MMD Aggregated Two-Sample Test & KSD Aggregated Goodness-of-fit Test

MMDAgg: Antonin Schrab I. Kim M. Albert B. Laurent B. Guedj A. Gretton

KSDAgg: Antonin Schrab B. Guedj A. Gretton

UCL

UCL CENTRE FOR ARTIFICIAL INTELLIGENCE

GATSBY

Overview

MMDAgg: Theoretical contributions

- aggregate MMD tests with different kernel bandwidths: no data splitting
- minimax adaptive over Sobolev balls: general kernels, estimated quantiles

Background

Two-sample problem

samples $\mathbb{X}_m := (X_1, \dots, X_m)$ $\quad X_i \overset{\text{iid}}{\sim} p$ in $\mathbb{R}^d$

samples $\mathbb{Y}_n := (Y_1, \dots, Y_n)$ $\quad Y_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

$\mathcal{H}_0\colon p = q$ against $\mathcal{H}_a\colon p \neq q$

Maximum Mean Discrepancy

Kernel: $$k_\lambda(x,y) := \prod_{i=1}^d \frac{1}{\lambda_i} K_i\left(\frac{x_i - y_i}{\lambda_i}\right)$$

$$\mathrm{MMD}^2_\lambda(p,q) := \mathbb{E}_{p,p}[k_\lambda(X,X')] - 2\,\mathbb{E}_{p,q}[k_\lambda(X,Y)] + \mathbb{E}_{q,q}[k_\lambda(Y,Y')]$$

$$\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) := \frac{1}{m(m-1)}\sum_{1\le i\ne i'\le m} k_\lambda(X_i,X_{i'}) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k_\lambda(X_i,Y_j) + \frac{1}{n(n-1)}\sum_{1\le j\ne j'\le n} k_\lambda(Y_j,Y_{j'})$$

MMD test for fixed bandwidth λ

$$\Delta^\lambda_\alpha(\mathbb{X}_m,\mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-\alpha}\right)$$

Quantile: $\widehat{q}^{\,\lambda}_{1-\alpha}$ is the $\lceil (B+1)(1-\alpha)\rceil$-th largest value of $\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n)$ and B $\mathcal{H}_0$-simulated test statistics

Permutations: $\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}^\sigma_m,\mathbb{Y}^\sigma_n) \qquad (\mathbb{X}^\sigma_m,\mathbb{Y}^\sigma_n) = \sigma(\mathbb{X}_m \cup \mathbb{Y}_n)$

Aggregation

MMDAgg for collection of bandwidths Λ

$$\Delta^\Lambda_\alpha(\mathbb{X}_m,\mathbb{Y}_n) := \mathbb{1}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) > \widehat{q}^{\,\lambda}_{1-u_\alpha w_\lambda} \text{ for some } \lambda\in\Lambda\right)$$

with positive weights $(w_\lambda)_{\lambda\in\Lambda}$ satisfying $\sum_{\lambda\in\Lambda} w_\lambda \le 1$ and correction

$$u_\alpha = \sup\left\{u>0 : \mathbb{P}_{p\times p}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{MMD}}^2_\lambda(\mathbb{X}_m,\mathbb{Y}_n) - \widehat{q}^{\,\lambda}_{1-uw_\lambda}\right) > 0\right) \le \alpha\right\}$$

Theory

Minimax adaptivity over Sobolev balls

$$\mathcal{S}^s_d(R) := \left\{ f\in L^1(\mathbb{R}^d)\cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d}\|\xi\|_2^{2s}|\widehat{f}(\xi)|^2\,\mathrm{d}\xi \le (2\pi)^d R^2\right\}$$

Minimax rate over Sobolev balls: $(m+n)^{-2s/(4s+d)}$

$$\Lambda^* := \left\{2^{-\ell}\mathbf{1}_d : \ell\in\left\{1,\dots,\left\lceil\frac{2}{d}\log_2\left(\frac{m+n}{\ln(\ln(m+n))}\right)\right\rceil\right\}\right\} \qquad w_\lambda := \frac{6}{\pi^2\ell^2}$$

Assuming $p-q\in\mathcal{S}^s_d(R)$, the condition

$$\|p-q\|_2 \;\ge\; C\left(\frac{m+n}{\ln(\ln(m+n))}\right)^{-2s/(4s+d)}$$

guarantees control over the probability of type II error of MMDAgg

$$\mathbb{P}_{p\times q}\left(\Delta^{\Lambda^*}_\alpha(\mathbb{X}_m,\mathbb{Y}_n)=0\right) \;\le\; \beta.$$

KSDAgg: Theoretical contributions

- aggregate KSD tests with different kernel bandwidths: no data splitting
- uniform separation rate upper bound: general kernels, estimated quantiles

Goodness-of-fit problem

model: density p, score $\nabla \log p(z)$

samples $\mathbb{Z}_n := (Z_1,\dots,Z_n)$ $\quad Z_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

$\mathcal{H}_0\colon p = q$ against $\mathcal{H}_a\colon p\neq q$

Kernel Stein Discrepancy

$$h_{p,\lambda}(x,y) := \left(\nabla\log p(x)^\top\nabla\log p(y)\right)k_\lambda(x,y) + \nabla\log p(y)^\top\nabla_1 k_\lambda(x,y) + \nabla\log p(x)^\top\nabla_2 k_\lambda(x,y) + \textstyle\sum_{i=1}^d \frac{\partial}{\partial x_i\partial y_i}k_\lambda(x,y)$$

Stein identity: $$\mathbb{E}_p[h_{p,\lambda}(Z,\cdot)] = 0$$

$$\mathrm{KSD}^2_{p,\lambda}(q) := \mathrm{MMD}^2_{h_{p,\lambda}}(p,q) = \mathbb{E}_{q,q}[h_{p,\lambda}(Z,Z')]$$

$$\widehat{\mathrm{KSD}}^2_{p,\lambda}(\mathbb{Z}_n) := \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} h_{p,\lambda}(Z_i,Z_j)$$

KSD test for fixed bandwidth λ

$$\Delta^\lambda_\alpha(\mathbb{Z}_n) := \mathbb{1}\left(\widehat{\mathrm{KSD}}^2_{p,\lambda}(\mathbb{Z}_n) > \widehat{q}^{\,\lambda}_{1-\alpha}\right)$$

Quantile: $\widehat{q}^{\,\lambda}_{1-\alpha}$ is $\lceil B(1-\alpha)\rceil$-th largest of the B $\mathcal{H}_0$-simulated test statistics

Wild bootstrap: $$\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\epsilon_i\epsilon_j h_{p,\lambda}(Z_i,Z_j) \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{Unif}\{-1,1\}$$

Parametric bootstrap: $$\frac{1}{N(N-1)}\sum_{1\le i\ne j\le N} h_{p,\lambda}(\widetilde{Z}_i,\widetilde{Z}_j) \qquad \widetilde{Z}_i \overset{\text{iid}}{\sim} p$$

KSDAgg for collection of bandwidths Λ

$$\Delta^\Lambda_\alpha(\mathbb{Z}_n) := \mathbb{1}\left(\widehat{\mathrm{KSD}}^2_{p,\lambda}(\mathbb{Z}_n) > \widehat{q}^{\,\lambda}_{1-u_\alpha w_\lambda} \text{ for some } \lambda\in\Lambda\right)$$

with positive weights $(w_\lambda)_{\lambda\in\Lambda}$ satisfying $\sum_{\lambda\in\Lambda} w_\lambda\le 1$ and correction

$$u_\alpha = \sup\left\{u>0 : \mathbb{P}_{p\times p}\left(\max_{\lambda\in\Lambda}\left(\widehat{\mathrm{KSD}}^2_{p,\lambda}(\mathbb{Z}_n) - \widehat{q}^{\,\lambda}_{1-uw_\lambda}\right) > 0\right)\le\alpha\right\}$$

Uniform separation rate

Integral transform: $$(\kappa\diamond f)(y) := \int_{\mathbb{R}^d}\kappa(x,y)f(x)\,\mathrm{d}x$$

Kernel assumption: $$A_\lambda := \mathbb{E}_{q,q}[h_{p,\lambda}(Z,Z')^2] < \infty$$

The condition

$$\|p-q\|_2^2 \;\ge\; \min_{\lambda\in\Lambda}\left(\|(p-q) - h_{p,\lambda}\diamond(p-q)\|_2^2 + C\ln\left(\frac{1}{\alpha w_\lambda}\right)\frac{\sqrt{A_\lambda}}{\beta n}\right)$$

guarantees control over the probability of type II error of KSDAgg

$$\mathbb{P}_q\left(\Delta^\Lambda_{\alpha,p}(\mathbb{Z}_n)=0\right) \;\le\; \beta.$$

MMDAgg & KSDAgg: Experimental results

MMDAgg & KSDAgg obtain **higher power** than alternative state-of-the-art approaches to MMD & KSD kernel adaptation.

Median bandwidth: λ_{med} is the median L^2-distance between all the samples

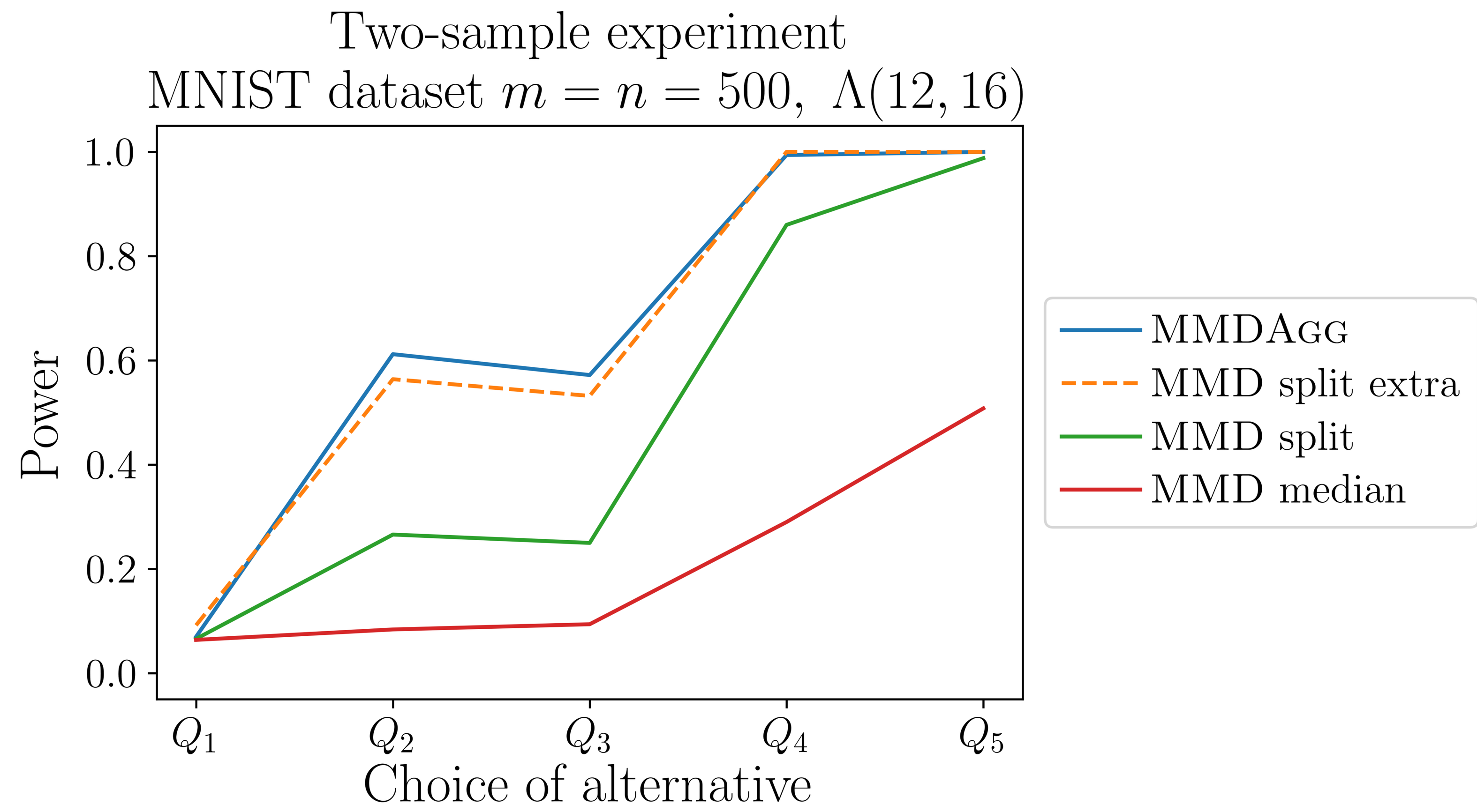
Collection of bandwidths with uniform weights:

$$\Lambda(\ell_-,\ell_+) := \left\{2^\ell\lambda_{med} : \ell\in\{\ell_-,\dots,\ell_+\}\right\} \qquad w_\lambda := 1/|\Lambda|$$

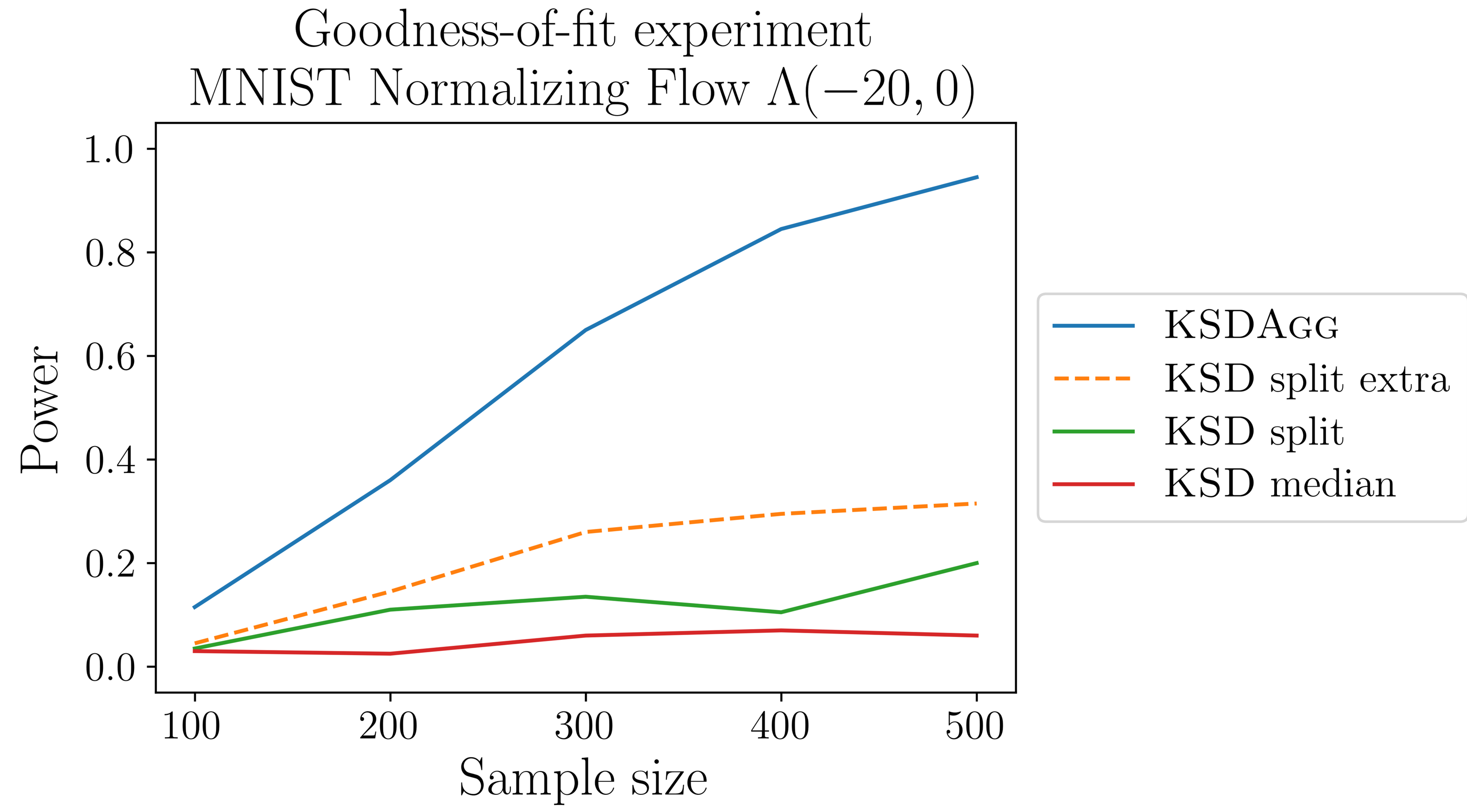
Split test: split the data in two parts

- **1st part:** select bandwidth λ^* maximizing a proxy for asymptotic power
- **2nd part:** perform test with selected bandwidth λ^*

Split extra test: select bandwidth λ^* using extra data



$\mathcal{P} := \{0,\dots,9\}$ $\quad \mathcal{Q}_1 := \mathcal{P}\setminus\{8\}$ $\quad \mathcal{Q}_2 := \mathcal{P}\setminus\{8,6\}$ $\quad \mathcal{Q}_3 := \mathcal{P}\setminus\{8,6,4\}$ $\quad \mathcal{Q}_4 := \mathcal{P}\setminus\{8,6,4,2\}$ $\quad \mathcal{Q}_5 := \mathcal{P}\setminus\{8,6,4,2,0\}$



model: Normalizing Flow density $\qquad$ samples: true MNIST digits

MMDAgg: paper, code

KSDAgg: paper, code