Robust Kernel Hypothesis Testing under Data Corruption

Antonin Schrab*
AI Centre, Gatsby Unit, Inria London
University College London, UK

Ilmun Kim*
Department of Statistics & Data Science
Yonsei University, South Korea

AISTATS 2025, Mai Khao, Thailand

*equal contributions

General Robust Testing Framework

- Space of distributions: $\mathscr{P}$ partitioned into disjoint $\mathscr{P}_0$ and $\mathscr{P}_1$

General Robust Testing Framework

- **Space of distributions:** $\mathscr{P}$ partitioned into disjoint $\mathscr{P}_0$ and $\mathscr{P}_1$
- **(Abstract) Goal:** given data related to some fixed $P \in \mathscr{P}$, determine whether

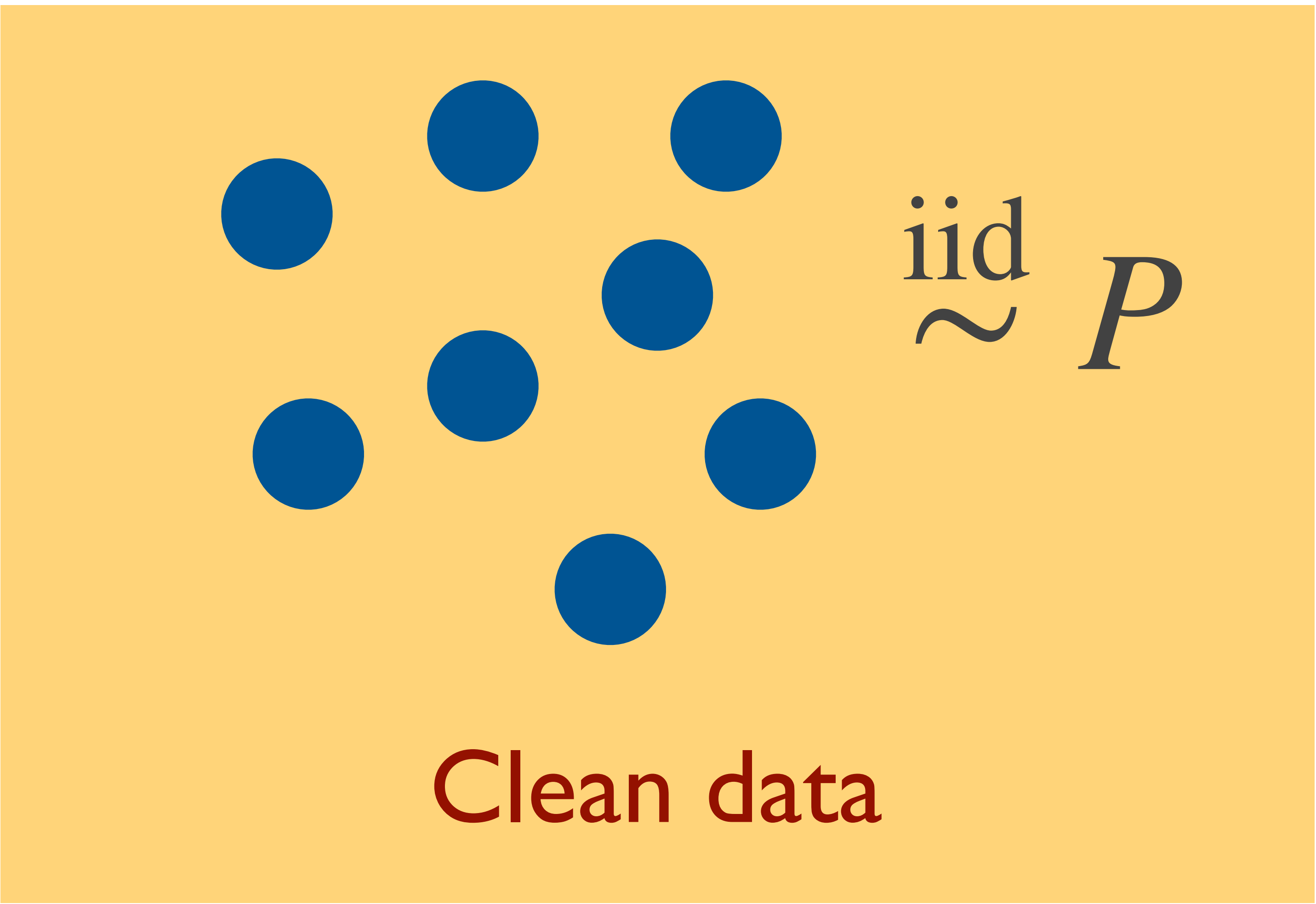
$$\mathscr{H}_0 : P \in \mathscr{P}_0 \quad \text{or} \quad \mathscr{H}_1 : P \in \mathscr{P}_1$$

General Robust Testing Framework

- Space of distributions: $\mathscr{P}$ partitioned into disjoint $\mathscr{P}_0$ and $\mathscr{P}_1$
- (Abstract) Goal: given data related to some fixed $P \in \mathscr{P}$, determine whether

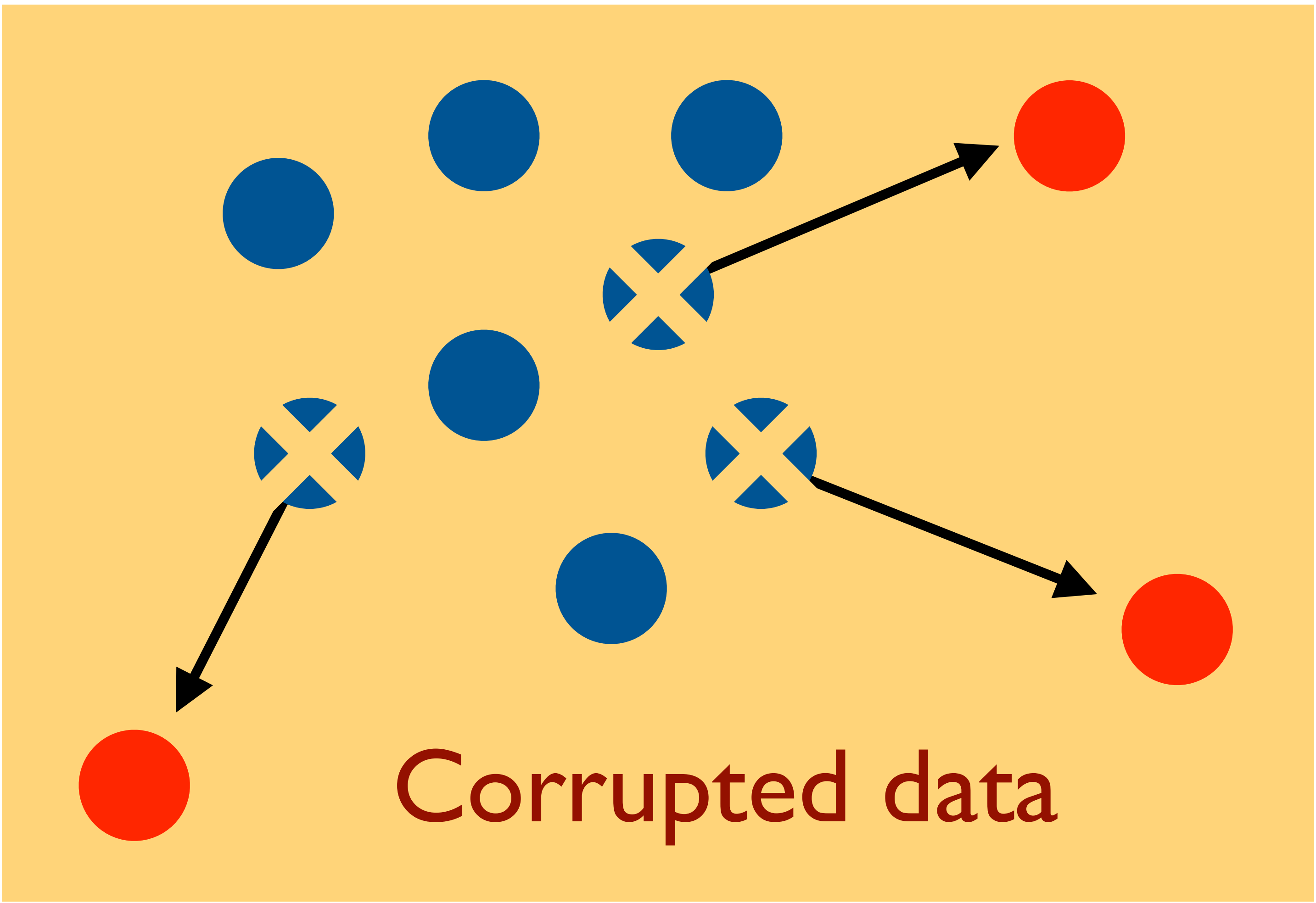
$$\mathscr{H}_0 : P \in \mathscr{P}_0 \quad \text{or} \quad \mathscr{H}_1 : P \in \mathscr{P}_1$$

Standard framework



vs

Robust framework



Up to r samples may have been corrupted arbitrarily

General Robust Testing Framework

- **Space of distributions:** $\mathscr{P}$ partitioned into disjoint $\mathscr{P}_0$ and $\mathscr{P}_1$
- **(Abstract) Goal:** given data related to some fixed $P \in \mathscr{P}$, determine whether

$$\mathscr{H}_0 : P \in \mathscr{P}_0 \quad \text{or} \quad \mathscr{H}_1 : P \in \mathscr{P}_1$$

- **(Specific) Goal:** given $\tilde{X}_1, \ldots, \tilde{X}_N$ related to some fixed $P \in \mathscr{P}$, determine whether

$\mathscr{H}_0 : \tilde{X}_1, \ldots, \tilde{X}_N$ are exchangeable or $\mathscr{H}_1 : \tilde{X}_1, \ldots, \tilde{X}_N$ are not exchangeable

General Robust Testing Framework

- **Space of distributions:** $\mathscr{P}$ partitioned into disjoint $\mathscr{P}_0$ and $\mathscr{P}_1$
- **(Abstract) Goal:** given data related to some fixed $P \in \mathscr{P}$, determine whether

$$\mathscr{H}_0 : P \in \mathscr{P}_0 \quad \text{or} \quad \mathscr{H}_1 : P \in \mathscr{P}_1$$

- **(Specific) Goal:** given $\tilde{X}_1, \ldots, \tilde{X}_N$ related to some fixed $P \in \mathscr{P}$, determine whether

$\mathscr{H}_0 : \tilde{X}_1, \ldots, \tilde{X}_N$ are exchangeable or $\mathscr{H}_1 : \tilde{X}_1, \ldots, \tilde{X}_N$ are not exchangeable

- **Challenge:** we don't observe $\tilde{X}_1, \ldots, \tilde{X}_N$ but observe $X_1, \ldots, X_N$ where

  - Up to r samples might have been corrupted arbitrarily
  - $N - r$ samples are from P

♣ **Robustness parameter r** is specified by the user depending on the application

DC Procedure for Robust Testing

- **Global sensitivity:** maximum possible change in T when one data point is arbitrarily changed

$$\Delta_T \coloneqq \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n : d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \leq 1} \big| T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}}) \big|$$

DC Procedure for Robust Testing

- **Global sensitivity:** maximum possible change in T when one data point is arbitrarily changed

$$\Delta_T := \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n : d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \leq 1} \left| T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}}) \right|$$

Input

Corrupted data
$X_1, \ldots, X_N$

DC Procedure for Robust Testing

- Global sensitivity: maximum possible change in T when one data point is arbitrarily changed

$$\Delta_T \coloneqq \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n : d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \leq 1} \left| T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}}) \right|$$

Input

Corrupted data
$X_1, \ldots, X_N$

For i in 1:B

Generate a permutation π of $[N]$
Compute $T_i = T(X_{\pi_1}, \ldots, X_{\pi_N})$

DC Procedure for Robust Testing

- Global sensitivity: maximum possible change in T when one data point is arbitrarily changed

$$\Delta_T \coloneqq \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n \,:\, d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \le 1} \left| T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}}) \right|$$

Input

Corrupted data
$X_1, \ldots, X_N$

For i in 1:B

Generate a permutation π of $[N]$
Compute $T_i = T(X_{\pi_1}, \ldots, X_{\pi_N})$

Quantile

Compute $(1 - \alpha)$ quantile q of
$T_0, T_1 \ldots, T_B$

DC Procedure for Robust Testing

- Global sensitivity: maximum possible change in T when one data point is arbitrarily changed

$$\Delta_T := \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n \,:\, d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \leq 1} \left| T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}}) \right|$$

Input

Corrupted data $X_1, \ldots, X_N$

For i in 1:B

Generate a permutation π of $[N]$

Compute $T_i = T(X_{\pi_1}, \ldots, X_{\pi_N})$

Quantile

Compute $(1 - \alpha)$ quantile q of $T_0, T_1 \ldots, T_B$

Output

Reject $\mathscr{H}_0$ if $T_0 > q + 2r\Delta_T$

DC Procedure for Robust Testing

- Global sensitivity: maximum possible change in T when one data point is arbitrarily changed

$$\Delta_T := \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n : d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \leq 1} \left| T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}}) \right|$$

Input

Corrupted data
$X_1, \ldots, X_N$

For i in 1:B

Generate a permutation π of $[N]$
Compute $T_i = T(X_{\pi_1}, \ldots, X_{\pi_N})$

Quantile

Compute $(1 - \alpha)$ quantile q of
$T_0, T_1 \ldots, T_B$

Output

Reject $\mathscr{H}_0$ if
$T_0 > q + 2r\Delta_T$

Adjustment factor for data corruption

DC Procedure for Robust Testing

- Global sensitivity: maximum possible change in T when one data point is arbitrarily changed

$$\Delta_T := \sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n : d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \leq 1} \left| T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}}) \right|$$

Input

Corrupted data
$X_1, \ldots, X_N$

For i in 1:B

Generate a permutation π of $[N]$
Compute $T_i = T(X_{\pi_1}, \ldots, X_{\pi_N})$

Quantile

Compute $(1-\alpha)$ quantile q of
$T_0, T_1 \ldots, T_B$

Output

Reject $\mathscr{H}_0$ if
$T_0 > q + 2r\Delta_T$

Adjustment factor for data corruption

We coin this as the "DC test":

A permutation test under data corruption

DC Procedure for Robust Testing

We prove two fundamental results for the DC test

- Level: the DC test is well-calibrated non-asymptotically

$$\mathbb{P}_{P_0}\big(\text{DC rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \ \leq \ \alpha \quad \begin{cases} \text{for any distribution } P_0 \in \mathscr{P}_0 \\ \text{for any value } \alpha \in (0,1) \\ \text{for any } N \geq 1 \end{cases}$$

DC Procedure for Robust Testing

We prove two fundamental results for the DC test

- Level: the DC test is well-calibrated non-asymptotically

$$\mathbb{P}_{P_0}\big(\text{DC rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \le \alpha \quad \begin{cases} \text{for any distribution } P_0 \in \mathscr{P}_0 \\ \text{for any value } \alpha \in (0,1) \\ \text{for any } N \ge 1 \end{cases}$$

- Consistency: the DC test is consistent in the sense that

$$\lim_{N\to\infty} \mathbb{P}_{P_1}\big(\text{DC rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) = 1 \text{ for any fixed distribution } P_1 \in \mathscr{P}_1$$

whenever $\lim_{N\to\infty} \mathbb{P}_{P_1}\big(T(\mathbb{X}_n) > T(\mathbb{X}_n^{\pi}) + 4r\Delta_T\big) = 1$

1. Two-Sample Testing
2. Independence Testing

1. $\boxed{\text{Two-Sample Testing}}$
2. Independence Testing

Robust Two-Sample Testing

- Two-sample problem: Given mutually independent
  - i.i.d. samples $\tilde{X}_1, \ldots, \tilde{X}_m$ from a distribution P
  - i.i.d. samples $\tilde{Y}_1, \ldots, \tilde{Y}_n$ from a distribution Q

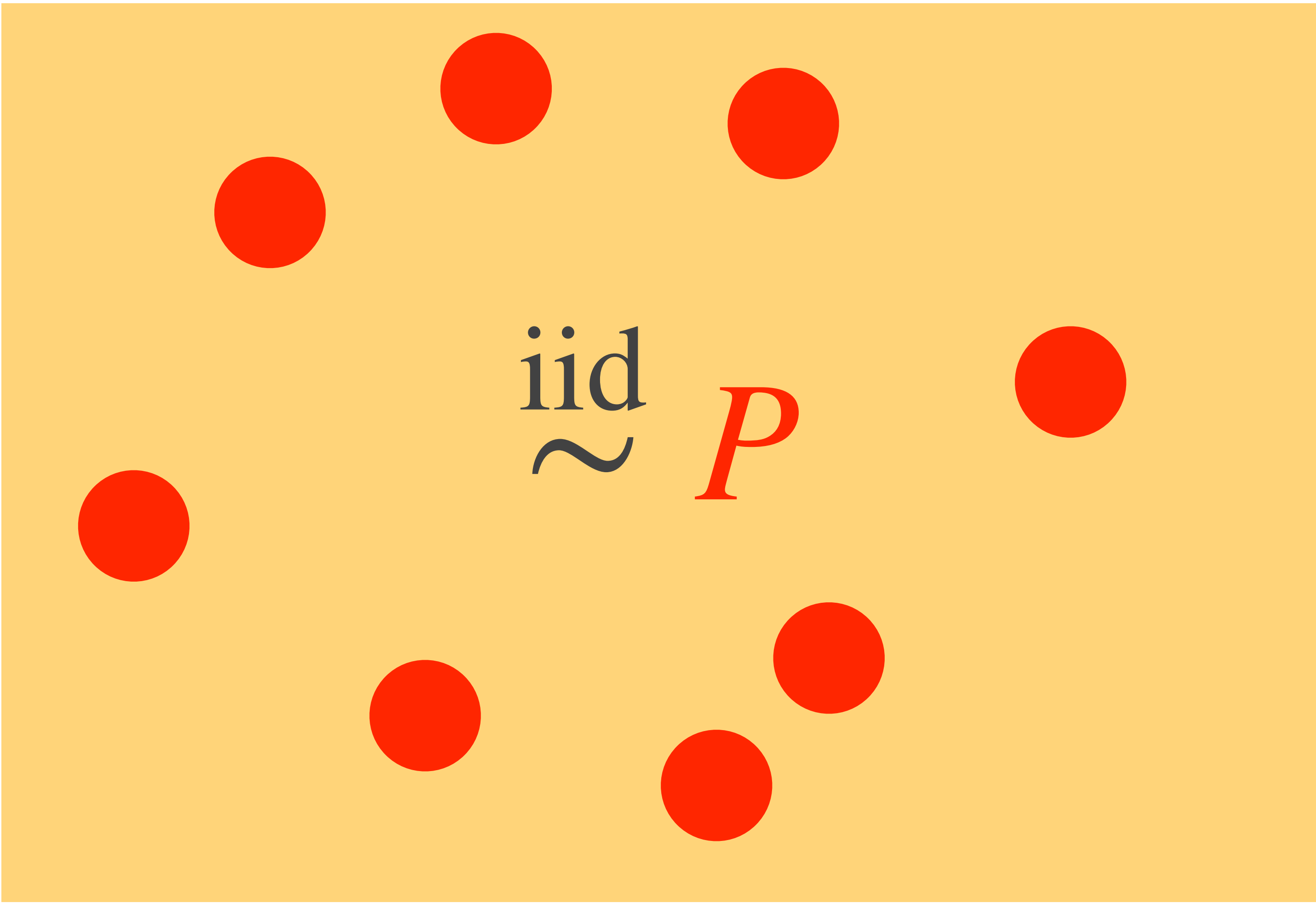
test whether $\mathscr{H}_0 : P = Q$ or $\mathscr{H}_1 : P \neq Q$

Robust Two-Sample Testing

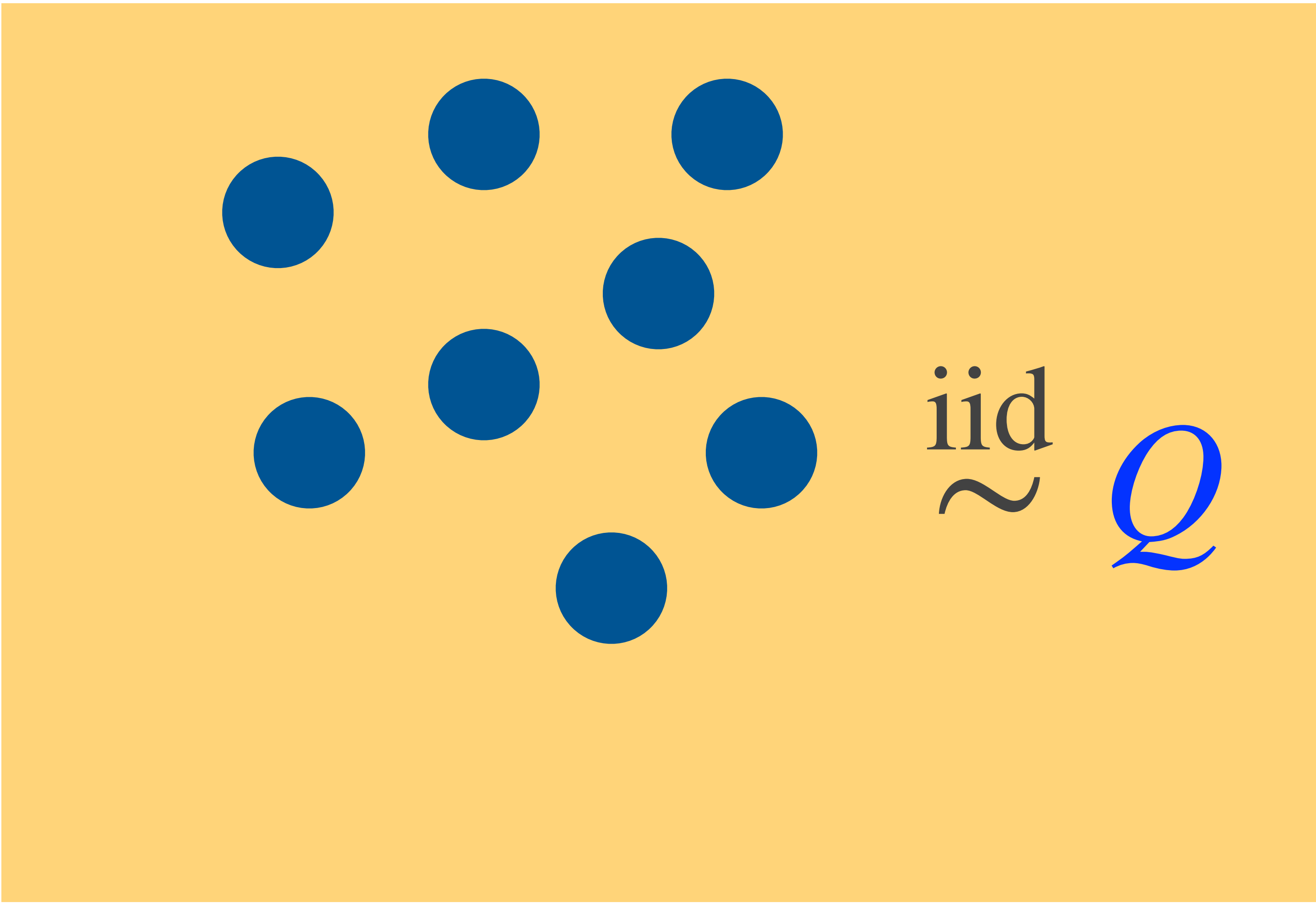
- Two-sample problem: Given mutually independent
  - i.i.d. samples $\tilde{X}_1, \ldots, \tilde{X}_m$ from a distribution P
  - i.i.d. samples $\tilde{Y}_1, \ldots, \tilde{Y}_n$ from a distribution Q

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

- Robust testing: Up to r samples from either P or Q can be corrupted



i.i.d. Samples from P



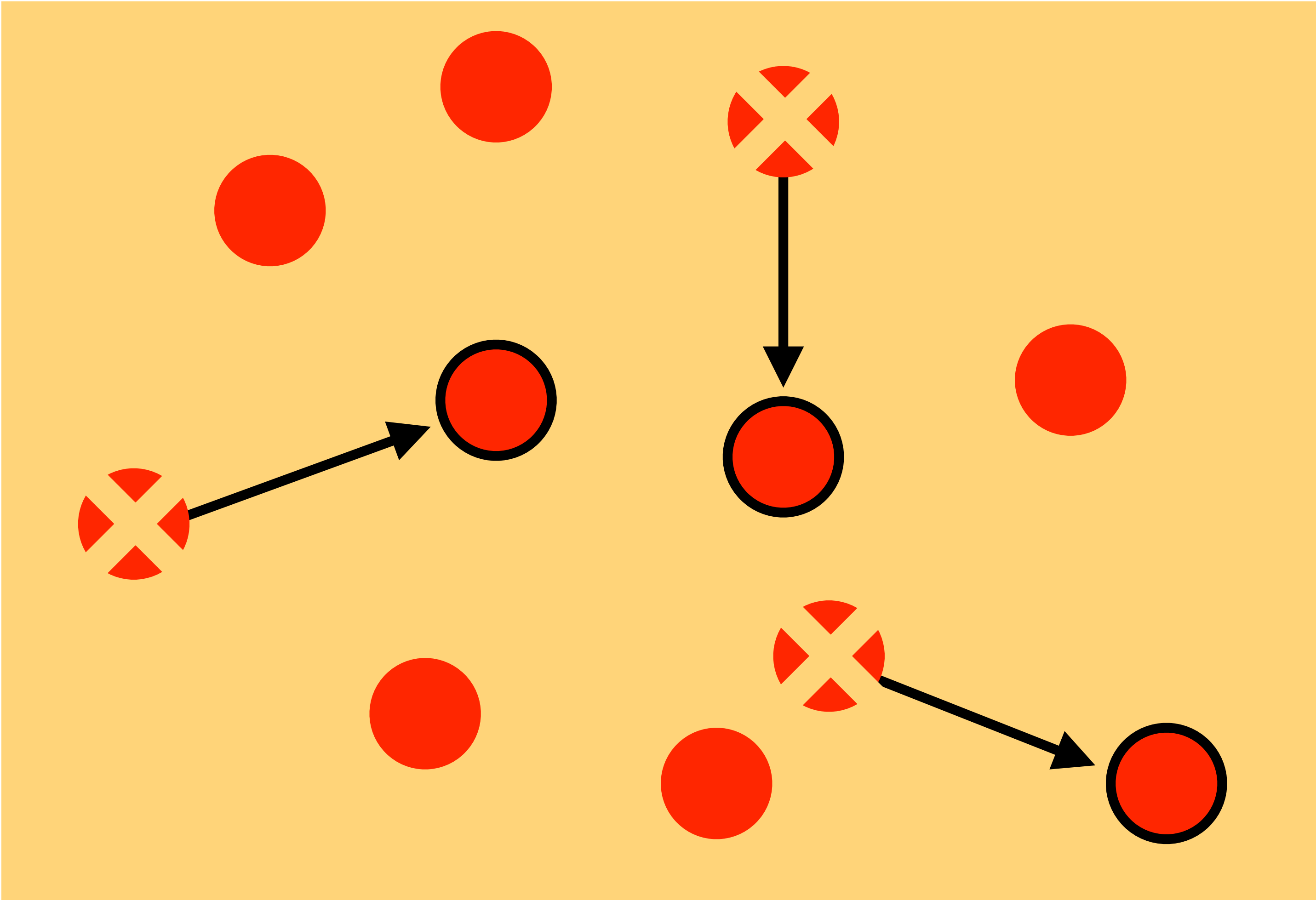
i.i.d. Samples from Q

Robust Two-Sample Testing

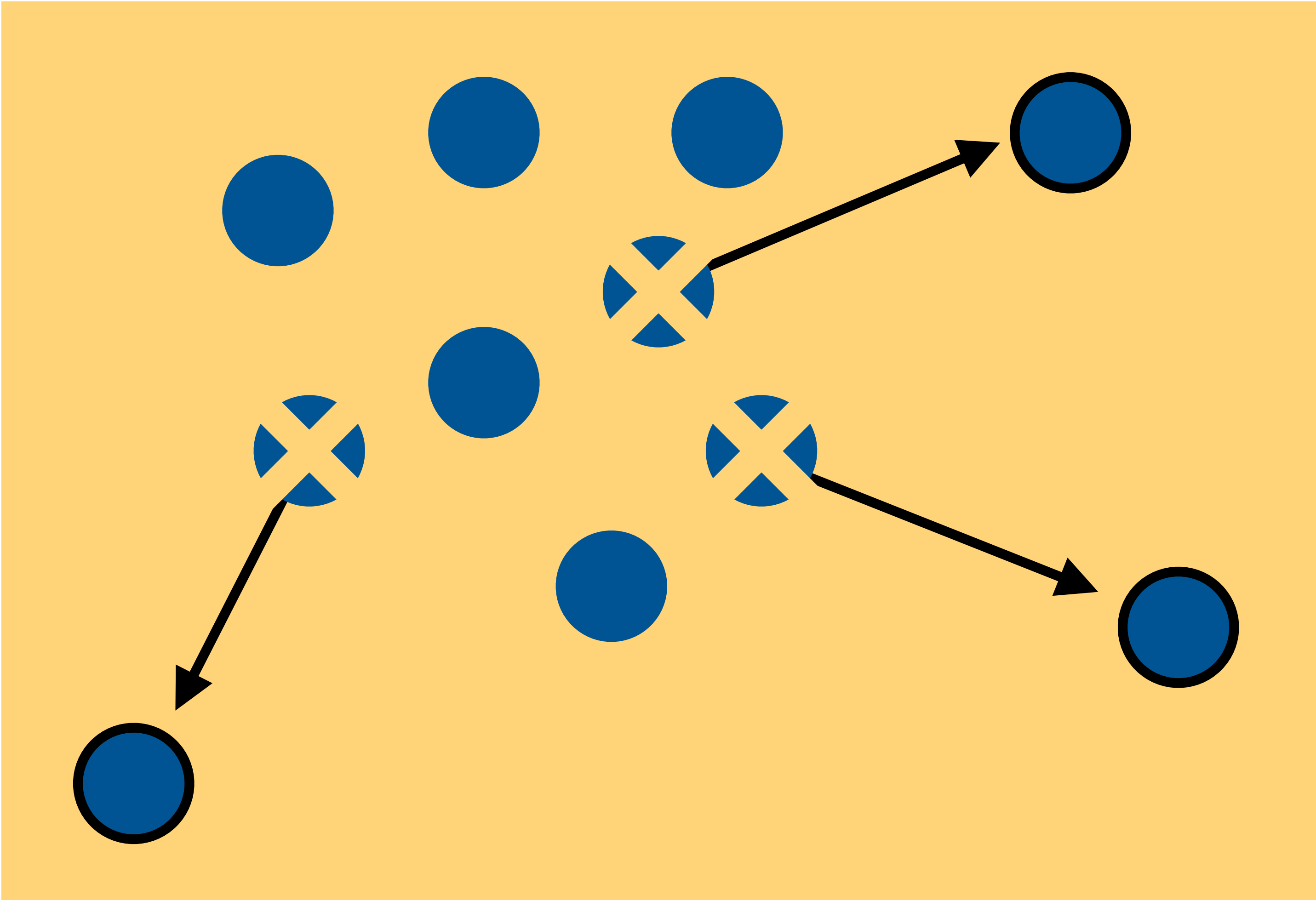
- Two-sample problem: Given mutually independent
  - i.i.d. samples $\tilde{X}_1, \ldots, \tilde{X}_m$ from a distribution P
  - i.i.d. samples $\tilde{Y}_1, \ldots, \tilde{Y}_n$ from a distribution Q

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

- Robust testing: Up to r samples from either P or Q can be corrupted



Corrupted Samples from P



Corrupted Samples from Q

dcMMD Procedure

- Maximum Mean Discrepancy:

$$\mathrm{MMD} = \sqrt{\mathbb{E}_{P,P}[k(X, X')] - 2\mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_{Q,Q}[k(Y, Y')]}$$

dcMMD Procedure

- Maximum Mean Discrepancy:

$$\mathrm{MMD} = \sqrt{\mathbb{E}_{P,P}[k(X, X')] - 2\mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_{Q,Q}[k(Y, Y')]}$$

- Statistic (plug-in estimator):

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

dcMMD Procedure

- Maximum Mean Discrepancy:

$$\mathrm{MMD} = \sqrt{\mathbb{E}_{P,P}[k(X, X')] - 2\mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_{Q,Q}[k(Y, Y')]}$$

- Statistic (plug-in estimator):

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

- Global sensitivity of $\widehat{\mathrm{MMD}}$: $\Delta_{\widehat{\mathrm{MMD}}} = \sqrt{2K}/N$ where K: kernel bound and $N = \min(m, n)$

dcMMD Procedure

- Maximum Mean Discrepancy:

$$\mathrm{MMD} = \sqrt{\mathbb{E}_{P,P}[k(X, X')] - 2\mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_{Q,Q}[k(Y, Y')]}$$

- Statistic (plug-in estimator):

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

- Global sensitivity of $\widehat{\mathrm{MMD}}$: $\Delta_{\widehat{\mathrm{MMD}}} = \sqrt{2K}/N$ where K: kernel bound and $N = \min(m, n)$
- dcMMD test: Apply DC procedure with $\widehat{\mathrm{MMD}}$ and $\Delta_{\widehat{\mathrm{MMD}}}$

dcMMD Guarantees

- **Level:** for any distribution P and any sample size

$$\mathbb{P}_{P,P}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\le\; \alpha$$

dcMMD Guarantees

- **Level:** for any distribution P and any sample size

$$\mathbb{P}_{P,P}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\le\; \alpha$$

- **Pointwise Power / Consistency:** for any fixed $P \neq Q$ and $r/N \to 0$

$$\lim_{m,n\to\infty} \mathbb{P}_{P,Q}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;=\; 1$$

dcMMD Guarantees

- Level: for any distribution P and any sample size

$$\mathbb{P}_{P,P}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\leq\; \alpha$$

- Pointwise Power / Consistency: for any fixed $P \neq Q$ and $r/N \to 0$

$$\lim_{m,n\to\infty} \mathbb{P}_{P,Q}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;=\; 1$$

- Uniform Power: for any distributions P and Q separated as

$$\mathrm{MMD}(P,Q) \;\gtrsim\; \max\left\{\sqrt{\frac{\max\{\log(1/\alpha), \log(1/\beta)\}}{\min(m,n)}}, \frac{r}{\min(m,n)}\right\}$$

dcMMD achieves high power

$$\mathbb{P}_{P,Q}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\geq\; 1-\beta$$

This rate is minimax optimal with respect to m, n, r, α, β.

dcMMD Guarantees

- **Level:** for any distribution P and any sample size

$$\mathbb{P}_{P,P}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\le\; \alpha$$

- **Pointwise Power / Consistency:** for any fixed $P \neq Q$ and $r/N \to 0$

$$\lim_{m,n\to\infty} \mathbb{P}_{P,Q}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;=\; 1$$

- **Uniform Power:** for any distributions P and Q separated as

$$\mathrm{MMD}(P,Q) \;\gtrsim\; \max\left\{\sqrt{\frac{\max\{\log(1/\alpha), \log(1/\beta)\}}{\min(m,n)}}, \frac{r}{\min(m,n)}\right\}$$

dcMMD achieves high power

$$\mathbb{P}_{P,Q}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\ge\; 1-\beta$$

This rate is minimax optimal with respect to m, n, r, α, β.

dcMMD Guarantees

- **Level:** for any distribution P and any sample size

$$\mathbb{P}_{P,P}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\le\; \alpha$$

- **Pointwise Power / Consistency:** for any fixed $P \neq Q$ and $r/N \to 0$

$$\lim_{m,n\to\infty} \mathbb{P}_{P,Q}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;=\; 1$$

- **Uniform Power:** for any distributions P and Q separated as

$$\mathrm{MMD}(P,Q) \;\gtrsim\; \max\left\{\sqrt{\frac{\max\{\log(1/\alpha),\, \log(1/\beta)\}}{\min(m,n)}},\, \frac{r}{\min(m,n)}\right\}$$

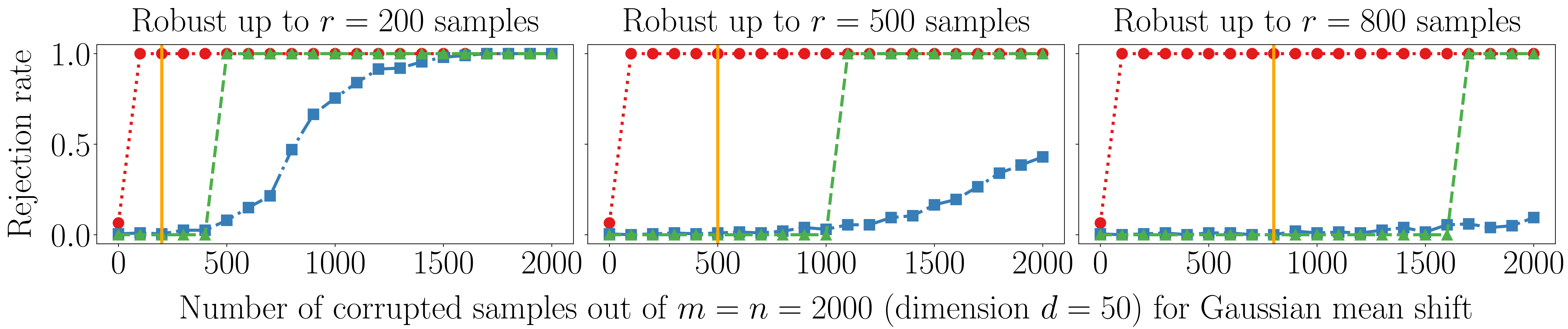
dcMMD achieves high power

$$\mathbb{P}_{P,Q}\big(\text{dcMMD rejects } \mathscr{H}_0 \mid r \text{ corrupted data}\big) \;\ge\; 1-\beta$$

This rate is minimax optimal with respect to m, n, r, α, β.

Experiments

dcMMD Experiments: Gaussian Mean Shift



- Generate two samples

  $\tilde{X}_1, \ldots, \tilde{X}_m \overset{\text{iid}}{\sim} N_d(0, 0.1)$

  $\tilde{Y}_1, \ldots, \tilde{Y}_n \overset{\text{iid}}{\sim} N_d(0, 0.1)$

- Corrupt one sample using

  $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} N_d(1000, 0.1)$

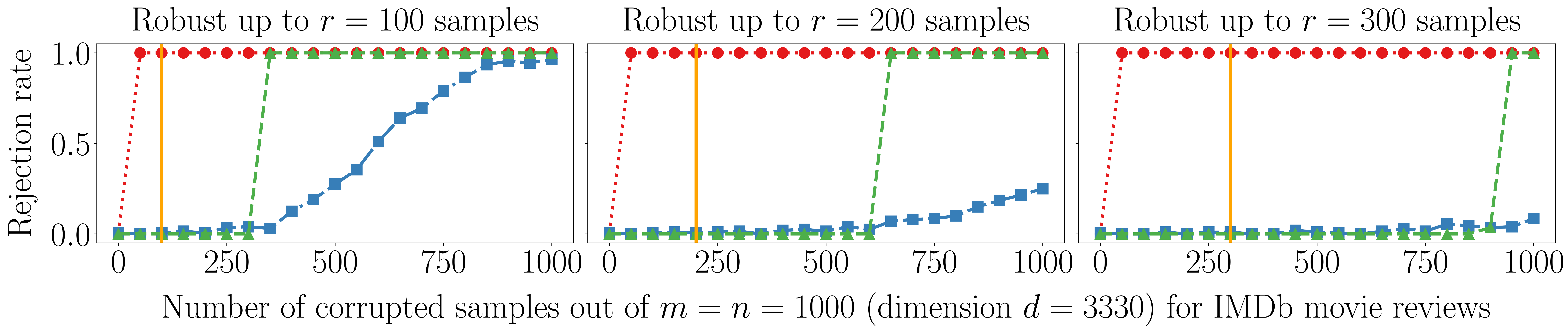
dcMMD: Our proposal

dpMMD: Procedure via differential privacy (Kim & Schrab)

MMD: Standard non-robust MMD

r : Robustness parameter

dcMMD Experiments: IMDb movie reviews



- Generate two samples

  $\tilde{X}_1, \ldots, \tilde{X}_m \overset{\text{iid}}{\sim}$ IMDb(3330)

  $\tilde{Y}_1, \ldots, \tilde{Y}_n \overset{\text{iid}}{\sim}$ IMDb(3330)

- Corrupt one sample using

  $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim}$ Geometric(3330)

dcMMD: Our proposal

dpMMD: Procedure via differential privacy (Kim & Schrab)

MMD: Standard non-robust MMD

r : Robustness parameter

Summary

Summary

- DC procedure: a general approach for constructing robust tests under data corruption
- Non-asymptotic validity and consistency under r data corruption
- Construct dcMMD and dcHSIC for two-sample and independence robust testing
- Prove that dcMMD/dcHSIC are minimax rate optimal
- Provide public implementations and illustrate the practicality

Any Question?

Paper:

Code: