

Differentially Private Maximal Information Coefficients

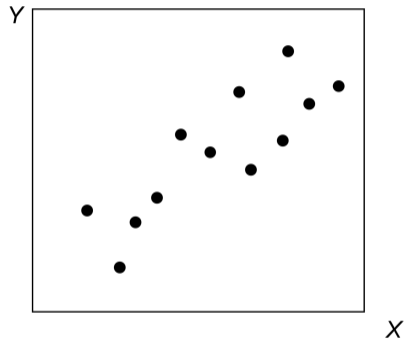
John Lazarsfeld¹, Aaron Johnson², Emmanuel Adeniran¹

¹Yale University ²U.S. Naval Research Laboratory

[ICML 2022]

Maximal Information Coefficients

Q: how correlated are X and Y ?

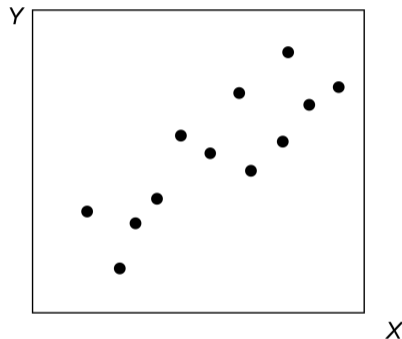


Maximal Information Coefficients

Q: how correlated are X and Y ?

MIC statistic [Reshef+11]:

- (i) find best $k \times l$ grid $G \rightarrow \mathbf{M}_{k,l}$
 \rightarrow *best*: largest normalized mutual info.
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$
 $\rightarrow B$: maximum grid size parameter.

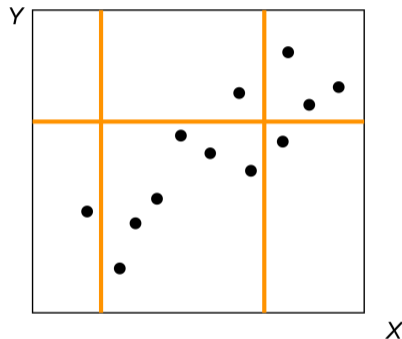


Maximal Information Coefficients

Q: how correlated are X and Y ?

MIC statistic [Reshef+11]:

- (i) find best $k \times l$ grid $G \rightarrow \mathbf{M}_{k,l}$
 \rightarrow *best*: largest normalized mutual info.
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$
 $\rightarrow B$: maximum grid size parameter.

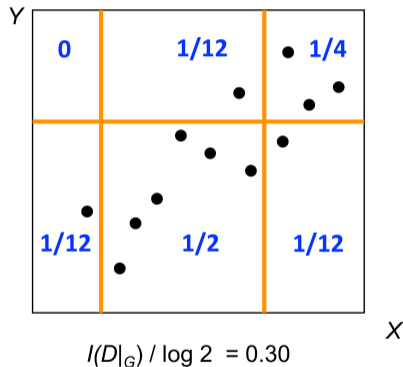


Maximal Information Coefficients

Q: how correlated are X and Y ?

MIC statistic [Reshef+11]:

- (i) find best $k \times l$ grid $G \rightarrow \mathbf{M}_{k,l}$
 \rightarrow *best*: largest normalized mutual info.
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$
 $\rightarrow B$: maximum grid size parameter.

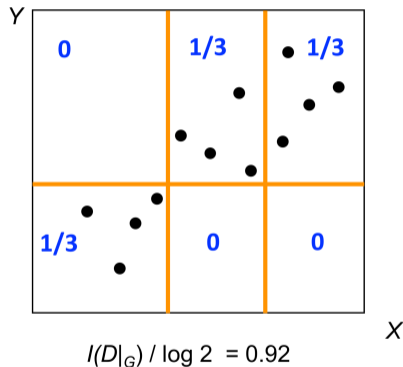


Maximal Information Coefficients

Q: how correlated are X and Y ?

MIC statistic [Reshef+11]:

- (i) find best $k \times l$ grid $G \rightarrow \mathbf{M}_{k,l}$
 \rightarrow *best*: largest normalized mutual info.
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$
 $\rightarrow B$: maximum grid size parameter.

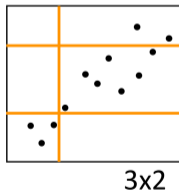
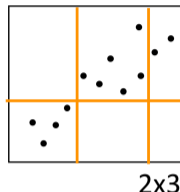
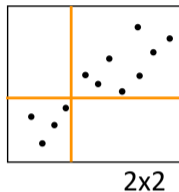


Maximal Information Coefficients

Q: how correlated are X and Y ?

MIC statistic [Reshef+11]:

- (i) find best $k \times l$ grid $G \rightarrow \mathbf{M}_{k,l}$
 \rightarrow *best*: largest normalized mutual info.
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$
 $\rightarrow B$: maximum grid size parameter.



$$\mathbf{M}_{2,2} = 0.92$$

$$\mathbf{M}_{2,3} = 0.92$$

$$\mathbf{M}_{3,2} = 0.81$$

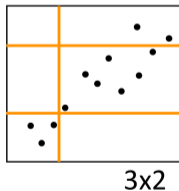
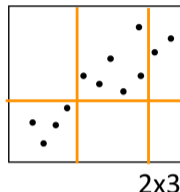
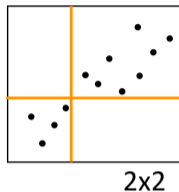
Maximal Information Coefficients

Q: how correlated are X and Y ?

MIC statistic [Reshef+11]:

- (i) find best $k \times l$ grid $G \rightarrow \mathbf{M}_{k,l}$
 \rightarrow *best*: largest normalized mutual info.
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$
 $\rightarrow B$: maximum grid size parameter.

Useful for data mining; but data may be sensitive!



$$\mathbf{M}_{2,2} = 0.92$$

$$\mathbf{M}_{2,3} = 0.92$$

$$\mathbf{M}_{3,2} = 0.81$$

Goal: differentially private approximations of MIC

Differentially Private MIC

ϵ -DP: for $D \sim D'$, need $A(D) \approx_{\epsilon} A(D')$

Laplace Mechanism: (1) bound sensitivity Δ (2) output $MIC(D) + \text{Lap}(\Delta/\epsilon)$

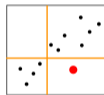
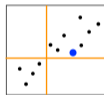
Differentially Private MIC

$$\epsilon\text{-DP: for } D \sim D', \text{ need } A(D) \approx_{\epsilon} A(D')$$

Laplace Mechanism: (1) bound sensitivity Δ (2) output $\text{MIC}(D) + \text{Lap}(\Delta/\epsilon)$

For MIC, sensitivity is small! \rightarrow for $|D| = n$, $\Delta = O(\log n/n)$

Intuition: changing one point
 \rightarrow small changes in new PMF



0	2/3
1/3	0

0	7/12
1/3	1/12

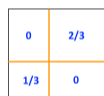
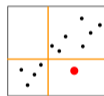
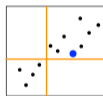
Differentially Private MIC

ϵ -DP: for $D \sim D'$, need $A(D) \approx_{\epsilon} A(D')$

Laplace Mechanism: (1) bound sensitivity Δ (2) output $\text{MIC}(D) + \text{Lap}(\Delta/\epsilon)$

For MIC, sensitivity is small! \rightarrow for $|D| = n$, $\Delta = O(\log n/n)$

Intuition: changing one point
 \rightarrow small changes in new PMF



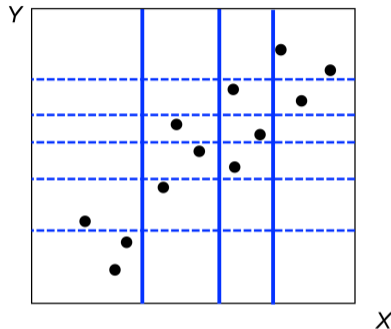
Issue: MIC is hard to compute. Solution(?) use MICe [Reshef+16]

\rightarrow MICe is (a) efficient to compute (b) consistent wrt MIC*

Differentially Private MICe (?)

MICe statistic [Reshef+16]:

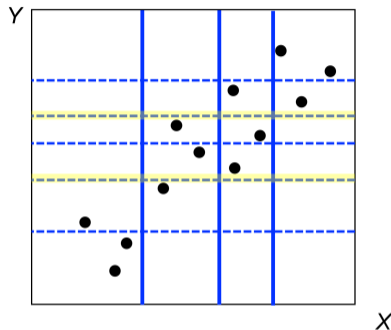
- (i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$
 - $\rightarrow P$ is mass equipartition
 - $\rightarrow Q \subseteq$ mass equipartition
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$



Differentially Private MICe (?)

MICe statistic [Reshef+16]:

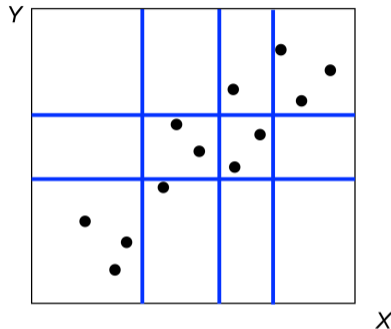
- (i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$
 - $\rightarrow P$ is mass equipartition
 - $\rightarrow Q \subseteq$ mass equipartition
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$



Differentially Private MICe (?)

MICe statistic [Reshef+16]:

- (i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$
 - $\rightarrow P$ is mass equipartition
 - $\rightarrow Q \subseteq$ mass equipartition
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$



Differentially Private MICe (?)

MICe statistic [Reshef+16]:

(i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$

$\rightarrow P$ is mass equipartition

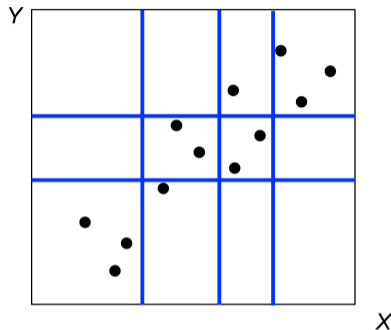
$\rightarrow Q \subseteq$ mass equipartition

(ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$

Constraint in (i): dataset dependent

\rightarrow fast to compute + consistent

\rightarrow larger bound on sensitivity Δ



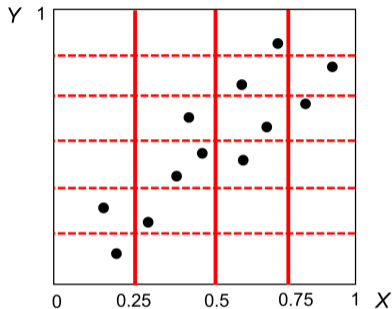
Thm: for MICe, $\Delta_e = O(B \log n/n)$

For typical $B=n^{1/a}$, $\Delta=\omega(\log n/n) \implies$ variance of $\text{Lap}(\Delta/\epsilon)$ too large!

New Statistic: MICr

MICr statistic:

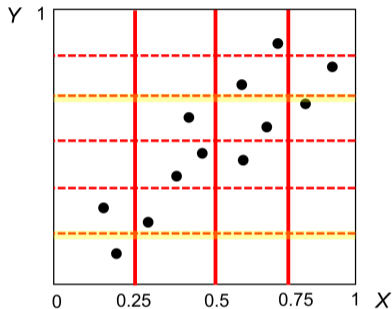
- (i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$
 - $\rightarrow P$ is range equipartition
 - $\rightarrow Q \subseteq$ range equipartition
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$



New Statistic: MICr

MICr statistic:

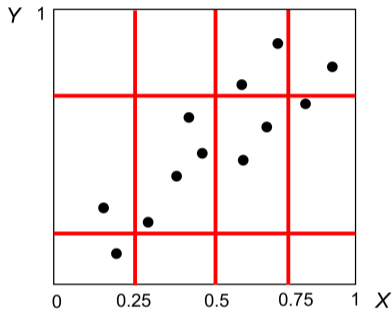
- (i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$
 - $\rightarrow P$ is range equipartition
 - $\rightarrow Q \subseteq$ range equipartition
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$



New Statistic: MICr

MICr statistic:

- (i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$
 - $\rightarrow P$ is range equipartition
 - $\rightarrow Q \subseteq$ range equipartition
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$

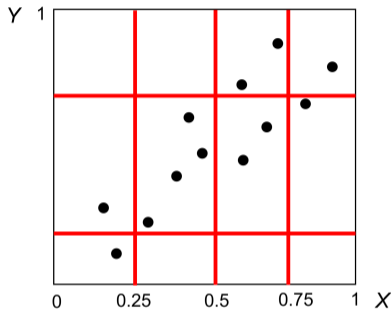


New Statistic: MICr

MICr statistic:

- (i) find best $k \times l$ grid $G=(P, Q) \rightarrow \mathbf{M}_{k,l}$
 - $\rightarrow P$ is range equipartition
 - $\rightarrow Q \subseteq$ range equipartition
- (ii) return $\max \mathbf{M}_{k,l}$ where $kl \leq B$

Constraint in (i): dataset independent
 \rightarrow requires bound on range of vars.
 \rightarrow still fast to compute



Thm: MICr is a consistent estimator (same as MICe)

Thm: for MICr, $\Delta_r = O(\log n/n)$ (same as MIC)

\implies MICr is a better base statistic for DP variants

Differentially Private MICr

Differentially Private MICr

MICr-Lap: *output perturbation*

$$[\text{MICr}(D) + \text{Lap}(\Delta_r/\epsilon)]_{0,1}$$

Thm: MICr-Lap \rightarrow MIC*

(since Δ_r vanishes with n)

Differentially Private MICr

MICr-Lap: *output perturbation*

$$[\text{MICr}(D) + \text{Lap}(\Delta_r/\epsilon)]_{0,1}$$

MICr-Geom: *input perturbation*

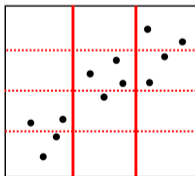
$\mathbf{M}_{k,\ell}$ computed using *noisy* counts
via Truncated Geometric [Ghosh+12]

Thm: MICr-Lap \rightarrow MIC*

(since Δ_r vanishes with n)

Thm: MICr-Geom \rightarrow MIC*

(error from noise vanishes with n)



Differentially Private MICr

MICr-Lap: *output perturbation*

$$[\text{MICr}(D) + \text{Lap}(\Delta_r/\epsilon)]_{0,1}$$

MICr-Geom: *input perturbation*

$\mathbf{M}_{k,\ell}$ computed using *noisy* counts
via Truncated Geometric [Ghosh+12]

0	0	2
0	3	2
2	1	0
2	0	0

Thm: MICr-Lap \rightarrow MIC*

(since Δ_r vanishes with n)

Thm: MICr-Geom \rightarrow MIC*

(error from noise vanishes with n)

Differentially Private MICr

MICr-Lap: *output perturbation*

$$[\text{MICr}(D) + \text{Lap}(\Delta_r/\epsilon)]_{0,1}$$

MICr-Geom: *input perturbation*

$\mathbf{M}_{k,\ell}$ computed using *noisy* counts
via Truncated Geometric [Ghosh+12]

Thm: MICr-Lap \rightarrow MIC*

(since Δ_r vanishes with n)

Thm: MICr-Geom \rightarrow MIC*

(error from noise vanishes with n)

0	0	2
0	3	2
2	1	0
2	0	0

→

1	0	2
0	2	3
1	0	0
4	0	1

Differentially Private MICr

MICr-Lap: *output perturbation*

$$[\text{MICr}(D) + \text{Lap}(\Delta_r/\epsilon)]_{0,1}$$

MICr-Geom: *input perturbation*

$\mathbf{M}_{k,\ell}$ computed using *noisy* counts
via Truncated Geometric [Ghosh+12]


Thm: MICr-Lap \rightarrow MIC*

(since Δ_r vanishes with n)

Thm: MICr-Geom \rightarrow MIC*

(error from noise vanishes with n)

0	0	2
0	3	2
2	1	0
2	0	0



1	0	2
0	2	3
1	0	0
4	0	1

Baseline: MICE-Lap Goal: empirical evaluation of utility

Experimental Results

Synthetic Data:

- 21 families of distributions [Reshef+11, 16]
- Tune hyperparams. for MICr DP variants
- B/V tradeoff for MICr-Lap / MICr-Geom

Real Data:

- *Spellman* and *Baseball* datasets [Reshef+11]
- MICr variants \gg MICE-Lap
- Useable error for MICr variants for $n \geq 4300$

