Machine learning meets false discovery rate

Ariane Marandon
Joint work with Lihua Lei, David Mary and Etienne Roquain

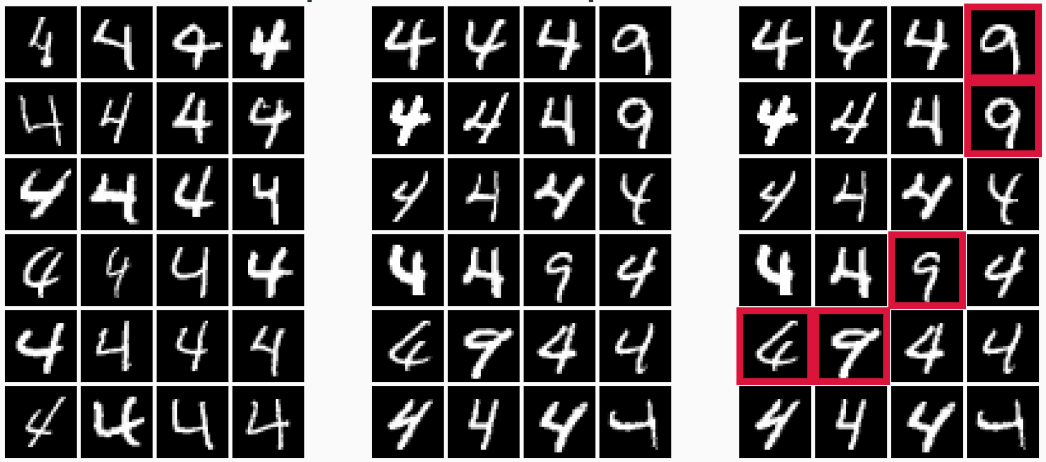
9 March 2023

LARGR Workshop, INRIA Lille

Motivation: novelty detection

- Nominal sample: $Y_1, \ldots, Y_n \sim P_0$
- Test sample: $X_1, \ldots, X_m$
  - m_0 nominals
  - m_1 novelties
- P_0 unknown
- No assumption on novelties

Nominal sample Test sample Procedure



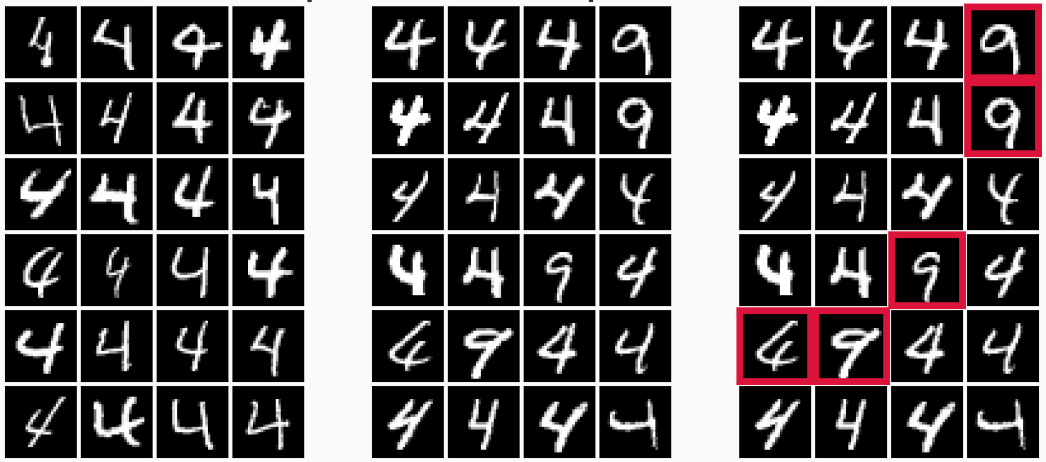
$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{discoveries}}\right] \leq \alpha$$

$$\text{TDR} = \mathbb{E}\left[\frac{\#\text{true discoveries}}{\#\text{possible discoveries}}\right] \quad \text{'large'}$$

Motivation: novelty detection

- Nominal sample: $Y_1, \ldots, Y_n \sim P_0$
- Test sample: $X_1, \ldots, X_m$
  - m_0 nominals
  - m_1 novelties
- P_0 unknown
- No assumption on novelties

Nominal sample Test sample Procedure



$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{discoveries}}\right] \leq \alpha$$

$$\text{TDR} = \mathbb{E}\left[\frac{\#\text{true discoveries}}{\#\text{possible discoveries}}\right] \quad \text{'large'}$$

Outline

Previous work: **Conformal Anomaly Detection**[1]

- Uses $Y_1, \ldots, Y_n$ to learn P_0
- Finite sample FDR control guarantee (!)

New: extension of previous work to **increase power**

- Leverages $X_1, \ldots, X_m$ to also learn P_1
- FDR control is retained
- Power analysis

[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Outline

Previous work: **Conformal Anomaly Detection**[1]

- Uses $Y_1, \ldots, Y_n$ to learn P_0
- Finite sample FDR control guarantee (!)

New: extension of previous work to **increase power**

- Leverages $X_1, \ldots, X_m$ to also learn P_1
- FDR control is retained
- Power analysis

[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Outline

Conformal Anomaly Detection[1]

- Nominal sample: $Y_1, \ldots, Y_n$
- Test sample $\mathcal{D}_{\text{test}}$: $X_1, \ldots, X_m$

General idea

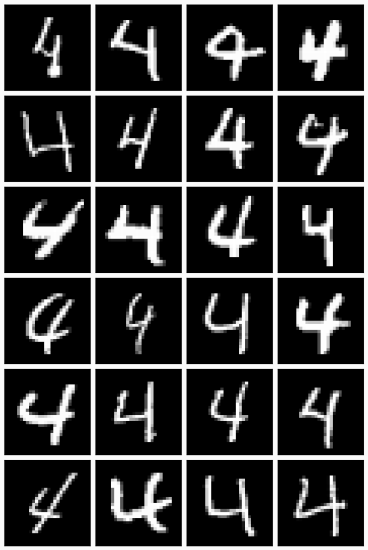
1. Split $Y_1, \ldots Y_n$ into $\underbrace{Y_1, \ldots, Y_k}_{\mathcal{D}_{\text{train}}}$ and $\underbrace{Y_{k+1}, \ldots, Y_n}_{\mathcal{D}_{\text{cal}}}$

[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

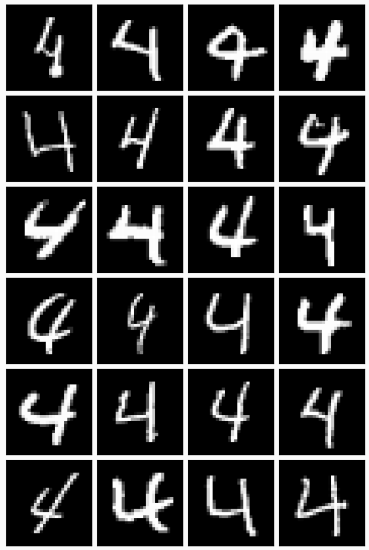
1. Split $Y_1, \dots Y_n$

Nominal sample

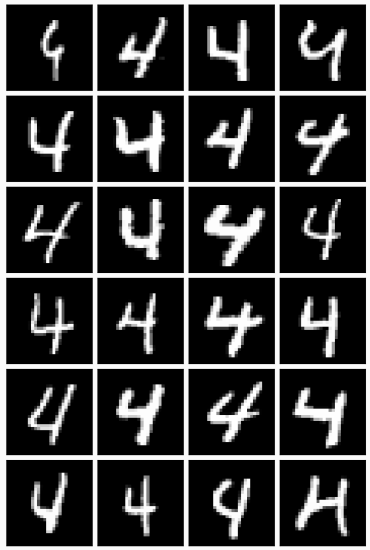


Split

$\mathcal{D}_{\text{train}}$



$\mathcal{D}_{\text{cal}}$



[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

- Nominal sample: $Y_1, \ldots, Y_n$
- Test sample $\mathcal{D}_{\text{test}}$: $X_1, \ldots, X_m$

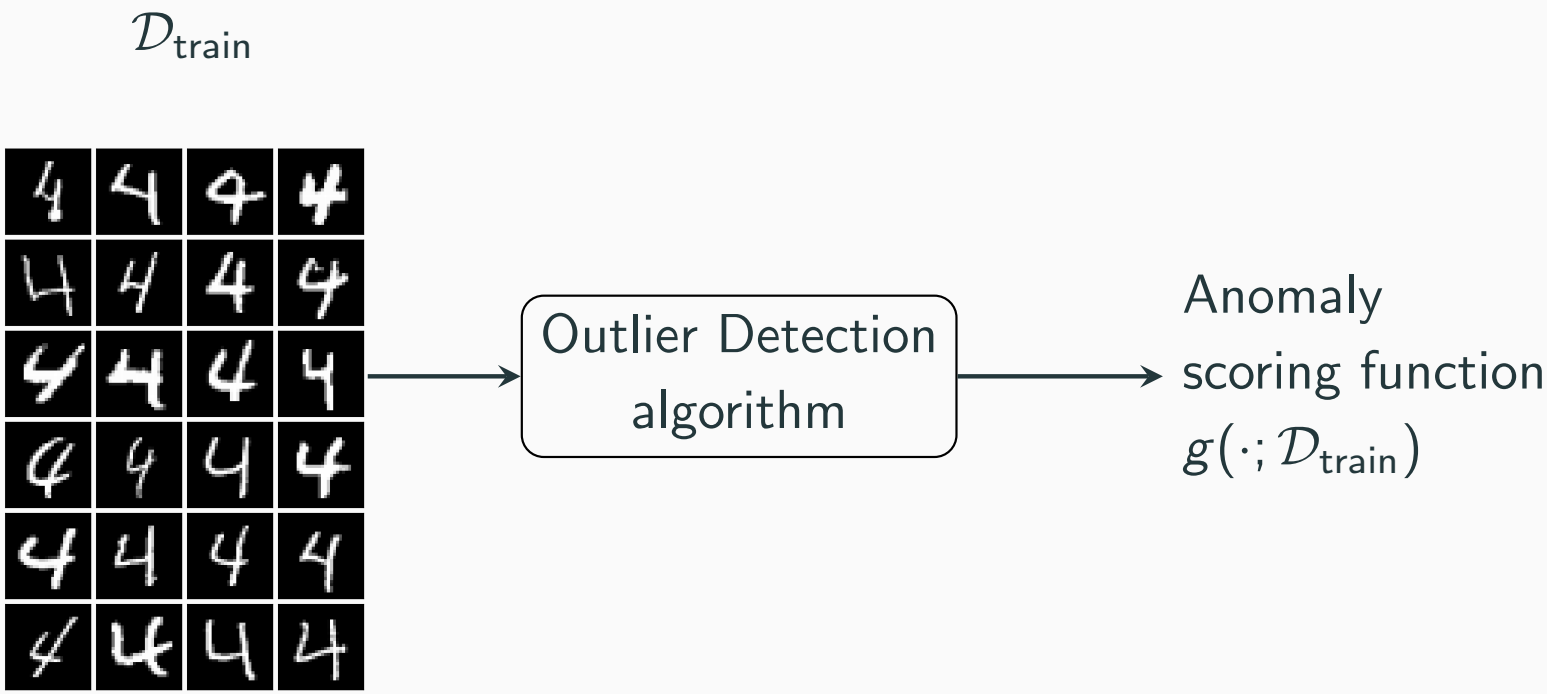
General idea

1. Split $Y_1, \ldots Y_n$ into $\underbrace{Y_1, \ldots, Y_k}_{\mathcal{D}_{\text{train}}}$ and $\underbrace{Y_{k+1}, \ldots, Y_n}_{\mathcal{D}_{\text{cal}}}$
2. Learn *anomaly score* function g on $\mathcal{D}_{\text{train}}$

[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

2. Learn anomaly score function g on $\mathcal{D}_{\text{train}}$



[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

- Nominal sample: $Y_1, \ldots, Y_n$
- Test sample $\mathcal{D}_{\text{test}}$: $X_1, \ldots, X_m$

General idea

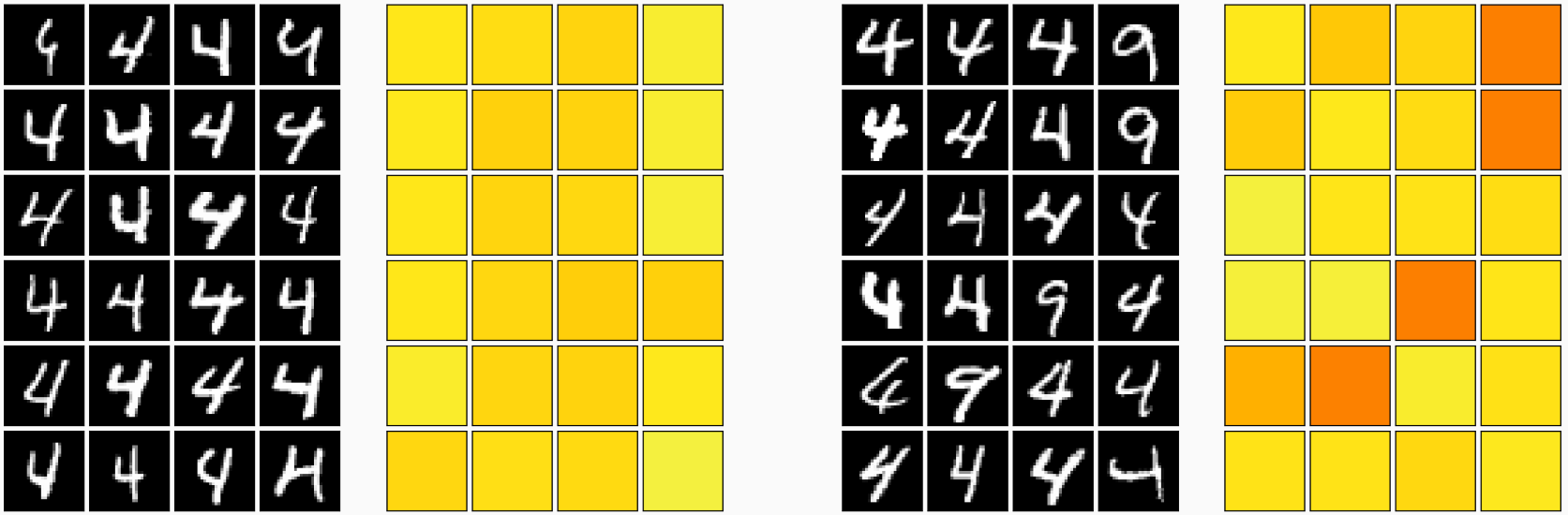
1. Split $Y_1, \ldots Y_n$ into $\underbrace{Y_1, \ldots, Y_k}_{\mathcal{D}_{\text{train}}}$ and $\underbrace{Y_{k+1}, \ldots, Y_n}_{\mathcal{D}_{\text{cal}}}$
2. Learn anomaly score function g on $\mathcal{D}_{\text{train}}$
3. Get scores for $\mathcal{D}_{\text{cal}}$ and for $\mathcal{D}_{\text{test}}$

[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

3. Get scores

$\mathcal{D}_{\text{cal}}$ Anomaly scores $\mathcal{D}_{\text{test}}$ Anomaly scores



[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

- Nominal sample: $Y_1, \ldots, Y_n$
- Test sample $\mathcal{D}_{\text{test}}$: $X_1, \ldots, X_m$

General idea

1. Split $Y_1, \ldots Y_n$ into $\underbrace{Y_1, \ldots, Y_k}_{\mathcal{D}_{\text{train}}}$ and $\underbrace{Y_{k+1}, \ldots, Y_n}_{\mathcal{D}_{\text{cal}}}$
2. Learn anomaly score function g on $\mathcal{D}_{\text{train}}$
3. Get scores for $\mathcal{D}_{\text{cal}}$ and for $\mathcal{D}_{\text{test}}$
4. For $i = 1, \ldots, m$, label X_i as novelty if

$$g(X_i) \text{ 'large' w.r.t. } g(Y_{k+1}), \ldots, g(Y_m)$$

using Counting Knockoff algorithm[2]

[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

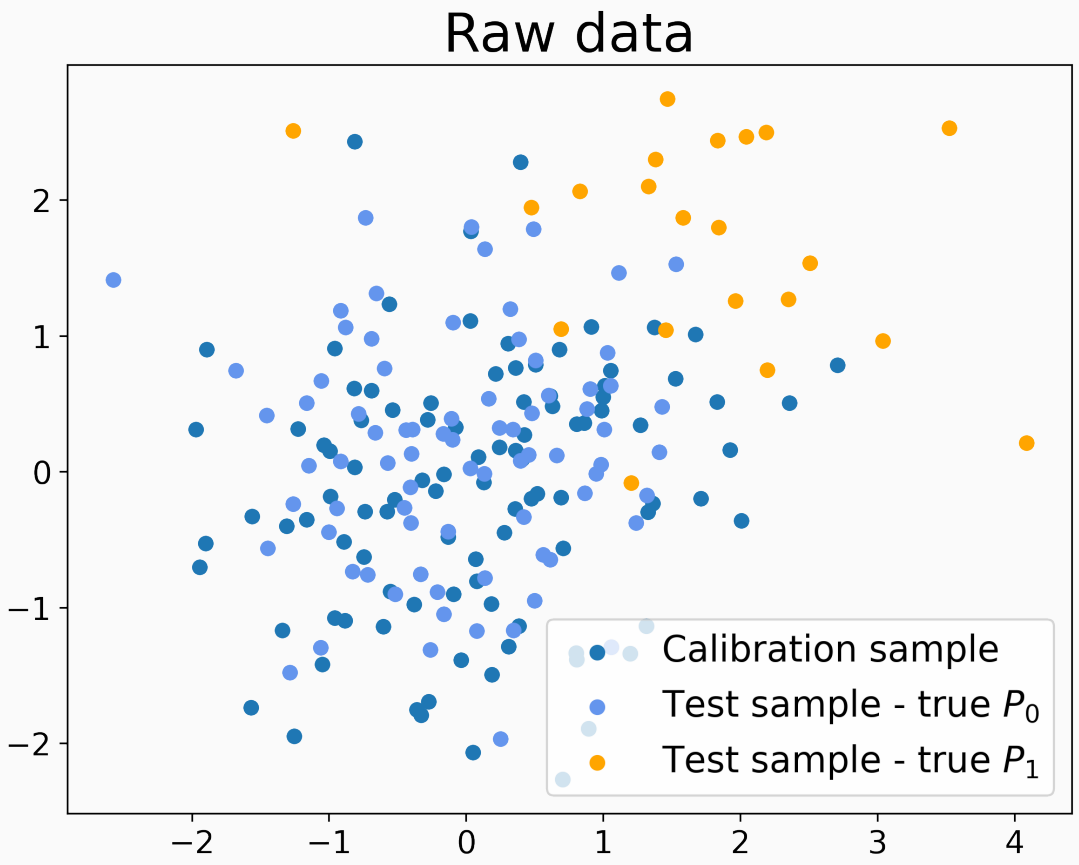
[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

4. Counting Knockoff

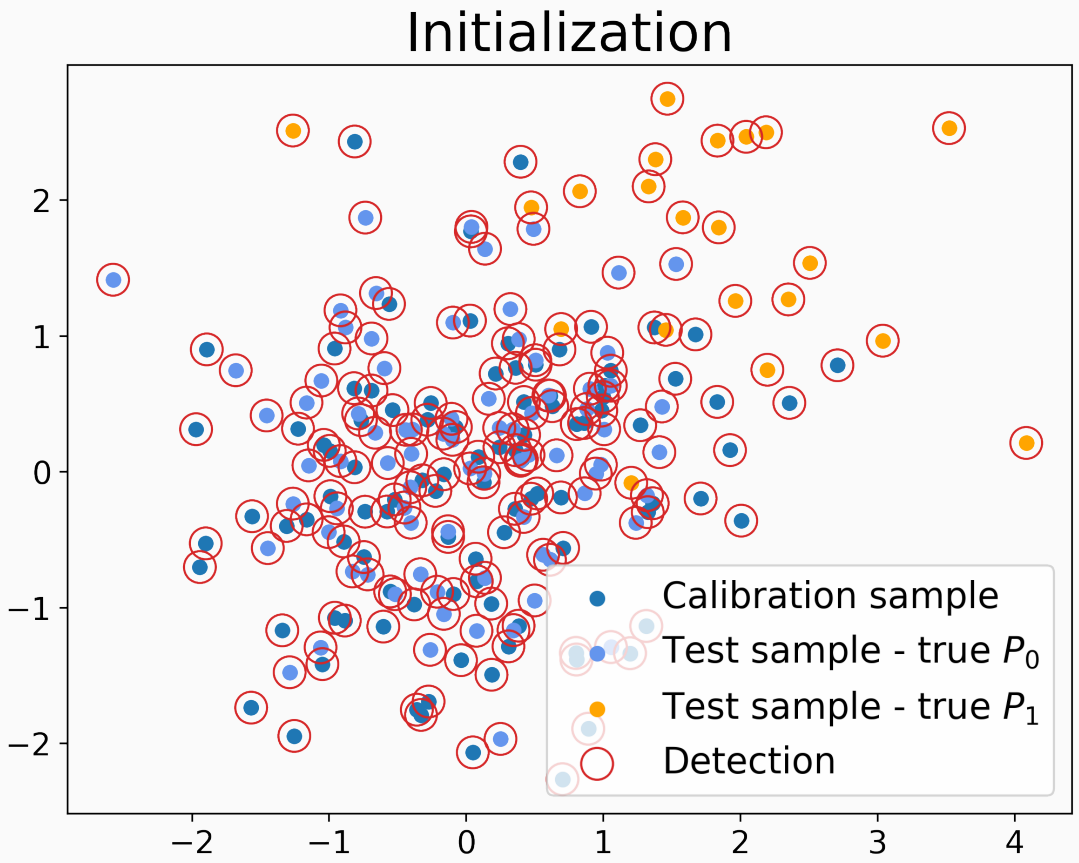
[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Counting knockoff[2]



[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

Counting knockoff[2]

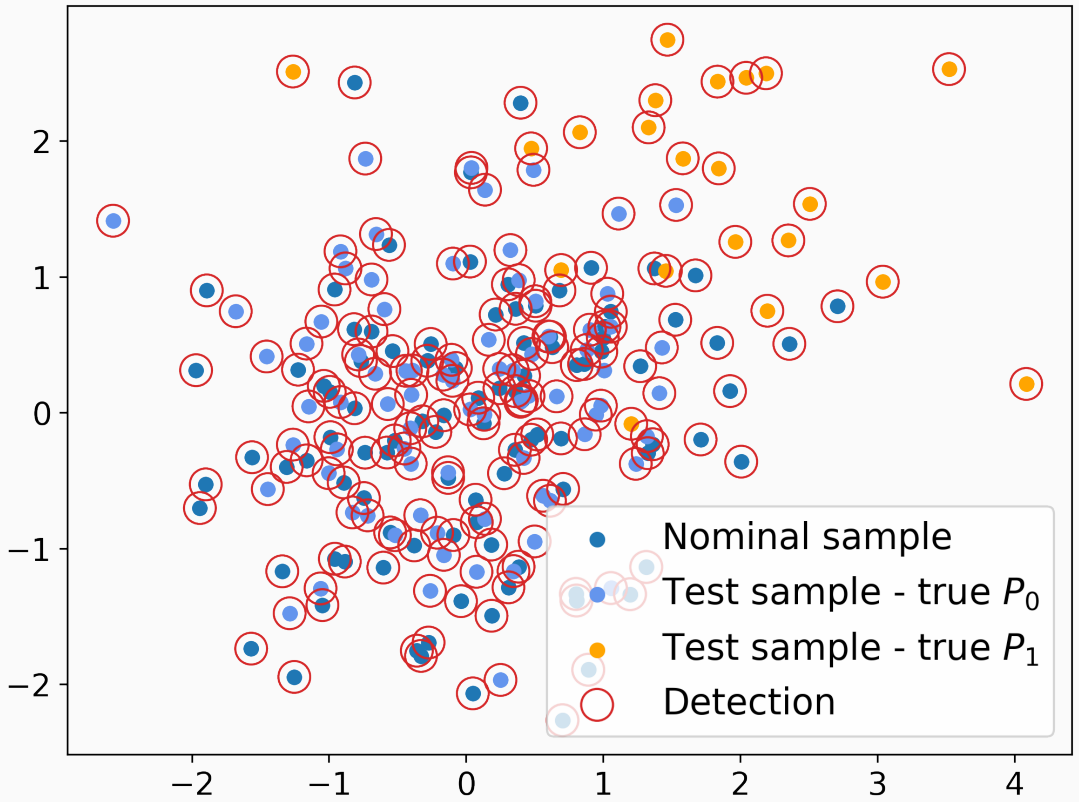


$$\widehat{\mathrm{FDP}} = \frac{\#\,\text{◉}}{\#\,\text{●} + \#\,\text{◉}} \frac{m}{n-k}$$

[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

Counting knockoff[2]

Remove observation with lowest anomaly score from detection set: Iteration=1

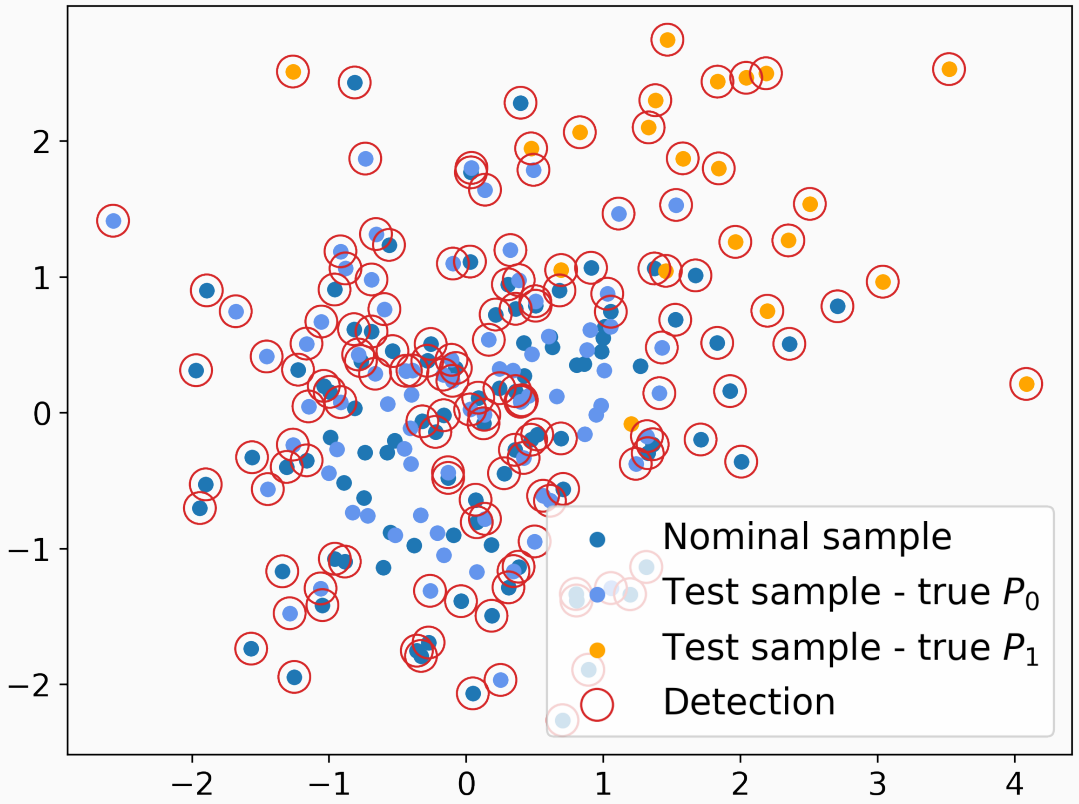


$$\widehat{\text{FDP}} = 100/99 \times 100/100 = 1.01 > \alpha = 0.2$$

[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

Counting knockoff[2]

Remove observation with lowest anomaly score from detection set: Iteration=50

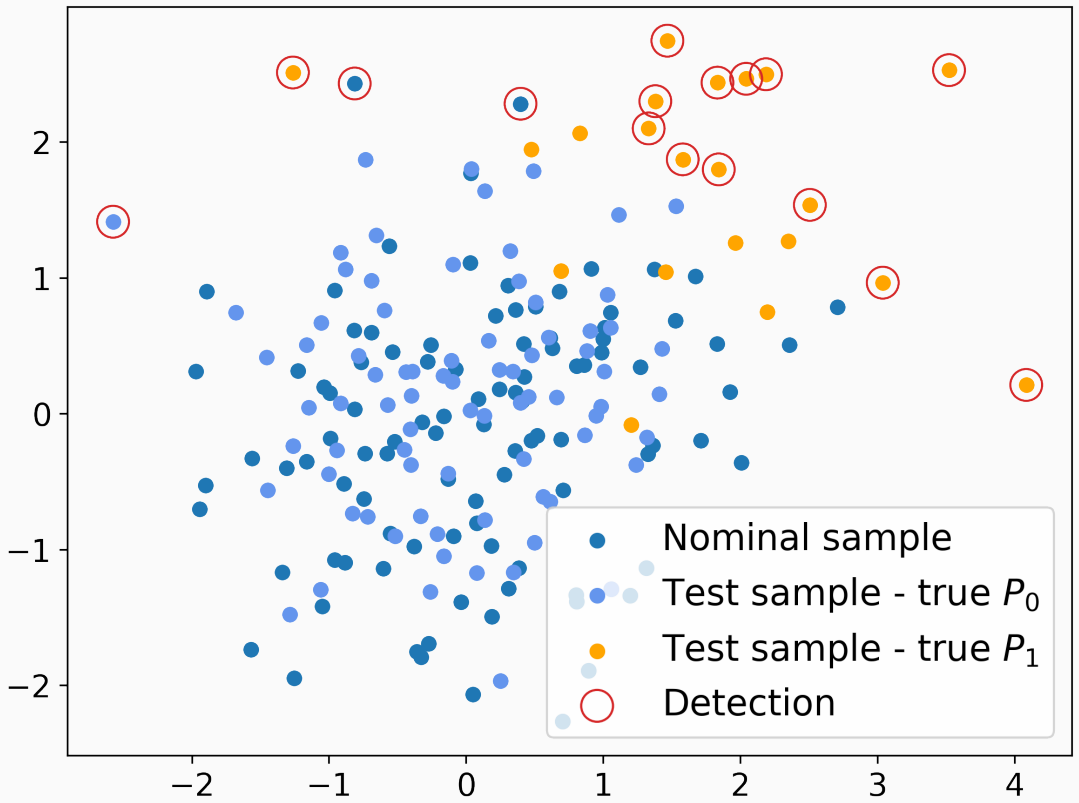


$$\widehat{\text{FDP}} = 80/77 \times 100/100 = 1.04 > \alpha = 0.2$$

[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

Counting knockoff[2]

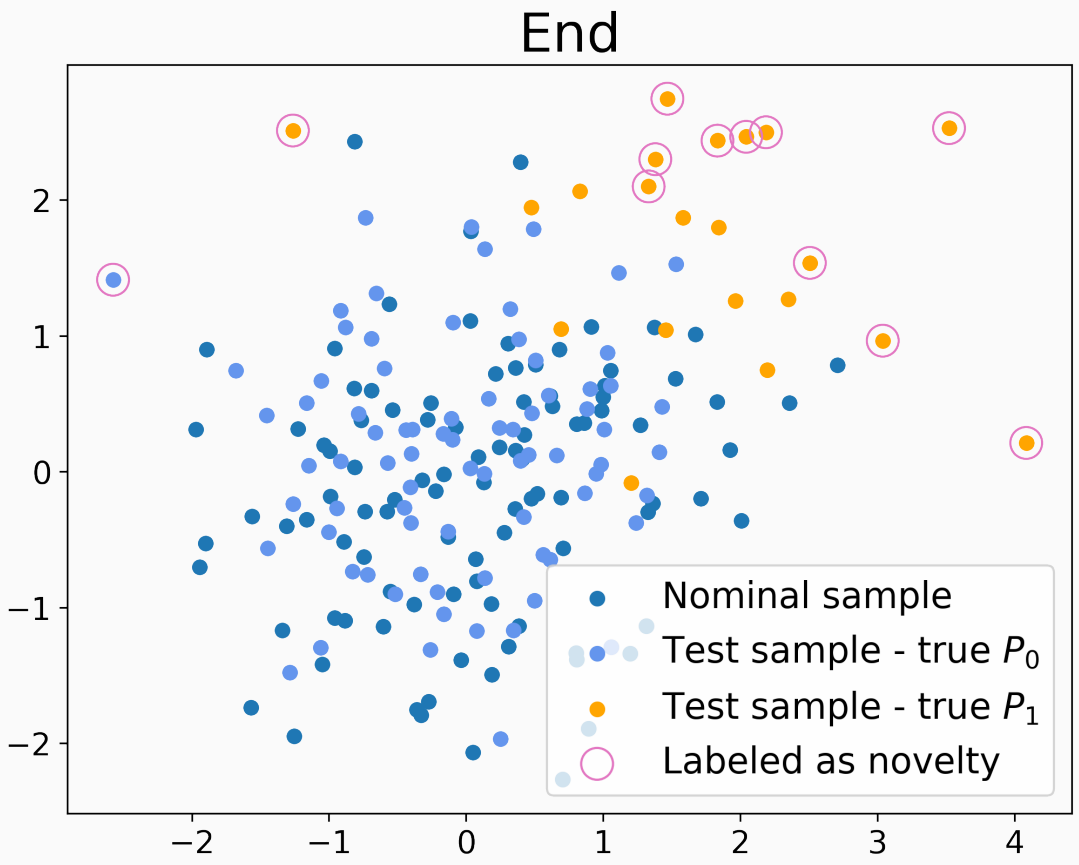
Remove observation with lowest anomaly score from detection set: Iteration=184



$$\widehat{\text{FDP}} = 2/13 \times 100/100 = 0.15 \leq \alpha = 0.2$$

[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

Counting knockoff[2]



[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

Conformal Anomaly Detection[1]

1. Split $Y_1, \dots Y_n$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$
2. Learn anomaly score function g on $\mathcal{D}_{\text{train}}$
3. Get scores for $\mathcal{D}_{\text{cal}}$ and for $\mathcal{D}_{\text{test}}$
4. Apply Counting Knockoff[2] to $\mathcal{D}_{\text{cal}}$ (as reference sample) and $\mathcal{D}_{\text{test}}$ (as test sample)

Comments

- FDR control is guaranteed under independence of the observations
  $\implies$ independence of the scores $S_k, \dots, S_{n+m}$ of $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$
- Key: g must not depend on the knowledge that $\mathcal{D}_{\text{cal}}$ are nominals

[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

Conformal Anomaly Detection[1]

1. Split $Y_1, \dots Y_n$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$
2. Learn anomaly score function g on $\mathcal{D}_{\text{train}}$
3. Get scores for $\mathcal{D}_{\text{cal}}$ and for $\mathcal{D}_{\text{test}}$
4. Apply Counting Knockoff[2] to $\mathcal{D}_{\text{cal}}$ (as reference sample) and $\mathcal{D}_{\text{test}}$ (as test sample)

Comments

- FDR control is guaranteed under independence of the observations
  $\implies$ independence of the scores $S_k, \dots, S_{n+m}$ of $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$
- Key: g must not depend on the knowledge that $\mathcal{D}_{\text{cal}}$ are nominals

[2] Asaf Weinstein, Rina Barber, and Emmanuel Candès. "A Power and Prediction Analysis for Knockoffs with Lasso Statistics". 2017

[1] Stephen Bates et al. "Testing for Outliers with Conformal p-values". 2021

New approach: AdaDetect

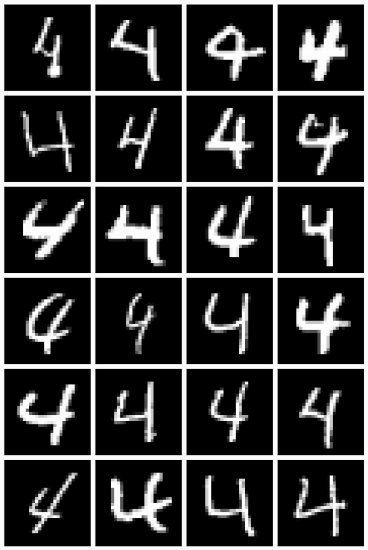
Previously: anomaly scoring function g learned from nominal data 'only'

New: learn g from **both nominal** data $(Y_i)_{1\leq i\leq n}$ and **unlabeled** data $(X_i)_{1\leq i\leq m}$ using classification

AdaDetect

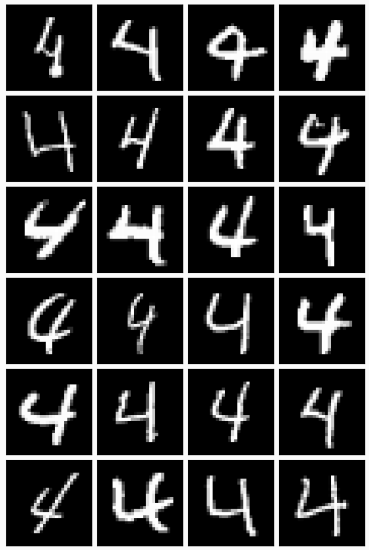
1. Split $Y_1, \dots Y_n$

Nominal sample

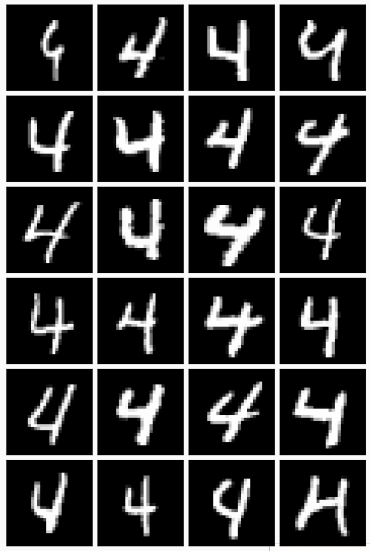


Split

$\mathcal{D}_{\text{train}}$



$\mathcal{D}_{\text{cal}}$

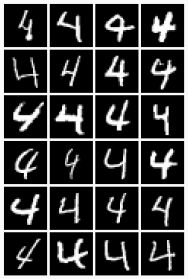


AdaDetect

2. Learn anomaly score function g (new)

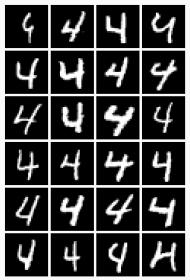
Class 0

$\mathcal{D}_{\text{train}}$

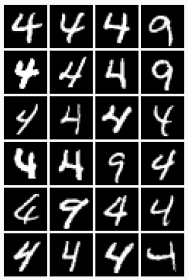


Class 1

$\mathcal{D}_{\text{cal}}$



$\mathcal{D}_{\text{test}}$



Classification algorithm

Class probability estimates (f_0, f_1)

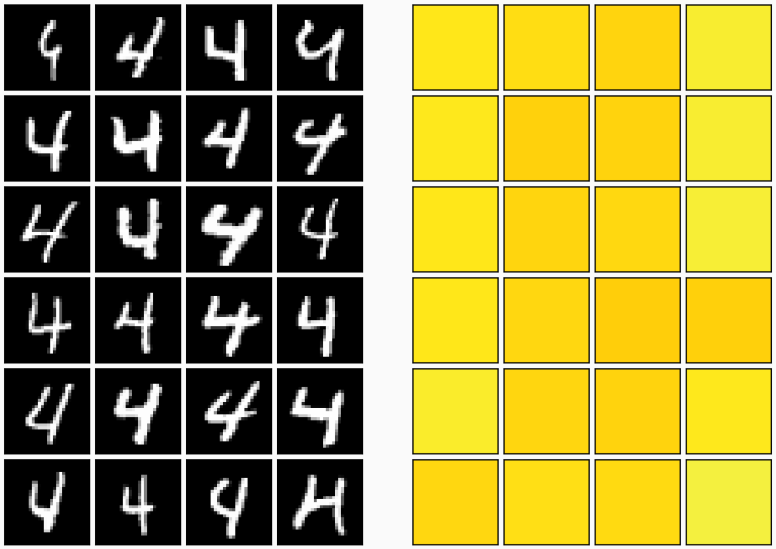
Anomaly score function

$g(\cdot\,; \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}) = f_1(\cdot)$

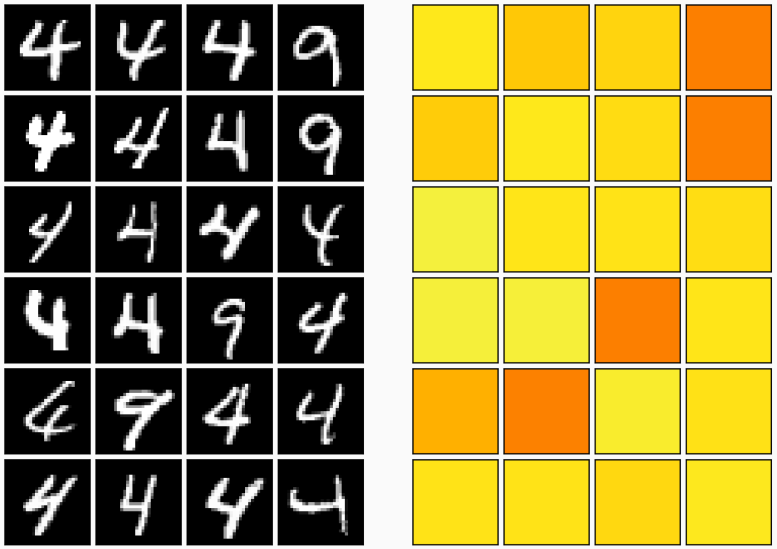
AdaDetect

3. Get scores

$\mathcal{D}_{\text{cal}}$ Anomaly scores



$\mathcal{D}_{\text{test}}$ Anomaly scores



4. Counting Knockoff

New approach: AdaDetect

1. Split $Y_1, \dots Y_n$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$
2. Learn anomaly score function g on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$
   E.g. by classifying $\mathcal{D}_{\text{train}}$ against $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$
3. Get scores for $\mathcal{D}_{\text{cal}}$ and for $\mathcal{D}_{\text{test}}$
4. Apply Counting Knockoff to $\mathcal{D}_{\text{cal}}$ (as reference sample) and $\mathcal{D}_{\text{test}}$ (as test sample)

Comments

- $\mathcal{D}_{\text{cal}}$ 'mixed' with $\mathcal{D}_{\text{test}} \implies$ we retain control
- More power: leverages unlabeled data
- Flexible: works with any classification algorithm

New approach: AdaDetect

1. Split $Y_1, \dots Y_n$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$
2. Learn anomaly score function g on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$
   E.g. by classifying $\mathcal{D}_{\text{train}}$ against $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$
3. Get scores for $\mathcal{D}_{\text{cal}}$ and for $\mathcal{D}_{\text{test}}$
4. Apply Counting Knockoff to $\mathcal{D}_{\text{cal}}$ (as reference sample) and $\mathcal{D}_{\text{test}}$ (as test sample)

Comments

- $\mathcal{D}_{\text{cal}}$ 'mixed' with $\mathcal{D}_{\text{test}} \implies$ we retain control
- More power: leverages unlabeled data
- Flexible: works with any classification algorithm

Theoretical results

- FDR control?
- Optimality?

FDR control

Assumption

- $(Y_1, \ldots, Y_n, X_i, i : X_i \sim P_0)$ is exchangeable conditionally on $(X_i, i : X_i \nsim P_0)$ (e.g., independence, equi-correlation)

Theorem

$$\text{FDR}(\text{AdaDetect}_\alpha) \begin{cases} \leq & \alpha m_0/m \\ \geq & m_0 \lfloor \alpha(n-k+1)/m \rfloor / (n-k+1) \end{cases}$$

- Furthermore: FDR $\leq \alpha$ for version of AdaDetect that estimates m_0
- Proof: multiple testing tools (BH procedure with p-values that are PRDS)
- Extend [Weinstein et al. (2017), Bates et al. (2021), Roquain and Mary (2021), Yang et al. (2021)]

Power result

Assumptions

- $(Y_1, \ldots, Y_n, X_1, \ldots, X_m)$ are mutually independent
- nominal density f_0, novelty density f_1

Additional notation

- $\mathcal{G}$: class of functions $\mathbb{R}^d \to \mathbb{R}$; set of classifiers $= \{\text{sign}(g), g \in \mathcal{G}\}$
- R_i theoretical risk restricted to label i, for 0-1 loss
- $\widehat{R}_i$ empirical risk restricted to label i, for 0-1 loss

Comparison to the power of the optimal Likelihood Ratio Test at some level $\beta = \beta(\alpha, m_1, m, \mathcal{G}, f_1/f_0)$:

- Likelihood ratio test:

$$H_0 : X \in P_0 \quad \text{vs.} \quad H_1 : X \in P_1$$

- Optimal test statistic$= g^*$ any increasing transformation of f_1/f_0

Power result

Assumptions

- $(Y_1, \ldots, Y_n, X_1, \ldots, X_m)$ are mutually independent
- nominal density f_0, novelty density f_1

Additional notation

- $\mathcal{G}$: class of functions $\mathbb{R}^d \to \mathbb{R}$; set of classifiers $= \{\text{sign}(g), g \in \mathcal{G}\}$
- R_i theoretical risk restricted to label i, for 0-1 loss
- $\widehat{R}_i$ empirical risk restricted to label i, for 0-1 loss

Comparison to the power of the optimal Likelihood Ratio Test at some level $\beta = \beta(\alpha, m_1, m, \mathcal{G}, f_1/f_0)$:

- Likelihood ratio test:

$$H_0 : X \in P_0 \quad \text{vs.} \quad H_1 : X \in P_1$$

- Optimal test statistic$= g^*$ any increasing transformation of f_1/f_0

Power result

Neyman-Pearson Empirical risk minimizer (ERM)[3]

$$\widehat{g} \in \operatorname{argmin}\{\widehat{R}_1(g), \widehat{R}_0(g) \leq \beta, g \in \mathcal{G}\},$$

Theorem

For $\beta \lesssim \alpha m_1/m(1 - R_1(g^*) - \Delta - b)$, if $n - k \geq 2m/\alpha$,

$$\text{TDR(AdaDetect with } \widehat{g}) \gtrsim \underbrace{1 - R_1(g^*)}_{\text{power of LRT}} - \Delta - b$$

Two error terms:

- $b = R_1(g^*_{\mathcal{G}}) - R_1(g^*)$ (bias) where $g^*_{\mathcal{G}}$=NP th. risk minimizer over $\mathcal{G}$
- $\Delta \lesssim \frac{m+n-k}{m_1}\sqrt{V(\mathcal{G})}\left(\sqrt{\frac{\log k}{k}} + \sqrt{\frac{\log(n-k)}{n-k}}\right)$ where $V(\mathcal{G}) = \text{VC dim}$

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. "Semi-Supervised Novelty Detection". 2010

Power result

Neyman-Pearson Empirical risk minimizer (ERM)[3]

$$\widehat{g} \in \text{argmin}\{\widehat{R}_1(g), \widehat{R}_0(g) \leq \beta, g \in \mathcal{G}\},$$

Theorem

For $\beta \lesssim \alpha m_1/m(1 - R_1(g^*) - \Delta - b)$, if $n - k \geq 2m/\alpha$,

$$\text{TDR(AdaDetect with } \widehat{g}) \gtrsim \underbrace{1 - R_1(g^*)}_{\text{power of LRT}} - \Delta - b$$

Two error terms:

- $b = R_1(g^*_{\mathcal{G}}) - R_1(g^*)$ (bias) where $g^*_{\mathcal{G}}$=NP th. risk minimizer over $\mathcal{G}$
- $\Delta \lesssim \frac{m+n-k}{m_1}\sqrt{V(\mathcal{G})}\left(\sqrt{\frac{\log k}{k}} + \sqrt{\frac{\log(n-k)}{n-k}}\right)$ where $V(\mathcal{G}) = \text{VC dim}$

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. "Semi-Supervised Novelty Detection". 2010

Applications on graphs: 1/3

Graph-level anomaly detection

- $\mathcal{G}$ some space of graphs
- $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} P_0 \in \mathcal{G}$
- $X_1, \ldots, X_m \in \mathcal{G}$

$$H_{0,i} : X_i \sim P_0$$

In practice

- *graph* classifier[4, 5]

[4] Nino Shervashidze et al. "Weisfeiler-Lehman Graph Kernels". 2011

[5] Federico Errica et al. "A fair comparison of graph neural networks for graph classification". In: Proceedings of the 8th International Conference on Learning Representations (ICLR). 2020

Applications on graphs: 2/3

Node-level anomaly detection

- $Y_1, \ldots, Y_n \sim P_0 \in \mathbb{R}^d$
- $X_1, \ldots, X_m \in \mathbb{R}^d$
- $(A_{i,j})_{1 \leq i,j \leq n+m} \in \{0,1\}^{n+m}$

$$H_{0,i} : X_i \sim P_0$$

- $\neq$ framework
- Claim:

$Y^\pi, X^\pi, A^\pi \sim Y, X, A$ for any permutation π of two nulls $\implies$ exchangeability of the null scores holds $\implies$ FDR guarantee

In practice

- *node* classifier (node permutation-invariant) such as GCN[6]

[6] Thomas N Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". 2016

Applications on graphs: 3/3

Link prediction

- $(A_{i,j})_{1 \leq i,j \leq N} \in \{0,1\}^N$ adjacency matrix
- $(M_{i,j})_{1 \leq i,j \leq N} \in \{0,1\}^N$ missing data matrix
- m unobserved edges, n observed edges absent/present

$$H_{0,i,j} : A_{i,j} = 0 \quad \text{vs.} \quad H_{1,i,j} : A_{i,j} = 1$$

General idea:

- *learn* link predictor[7, 8];
- compare link prediction scores of *unobserved* edges to *known absent* edges.

(!) Completely $\neq$ framework: no guarantees, but empirical results

[7] Kevin Bleakley, Gérard Biau, and Jean-Philippe Vert. "Supervised reconstruction of biological networks with local models". July 2007

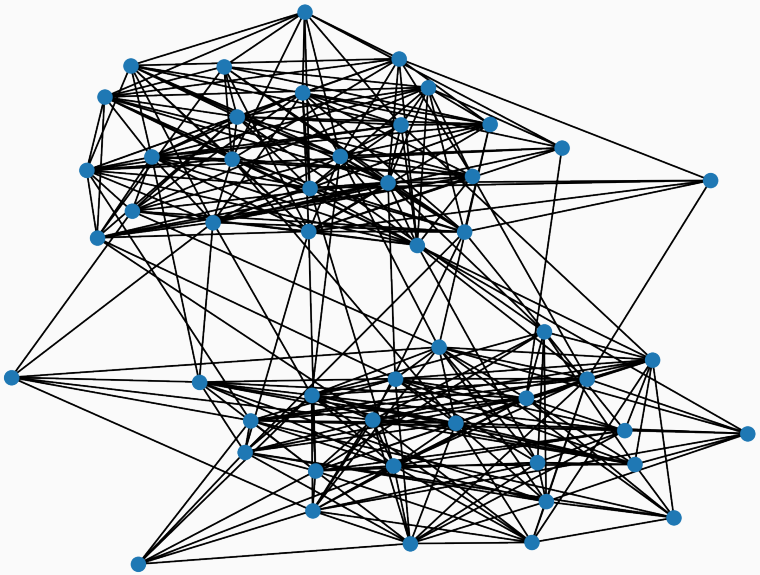
[8] Muhan Zhang and Yixin Chen. "Link prediction based on graph neural networks". In: Advances in Neural Information Processing Systems. 2018, pp. 5165–5175

Num. results: graph-level anomaly detection

Inliers

$P_0 = \mathrm{SBM}(\pi_0, \gamma_0)$

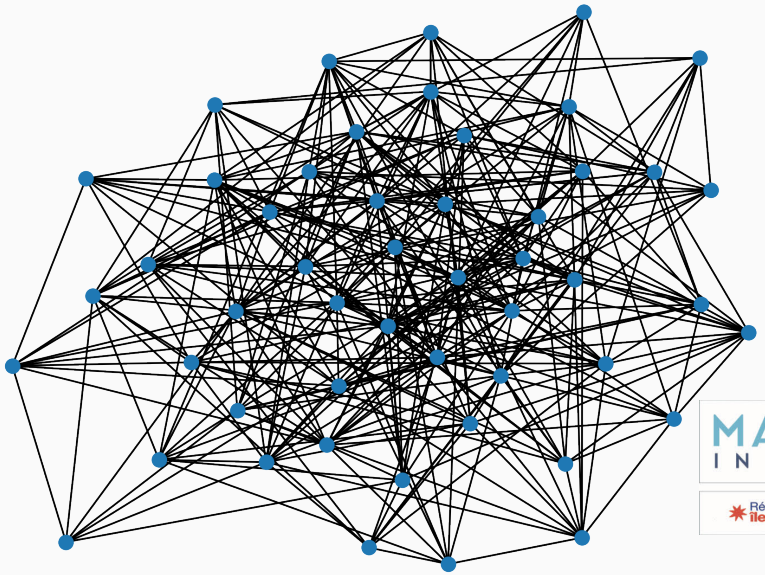
$$\pi_0 = (1/2, 1/2), \gamma_0 = \begin{pmatrix} 0.5 & 0.05 \\ 0.05 & 0.5 \end{pmatrix}$$



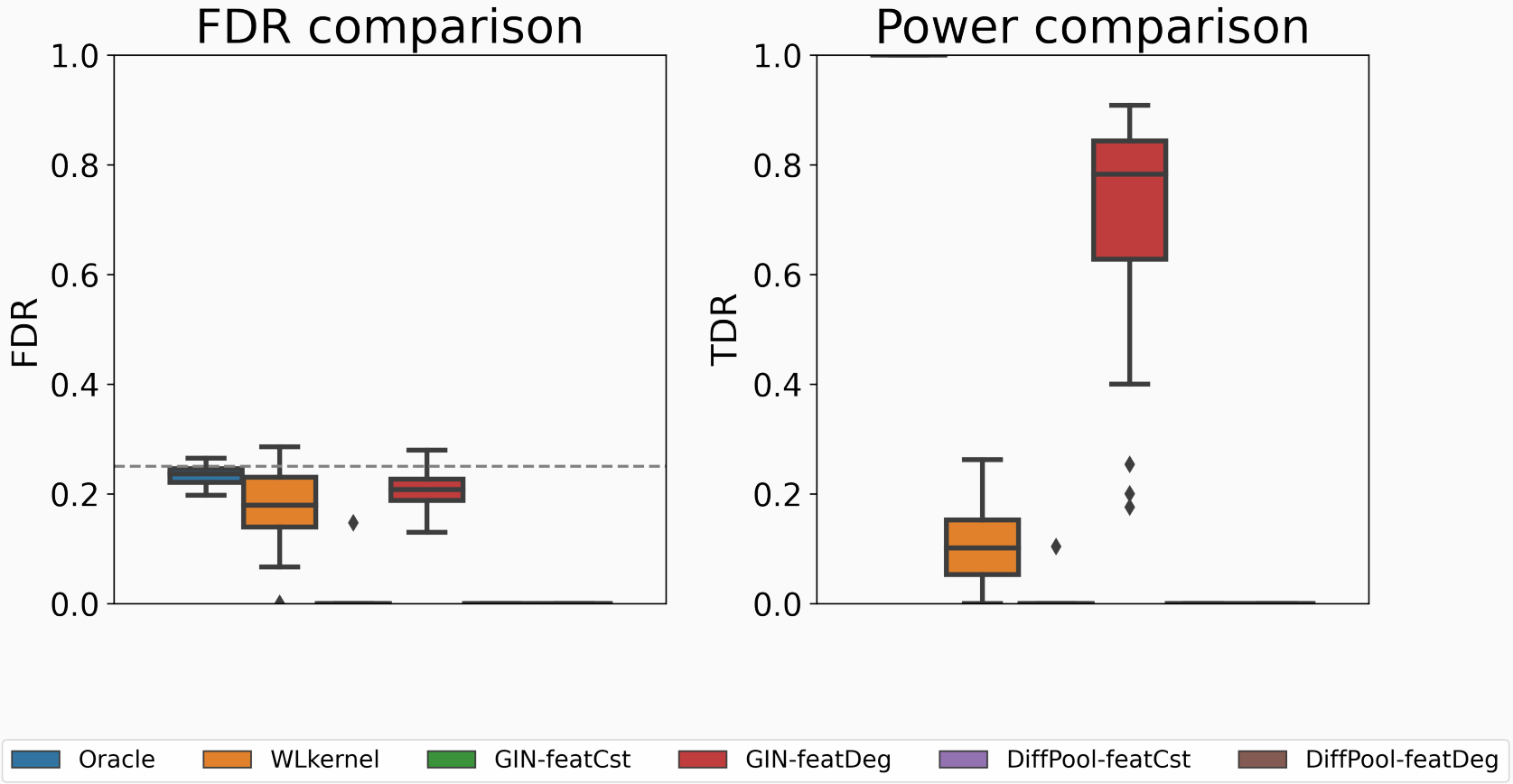
Outliers

$P_1 = \mathrm{SBM}(\pi_1, \gamma_1)$

$$\pi_1 = (1/2, 1/2), \gamma_1 = \begin{pmatrix} 0.05 & 0.5 \\ 0.5 & 0.05 \end{pmatrix}$$

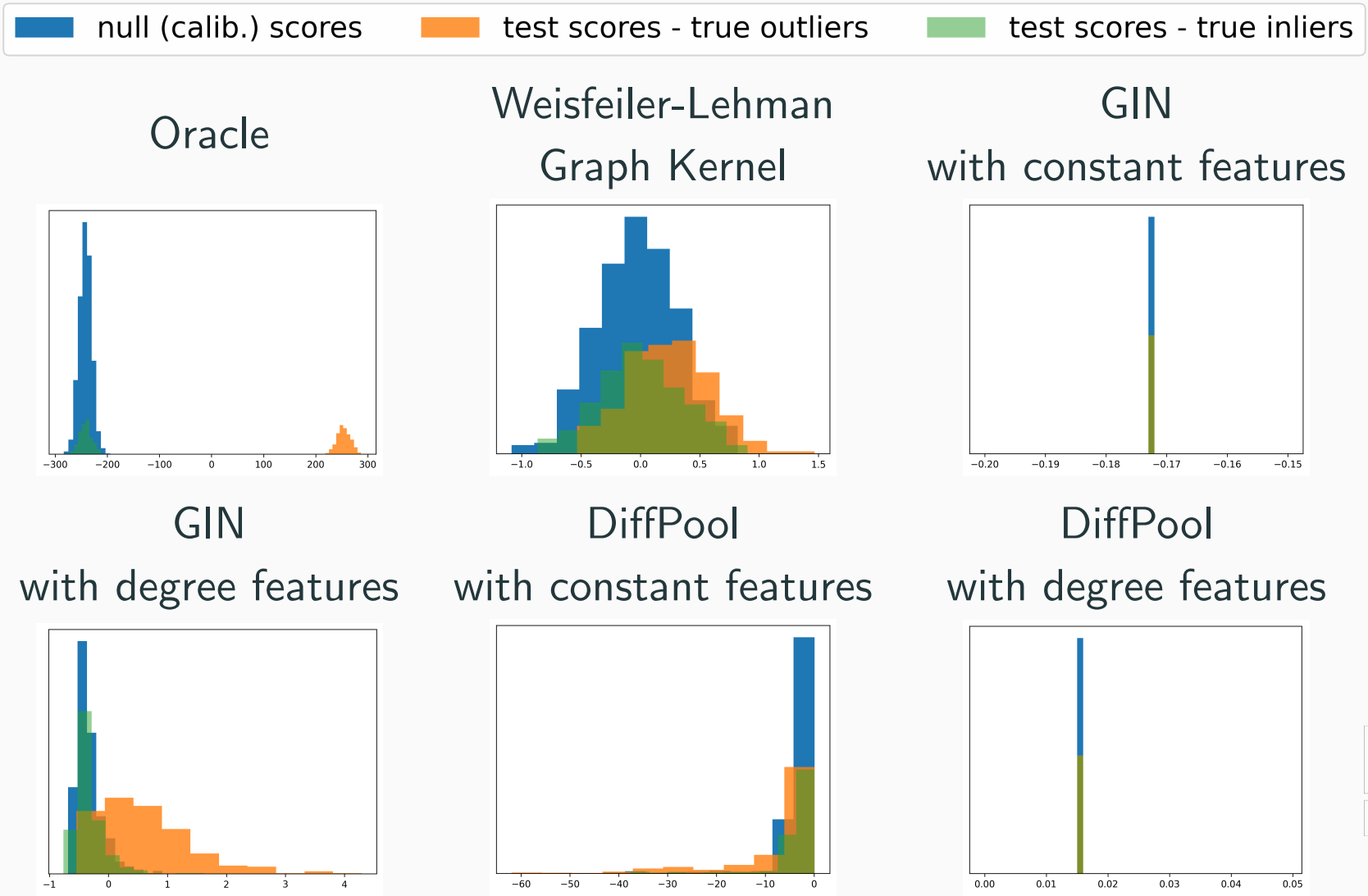


Num. results: graph-level anomaly detection



- For GNN: default hyperparameters as in *"A fair comparison of graph neural networks for graph classification"* (Errica et al. , ICLR, 2020)
- Possible hyper-parameter tuning

Num. results: graph-level anomaly detection



Take-away

- New powerful method for novelty detection, with finite-sample FDR control
- Combines classification (learn score function) and multiple testing (threshold score function)