_{P_1 \in \mathscr{P}_1} \mathbb{P}_{P_1}\big(\text{fail to reject } \mathscr{H}_0\big) \ \leq\ \beta$$

Pointwise power / consistency: for fixed $P_1 \in \mathscr{P}_1$

$$\lim_{N \to \infty} \mathbb{P}_{P_1}\big(\text{reject } \mathscr{H}_0\big) \ =\ 1$$

Hypothesis Testing: Power

Uniform power over $\mathcal{S}_1 \subset \mathcal{P}_1$:

$$\sup_{P_1 \in \mathcal{S}_1} \mathbb{P}_{P_1}\big(\text{reject } \mathscr{H}_0\big) \;\geq\; 1-\beta \qquad (\bigstar)$$

Example of subset $\mathcal{S}_1$:

$$\mathcal{S}_1 \;=\; \big\{P_1 \in \mathcal{P}_1 \;:\; \mathrm{Disc}(P_1) \;\geq\; f(N,\alpha,\beta)\big\}$$

Upper bound: given a test, find smallest separation rate $f(N,\alpha,\beta)$ s.t. $(\bigstar)$ holds

Lower bound: find largest separation rate $f(N,\alpha,\beta)$ such that no level-α test satisfy $(\bigstar)$

Minimax optimality: test for which upper bound matches the lower bound

Sobolev regularity: smoothness restriction on P_1

$$\int_{\mathbb{R}^d} \|\xi\|_2^{2s}\, |\hat{f}(\xi)|^2 \,\mathrm{d}\xi \;\leq\; (2\pi)^d$$

Aggregation
Kernel Adaptivity

Aggregation (Kernel Adaptivity)

Multiple testing:

reject $\mathscr{H}_0$ if one of the tests $T_{k_1}, \ldots, T_{k_{|K|}}$ rejects

Bonferroni level correction: $\alpha / |K|$

$$\mathbb{P}\big(\text{reject } \mathscr{H}_0\big) \;=\; \mathbb{P}\left(\bigcup_{i=1}^{|K|} \big\{T_{k_i} \text{ rejects } \mathscr{H}_0\big\}\right) \;\leq\; \sum_{i=1}^{|K|} \mathbb{P}\Big(T_{k_i} \text{ rejects } \mathscr{H}_0\Big) \;=\; \alpha$$

Aggregation: estimate level correction between $\alpha / |K|$ and α such that

$$\mathbb{P}\big(\text{reject } \mathscr{H}_0\big) \;=\; \alpha$$

with bisection method using null-simulated samples via permutations or wild bootstrap

Kernel Collection

Radial kernel with bandwidth λ:

$$k_\lambda(x, y) = f\left(\frac{\|x - y\|_r}{\lambda}\right)$$

Inter-sample distances:

$$\mathsf{D} = \left\{\|x - x'\|_r : x, x' \in \{X_1, \ldots, X_n\}\right\} \setminus \{0\}$$

Bandwidth collection:

$$\Lambda(k, M) = \left\{q^{\mathsf{D}}_{5\%} + i(q^{\mathsf{D}}_{95\%} - q^{\mathsf{D}}_{5\%})/M : i = 0, \ldots, M\right\}$$

Kernel collection:

$$K = \left\{k_\lambda : k \in \{\text{Gaussian, Laplace}\}, \lambda \in \Lambda(k, 10)\right\}$$

Testing Constraints:
Efficiency
Privacy
Robustness

Computational Efficiency

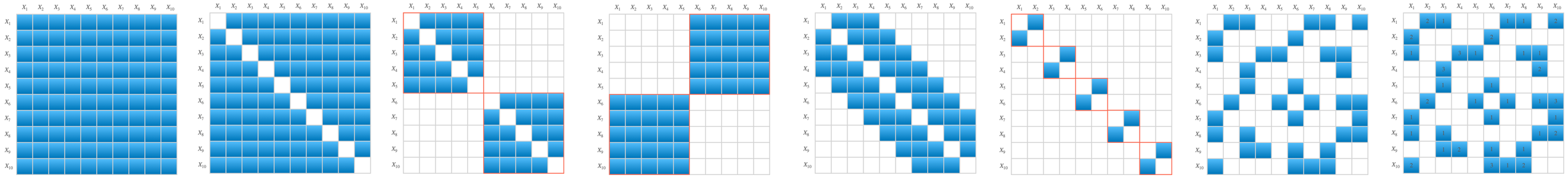
High computational cost: prohibitive for very large datasets

Goal: construct tests with lower computational complexity (e.g. sub-quadratic)

Trade-off: computational efficiency vs test power via uniform separation rates

Our approach: use incomplete statistics as efficient estimators

Others: Nyström approximation, random Fourier features, kernel thinning



Differential Privacy

(ε, δ)-differentially privacy of randomised test:

$$\mathbb{P}\left(\text{reject } \mathscr{H}_0 \text{ using } \mathbb{X}_N\right) \leq e^{\varepsilon}\, \mathbb{P}\left(\text{reject } \mathscr{H}_0 \text{ using } \widetilde{\mathbb{X}}_N\right) + \delta$$

$$\mathbb{P}\left(\text{fail to reject } \mathscr{H}_0 \text{ using } \mathbb{X}_N\right) \leq e^{\varepsilon}\, \mathbb{P}\left(\text{fail to reject } \mathscr{H}_0 \text{ using } \widetilde{\mathbb{X}}_N\right) + \delta$$

for any two datasets $\mathbb{X}_N$ and $\widetilde{\mathbb{X}}_N$ differing only in one entry

Intuition: probability of a test output remains roughly the same when the data of a single user is modified

Privatisation: inject Laplacian noise to original and permuted statistics

$$\widehat{\mathrm{MMD}}_{\pi_i} + \frac{2}{\xi}\frac{\sqrt{2}}{N}\zeta_i \qquad \widehat{\mathrm{HSIC}}_{\pi_i} + \frac{2}{\xi}\frac{4(N-1)}{N^2}\zeta_i \qquad \xi = \varepsilon + \log\left(\frac{1}{1-\delta}\right)$$

Robustness

Standard testing: Given i.i.d. samples $X_1, \ldots, X_N$ from $P \in \mathscr{P}$, determine whether $\mathscr{H}_0 \colon P \in \mathscr{P}_0$ or $\mathscr{H}_1 \colon P \in \mathscr{P}_1$

Robust testing: Given samples $X_1, \ldots, X_N$ where

- $N - r$ samples are i.i.d. from $P \in \mathscr{P}$
- r samples have been corrupted (possibly in a non i.i.d. adversarial manner)

determine whether $\mathscr{H}_0 \colon P \in \mathscr{P}_0$ or $\mathscr{H}_1 \colon P \in \mathscr{P}_1$

Intuition: null hypothesis is enlarged with only $N - r$ samples being i.i.d. from P

Robustisation: add shift to permuted quantile (not to original statistic)

$$q_{1-\alpha}^{\mathrm{MMD}} + 2\frac{\sqrt{2}}{N}r \qquad\qquad q_{1-\alpha}^{\mathrm{HSIC}} + 2\frac{4(N-1)}{N^2}r$$

Two-sample Testing

Two-sample testing

Two-sample problem: Given independent

- i.i.d. samples $X_1, \ldots, X_m$ from a distribution P,
- i.i.d. samples $Y_1, \ldots, Y_n$ from a distribution Q,

test whether $\mathscr{H}_0 \colon P = Q$ or $\mathscr{H}_1 \colon P \neq Q$. Let $N = \min(m, n)$.

General notation: $\mathscr{P}_0 = \{(P, Q) : P = Q\}$ and $\mathscr{P}_1 = \{(P, Q) : P \neq Q\}$

Testing with MMD: $\mathscr{H}_0 \colon \mathrm{MMD}(P, Q) = 0$ and $\mathscr{H}_1 \colon \mathrm{MMD}(P, Q) > 0$

Permutations (exchangeability): $Z_i = X_i$, $Z_{m+i} = Y_i$, $(\mathbb{X}_m^\pi, \mathbb{Y}_n^\pi) = \left(Z_{\pi(1)}, \ldots, Z_{\pi(m+n)}\right)$

Wild bootstrap: $\dfrac{1}{|\mathscr{D}|} \displaystyle\sum_{(i,j) \in \mathscr{D}} \varepsilon_i \varepsilon_j h(X_i, X_j)$, $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. Rademacher variables

Pointwise power / consistency: $\lim\limits_{N \to \infty} \mathbb{P}_{P,Q}\left(\text{reject } \mathscr{H}_0\right) = 1$ for fixed $P \neq Q$

Power in MMD Metric: Fixed Kernel

Standard testing:

$$\mathrm{MMD}_k \gtrsim \sqrt{\frac{\max\{\log(1/\alpha),\ \log(1/\beta)\}}{N}}$$

Efficient testing:

$$\mathrm{MMD}_k \gtrsim \sqrt{\frac{\max\{\log(1/\alpha),\ \log(1/\beta)\}}{N/B}}$$

Robust testing:

$$\mathrm{MMD}_k \gtrsim \max\left\{\sqrt{\frac{\max\{\log(1/\alpha),\ \log(1/\beta)\}}{N}},\ \frac{r}{N}\right\}$$

Differentially private testing:

$$\mathrm{MMD}_k \gtrsim \max\left\{\sqrt{\frac{\max\{\log(1/\alpha),\ \log(1/\beta)\}}{N}},\ \frac{\max\{\log(1/\alpha),\ \log(1/\beta)\}}{N\xi}\right\}$$

Power in MMD Metric: Kernel Pooling

Standard testing:

$$\max_{k \in K} \mathrm{MMD}_k \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K|)\right\}}{N}}$$

Efficient testing:

$$\max_{k \in K} \mathrm{MMD}_k \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K|)\right\}}{N/B}}$$

Robust testing:

$$\max_{k \in K} \mathrm{MMD}_k \gtrsim \max\left\{\sqrt{\frac{\max\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K|)\}}{N}},\ \frac{r}{N}\right\}$$

Power in L2 Sobolev Metric: Oracle Kernel

Standard testing:

$$\|p-q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\,\log(1/\beta)}{N}\right)^{2s/(4s+d)}$$

Efficient testing:

$$\|p-q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\,\log(1/\beta)}{|\mathscr{D}|/N}\right)^{2s/(4s+d)}$$

Differentially private testing:

$\|p-q\|_{L^2} \gtrsim N^{-2s/(4s+d)}$ in low privacy regime $\xi \gtrsim N^{-(2s-d/2)/(4s+d)}$

$\|p-q\|_{L^2} \gtrsim \left(N^{3/2}\xi\right)^{-s/(2s+d)}$ in mid privacy regime $N^{-1/2} \lesssim \xi \lesssim N^{-(2s-d/2)/(4s+d)}$

$\|p-q\|_{L^2} \gtrsim \left(N\xi\right)^{-2s/(2s+d)}$ in high privacy regime $\xi \lesssim N^{-1/2}$

Power in L2 Sobolev Metric: Kernel Aggregation

Standard testing:

$$\|p-q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\,\log(1/\beta)}{N/\,\log(\log(N))}\right)^{2s/(4s+d)}$$

Efficient testing:

$$\|p-q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\,\log(1/\beta)}{\big(|\mathscr{D}|/N\big)/\log\big(\log(|\mathscr{D}|/N)\big)}\right)^{2s/(4s+d)}$$

Independence Testing

Independence testing

Independence problem: Given

- paired samples $(X_1, Y_1), \ldots, (X_N, Y_N)$ drawn i.i.d. from a joint distribution P_{XY},

test whether the first and second components of the pairs are independent

$$\mathscr{H}_0 \colon P_{XY} = P_X \otimes P_Y \qquad \text{vs} \qquad \mathscr{H}_1 \colon P_{XY} \neq P_X \otimes P_Y$$

General notation: $\mathscr{P}_0 = \{P_{XY} : P_{XY} = P_X \otimes P_Y\}$ and $\mathscr{P}_1 = \{P_{XY} : P_{XY} \neq P_X \otimes P_Y\}$

Testing with HSIC: $\mathscr{H}_0 \colon \mathrm{HSIC}(P_{XY}) = 0$ and $\mathscr{H}_1 \colon \mathrm{HSIC}(P_{XY}) > 0$

Permutations (exchangeability): $\left(X_1, Y_{\pi(1)}\right), \ldots, \left(X_N, Y_{\pi(N)}\right)$

Wild bootstrap: $\dfrac{1}{|\mathscr{D}|} \displaystyle\sum_{(i,j)\in\mathscr{D}} \varepsilon_i \varepsilon_j h(X_i, X_j)$, $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. Rademacher variables

Pointwise power / consistency: $\lim\limits_{N\to\infty} \mathbb{P}_{P,Q}\left(\text{reject } \mathscr{H}_0\right) = 1$ for fixed $P_{XY} \neq P_X \otimes P_Y$

Power in HSIC Metric: Fixed Kernels

Standard testing:

$$\mathrm{HSIC}_{k,\ell} \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta)\right\}}{N}}$$

Efficient testing:

$$\mathrm{HSIC}_{k,\ell} \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta)\right\}}{N/B}}$$

Robust testing:

$$\mathrm{HSIC}_{k,\ell} \gtrsim \max\left\{\sqrt{\frac{\max\{\log(1/\alpha),\ \log(1/\beta)\}}{N}}, \frac{r}{N}\right\}$$

Differentially private testing:

$$\mathrm{HSIC}_{k,\ell} \gtrsim \max\left\{\sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta)\right\}}{N}}, \frac{\max\left\{\log(1/\alpha),\ \log(1/\beta)\right\}}{N\xi}\right\}$$

Power in MMD Metric: Kernel Pooling

Standard testing:

$$\max_{k\in K}\ \max_{\ell\in L}\ \mathrm{HSIC}_{k,\ell} \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K||L|)\right\}}{N}}$$

Efficient testing:

$$\max_{k\in K}\ \max_{\ell\in L}\ \mathrm{HSIC}_{k,\ell} \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K||L|)\right\}}{N/B}}$$

Robust testing:

$$\max_{k\in K}\ \max_{\ell\in L}\ \mathrm{HSIC}_{k,\ell} \gtrsim \max\left\{\sqrt{\frac{\max\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K||L|)\}}{N}},\ \frac{r}{N}\right\}$$

Power in L2 Sobolev Metric: Oracle Kernels

Standard testing:

$$\|p_{xy} - p_x \otimes p_y\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\ \log(1/\beta)}{N}\right)^{2s/(4s+d)}$$

Efficient testing:

$$\|p_{xy} - p_x \otimes p_y\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\ \log(1/\beta)}{|\mathscr{D}|\,/\,N}\right)^{2s/(4s+d)}$$

Differentially private testing:

$\|p - q\|_{L^2} \gtrsim N^{-2s/(4s+d)}$ in low privacy regime $\xi \gtrsim N^{-(2s-d/2)/(4s+d)}$

$\|p - q\|_{L^2} \gtrsim \left(N^{3/2}\xi\right)^{-s/(2s+d)}$ in mid privacy regime $N^{-1/2} \lesssim \xi \lesssim N^{-(2s-d/2)/(4s+d)}$

$\|p - q\|_{L^2} \gtrsim \left(N\xi\right)^{-2s/(2s+d)}$ in high privacy regime $\xi \lesssim N^{-1/2}$

Power in L2 Sobolev Metric: Kernel Aggregation

Standard testing:

$$\|p_{xy} - p_x \otimes p_y\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\ \log(1/\beta)}{N/\ \log(\log(N))} \right)^{2s/(4s+d)}$$

Efficient testing:

$$\|p_{xy} - p_x \otimes p_y\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\ \log(1/\beta)}{(|\mathscr{D}|/N)/\log(\log(|\mathscr{D}|/N))} \right)^{2s/(4s+d)}$$

Goodness-of-fit Testing

Goodness-of-fit testing

Goodness-of-fit problem: Given

- model distribution P (*e.g.* via its score function),
- i.i.d. samples $X_1, \ldots, X_N$ from a distribution Q,

test whether $\mathscr{H}_0 \colon P = Q$ or $\mathscr{H}_1 \colon P \neq Q$.

General notation: $\mathscr{P}_0 = \{(P, Q) : P = Q\}$ and $\mathscr{P}_1 = \{(P, Q) : P \neq Q\}$

Testing with KSD: $\mathscr{H}_0 \colon \mathrm{KSD}_P(Q) = 0$ and $\mathscr{H}_1 \colon \mathrm{KSD}_P(Q) > 0$

Permutations do not work: goodness-of-fit does **not** test for exchangeability

Wild bootstrap: $\dfrac{1}{|\mathscr{D}|} \displaystyle\sum_{(i,j)\in\mathscr{D}} \varepsilon_i \varepsilon_j h(X_i, X_j)$, $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. Rademacher variables

Pointwise power / consistency: $\lim\limits_{N\to\infty} \mathbb{P}_Q\big(\text{reject } \mathscr{H}_0\big) = 1$ for fixed $P \neq Q$

Power in KSD Metric: Fixed Kernel

Standard testing:

$$\mathrm{KSD}_k \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta)\right\}}{N}}$$

Efficient testing:

$$\mathrm{KSD}_k \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta)\right\}}{N/B}}$$

Power in KSD Metric: Kernel Pooling

Standard testing:

$$\max_{k \in K} \mathrm{KSD}_k \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K|)\right\}}{N}}$$

Efficient testing:

$$\max_{k \in K} \mathrm{KSD}_k \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta),\ \log(|K|)\right\}}{N/B}}$$

Power in L2 Sobolev Metric: Oracle Kernel

Standard testing:

$$\|(\nabla \log p - \nabla \log q)\, q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\, \log(1/\beta)}{N} \right)^{2s/(4s+d)}$$

Efficient testing:

$$\|(\nabla \log p - \nabla \log q)\, q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\, \log(1/\beta)}{|\mathscr{D}|\, / N} \right)^{2s/(4s+d)}$$

Power in L2 Sobolev Metric: Kernel Aggregation

Standard testing:

$$\|(\nabla \log p - \nabla \log q)\, q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\, \log(1/\beta)}{N/\, \log(\log(N))} \right)^{2s/(4s+d)}$$

Efficient testing:

$$\|(\nabla \log p - \nabla \log q)\, q\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\, \log(1/\beta)}{\big(|\mathscr{D}|/N\big) / \log\big(\log(|\mathscr{D}|/N)\big)} \right)^{2s/(4s+d)}$$

Experiments

Experiments

Synthetic data:

- Perturbed uniform distributions
- Gaussian (mean shift, mixture, etc.)
- Gamma distribution
- Gaussian-Bernoulli Restricted Boltzmann Machine

Real-world data:

- MNIST $d = 28 \times 28 = 784$ (digit distribution, image shift, normalising flow model)
- CIFAR-10 $d = 32 \times 32 = 1024$ (image shift, CIFAR-10 vs CIFAR-10.1)
- Galaxy MNIST $d = 3 \times 64 \times 64 = 12288$ (types of galaxies)
- CelebA $d = 3 \times 178 \times 218 = 116412$ (men/women)
- IMDb movie reviews $d = 3330$ (sentiment analysis)

Level: under the null type I error is at level $\alpha = 0.05$

Power: outperform other state-of-the-art kernel hypothesis tests

Open Problems

Open Problems

- L2 separation upper bound for the HSIC test with $s \in \left(0, (d_X + d_Y)/4\right)$
- L2 separation lower bounds under differential privacy constraint
- L2 and kernel separation lower bounds under efficiency constraint
- L2 separation upper and lower bounds under the robustness constraint
- L2 separation lower bounds for goodness-of-fit testing
- Kernel separation rates for normalised pooling
- Kernel pooling and aggregation procedure for differential private tests
- KSD private and robust test constructions and uniform power separation

Any Questions?

Proof Strategies

Concentration results

Exponential concentration for statistic:

$$\mathbb{P}\left(\left| \widehat{\mathrm{Kdisc}} - \mathrm{Kdisc} \right| > t \right) \leq \exp\left(- C t^2 N \right)$$

$$\left| \widehat{\mathrm{Kdisc}} - \mathrm{Kdisc} \right| \lesssim \sqrt{\frac{1}{N} \log\left(\frac{1}{\beta} \right)} \quad \text{with probability } \geq 1 - \beta/2$$

Exponential concentration for bootstrapped statistic (quantile):

$$\mathbb{P}\left(\widehat{\mathrm{Kdisc}}_{\mathrm{boot}} > t \right) \leq \exp\left(- C t^2 N \right)$$

$$q_{1-\alpha} \lesssim \sqrt{\frac{1}{N} \log\left(\frac{1}{\alpha} \right)} \quad \text{with probability } \geq 1 - \beta/2$$

Proof Separation in Kernel Metric

Type II error control:

$$\mathbb{P}\left(\widehat{\text{Kdisc}} \leq q_{1-\alpha}\right)$$

$$\leq \mathbb{P}\left(\text{Kdisc} \leq q_{1-\alpha} + C_1\sqrt{\frac{1}{N}\log\left(\frac{1}{\beta}\right)}\right) + \frac{\beta}{2}$$

$$\leq \mathbb{P}\left(\text{Kdisc} \leq C_2\sqrt{\frac{1}{N}\log\left(\frac{1}{\alpha}\right)} + C_1\sqrt{\frac{1}{N}\log\left(\frac{1}{\beta}\right)}\right) + \beta$$

$$= \beta$$

when there is kernel separation:

$$\text{Kdisc} \gtrsim \sqrt{\frac{\max\left\{\log(1/\alpha),\ \log(1/\beta)\right\}}{N}}$$

L2 Testing

L2 testing: $\mathscr{H}_0 \colon \|\psi\|_{L^2} = 0$ vs $\mathscr{H}_1 \colon \|\psi\|_{L^2} > 0$

Two-sample: $\psi = p - q$

Independence: $\psi = p_{xy} - p_x \otimes p_y$

Goodness-of-fit: $\psi = (\nabla \log p - \nabla \log q)\, q$

Kernel Integral Transform: $\big(S_\lambda f\big)(y) = \displaystyle\int_{\mathbb{R}^d} k_\lambda(x, y)\, f(x)\, dx$

Linking Kdisc and L2:

$$\text{Kdisc} = \langle \psi, S_\lambda \psi \rangle_{L^2} = \frac{1}{2}\left(\|\psi\|_{L^2}^2 + \|S_\lambda \psi\|_{L^2}^2 - \|\psi - S_\lambda \psi\|_{L^2}^2 \right)$$

MMD example:

$$\langle \psi, S_\lambda \psi \rangle_{L^2} \;=\; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k_\lambda(x, y) \big(p(x) - q(x)\big) \big(p(y) - q(y)\big) dx dy \;=\; \text{MMD}_\lambda^2(P, Q)$$

Concentrations with bandwidth dependency

Exponential concentration for statistic:

$$\left| \widehat{\text{Kdisc}} - \text{Kdisc} \right| \lesssim \frac{1}{2}\|S_\lambda \psi\|_{L^2}^2 + \frac{1}{N}\log\left(\frac{1}{\beta}\right)$$

with probability $\geq 1 - \beta/2$

Exponential concentration for bootstrapped statistic (quantile):

$$q_{1-\alpha} \lesssim \frac{1}{N\lambda^{d/2}}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{1}{\beta}\right)$$

with probability $\geq 1 - \beta/2$

Proof Separation in L2 Metric

Type II error control:

$$\mathbb{P}\left(\widehat{\text{Kdisc}} \leq q_{1-\alpha}\right)$$

$$\leq \mathbb{P}\left(\text{Kdisc} \leq \frac{1}{2}\|S_\lambda \psi\|_{L^2}^2 + C\frac{1}{N\lambda^{d/2}}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{1}{\beta}\right)\right) + \beta$$

$$= \beta$$

when there is kernel separation:

$$\text{Kdisc} \;\geq\; \frac{1}{2}\|S_\lambda \psi\|_{L^2}^2 + \frac{C}{N\lambda^{d/2}}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{1}{\beta}\right)$$

equivalently for L2 separation: e.g. use $\text{Kdisc} = \frac{1}{2}\left(\|\psi\|_{L^2}^2 + \|S_\lambda \psi\|_{L^2}^2 - \|\psi - S_\lambda \psi\|_{L^2}^2\right)$

$$\|\psi\|_{L^2}^2 \;\geq\; \|\psi - S_\lambda \psi\|_{L^2}^2 + \frac{C}{N\lambda^{d/2}}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{1}{\beta}\right)$$

Proof Separation in L2 Metric

Power is guaranteed when:

$$\|\psi\|_{L^2}^2 \geq \|\psi - S_\lambda \psi\|_{L^2}^2 + \frac{C}{N\lambda^{d/2}} \log\left(\frac{1}{\alpha}\right) \log\left(\frac{1}{\beta}\right)$$

Smoothness: Assuming ψ is s-Sobolev smooth: $\int_{\mathbb{R}^d} \|\xi\|_2^{2s} \, |\widehat{\psi}(\xi)|^2 \, \mathrm{d}\xi \leq (2\pi)^d$

$$\|\psi - S_\lambda \psi\|_{L^2}^2 \lesssim \lambda^{2s}$$

Optimal bandwidth: $\lambda = \left(\log(1/\alpha)\log(1/\beta) \,/\, N\right)^{2/(4s+d)}$

Power is guaranteed when:

$$\|\psi\|_{L^2} \gtrsim \left(\frac{\log(1/\alpha)\log(1/\beta)}{N}\right)^{2s/(4s+d)}$$

Proofs under testing constraints

Computational efficiency:

Keep track of design size $|\mathscr{D}|$

Differential privacy:

Keep track of privacy parameters ε and δ

Robustness:

Keep track of robustness parameter r

Kernel Pooling Asymptotic Validity

Kernel Pooling Asymptotic Validity (KSD)

Known: $\widehat{\mathrm{Kdisc}}^{(k)}_{\mathrm{wild}} \xrightarrow[n\to\infty]{D} \widehat{\mathrm{Kdisc}}^{(k)}$ for any fixed kernel k

Cramér–Wold Theorem: a sequence of d-dimensional variables $Z^{(n)} \xrightarrow{D} Z = (Z_1, \ldots, Z_d)$ if and only if

$$\sum_{i=1}^{d} t_i Z_i^{(n)} \xrightarrow[n\to\infty]{D} \sum_{i=1}^{d} t_i Z_i \qquad \text{for every } (t_1, \ldots, t_d) \in \mathbb{R}^d$$

Our setting:

$$\sum_{i=1}^{K} t_i \widehat{\mathrm{Kdisc}}^{(k_i)}_{\mathrm{wild}} = \sum_{i=1}^{K} t_i \frac{1}{N(N-1)} \sum_{1\le j\neq j'\le N} \epsilon_j \epsilon_{j'} k_i(X_j, X_{j'}) = \frac{1}{N(N-1)} \sum_{1\le j\neq j'\le N} \epsilon_j \epsilon_{j'} \left(\sum_{i=1}^{K} t_i k_i(X_j, X_{j'}) \right)$$

$$= \widehat{\mathrm{Kdisc}}^{(\sum_{i=1}^{K} t_i k_i)}_{\mathrm{wild}} \xrightarrow[N\to\infty]{D} \widehat{\mathrm{Kdisc}}^{(\sum_{i=1}^{K} t_i k_i)} = \sum_{i=1}^{K} t_i \widehat{\mathrm{Kdisc}}^{(k_i)}$$

Cramér–Wold Theorem: $\underset{k\in K}{\mathrm{joint}} \left(\widehat{\mathrm{Kdisc}}^{(k)}_{\mathrm{wild}} \right) \xrightarrow[N\to\infty]{D} \underset{k\in K}{\mathrm{joint}} \left(\widehat{\mathrm{Kdisc}}^{(k)} \right)$

Continuous Mapping Theorem: for any continuous pooling function (e.g. mean, max, fuse)

$$\underset{k\in K}{\mathrm{pool}} \left(\widehat{\mathrm{Kdisc}}^{(k)}_{\mathrm{wild}} \right) \xrightarrow[N\to\infty]{D} \underset{k\in K}{\mathrm{pool}} \left(\widehat{\mathrm{Kdisc}}^{(k)} \right)$$

Any Questions?