Binary Classifiers as Dilations

Filip Obradović
Northwestern University

Gabriel Ziegler
University of Edinburgh

27 November 2024

Applied Reading Group, University of Edinburgh

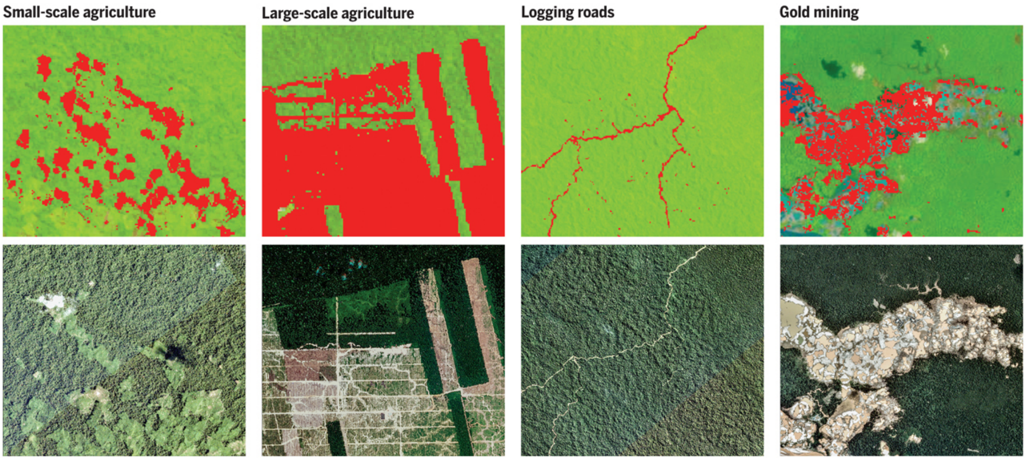
Binary Diagnostic Tests

- Medical diagnostic tests:
  - HIV test, caries, cancer screening, diabetes, pulmonary embolism...
  - Antigen tests: Covid, influenza, RSV, streptococcus, ...
  - Antibody/ELISA tests

Binary Diagnostic Tests

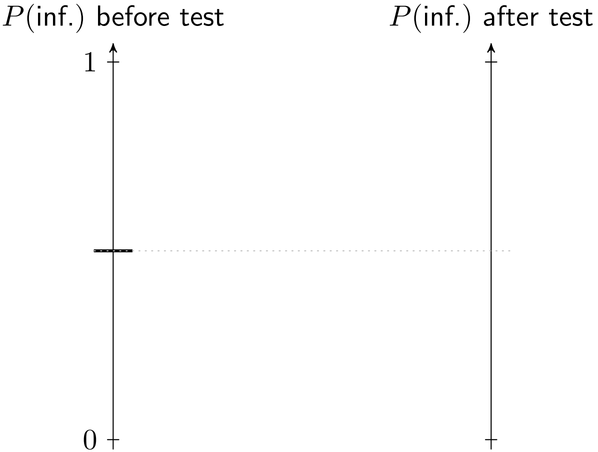
- Medical diagnostic tests:
  - HIV test, caries, cancer screening, diabetes, pulmonary embolism...
  - Antigen tests: Covid, influenza, RSV, streptococcus, ...
  - Antibody/ELISA tests
- Binary expert testing, binary classification, ...

Land Cover Change



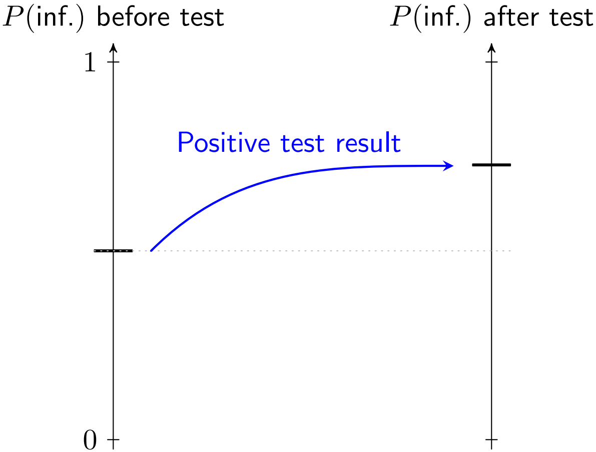
Finer, Matt, et al. "Combating deforestation: From satellite to intervention." *Science* (2018).

Usual Binary Classifiers



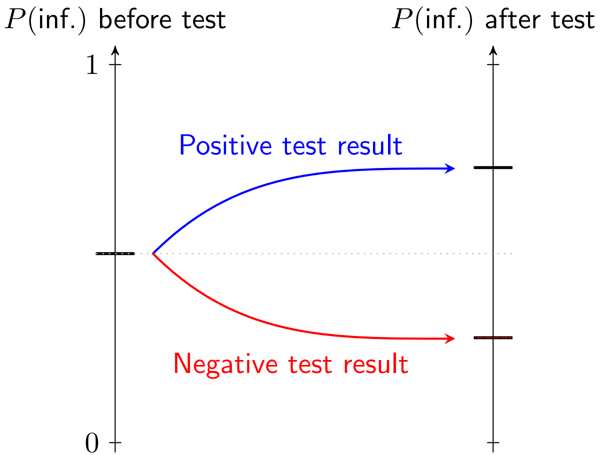
- Stick to medical interpretation:

Usual Binary Classifiers



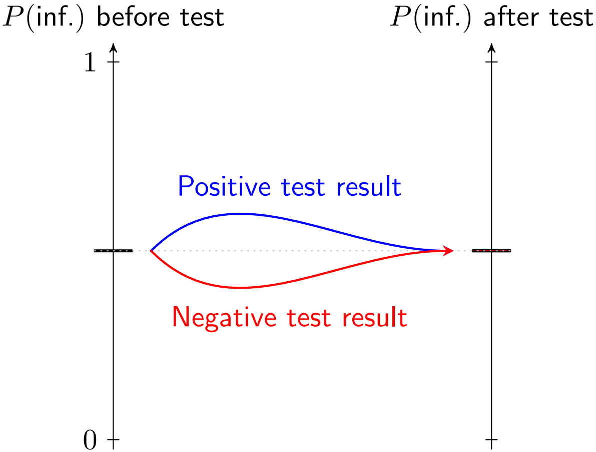
- Stick to medical interpretation:
  - Positive result, or

Usual Binary Classifiers



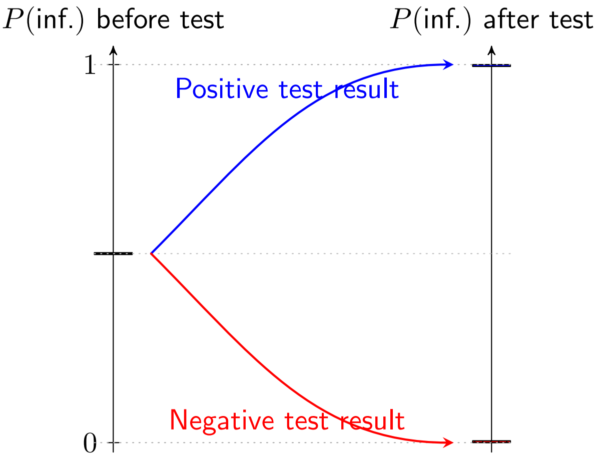
- Stick to medical interpretation:
  - Positive result, or
  - negative result

Usual Binary Classifiers



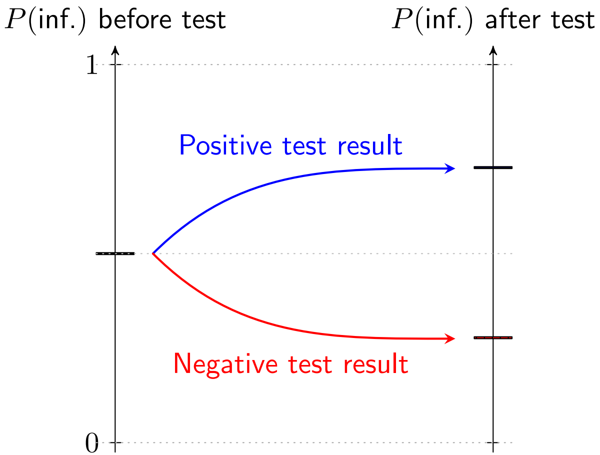
- Stick to medical interpretation:
  - Positive result, or
  - negative result
- Possible extremes:
  - No information

Usual Binary Classifiers



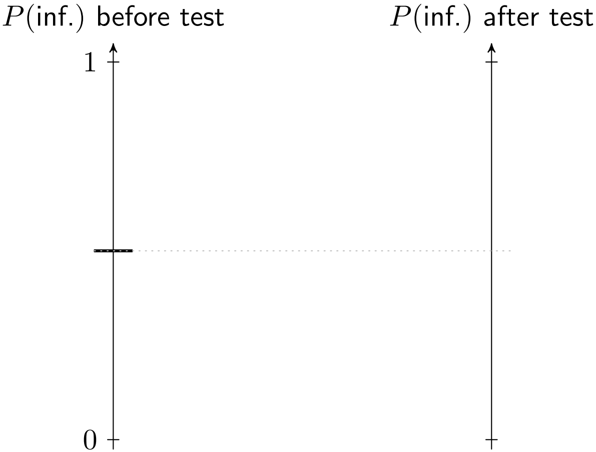
- Stick to medical interpretation:
  - Positive result, or
  - negative result
- Possible extremes:
  - No information
  - Full information

Usual Binary Classifiers

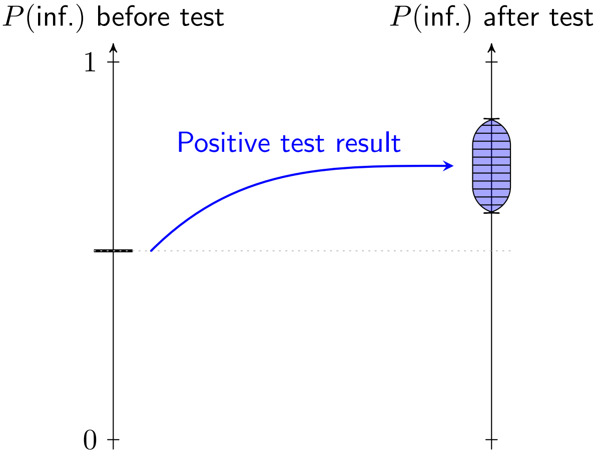


- Stick to medical interpretation:
  - Positive result, or
  - negative result
- Possible extremes:
  - No information
  - Full information
- always (weakly) informative à la Blackwell

Tests with ambiguous information



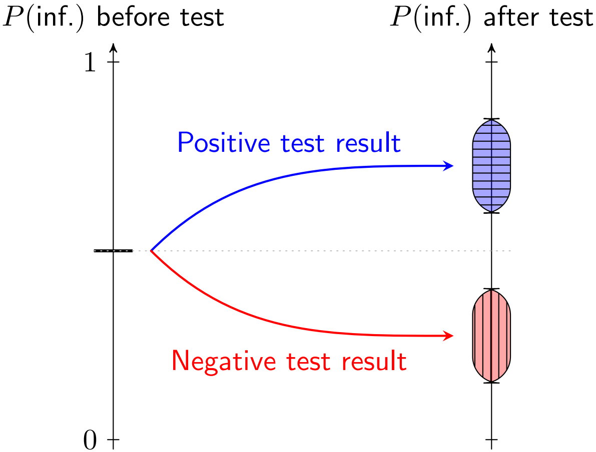
Tests with ambiguous information



After test result: set of posteriors

aka *ambiguity/Knightian uncertainty*

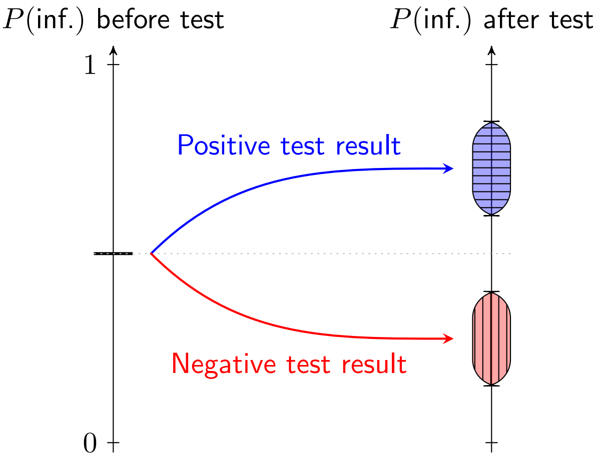
Tests with ambiguous information



After test result: set of posteriors

aka *ambiguity/Knightian uncertainty*

Tests with ambiguous information



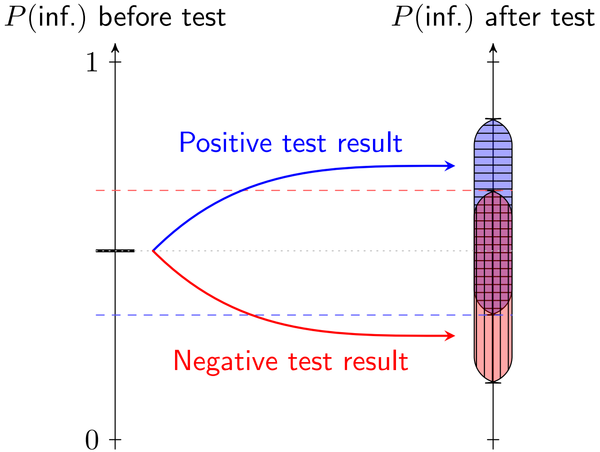
After test result: set of posteriors

aka *ambiguity/Knightian uncertainty*

Not a test à la Blackwell:

- Informativeness?

Tests with ambiguous information



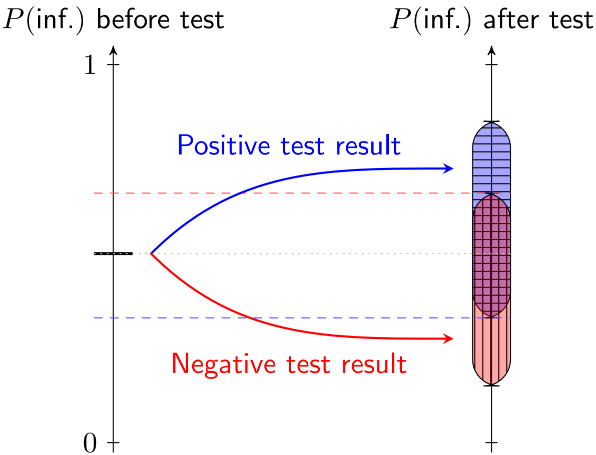
After test result: set of posteriors

aka *ambiguity/Knightian uncertainty*

Not a test à la Blackwell:

- Informativeness?
- Dilation

Tests with ambiguous information



After test result: set of posteriors

aka *ambiguity/Knightian uncertainty*

Not a test à la Blackwell:

- Informativeness?
- Dilation $\hat{=}$ not informative

Why ambiguous information?

Often the test is evaluated relative to a **reference** test

Why ambiguous information?

Often the test is evaluated relative to a **reference** test

- If reference test is **imperfect**: ref test result $\neq$ true underlying state
  - reference test result $\neq$ true underlying state
  - false positives and/or false negatives

Why ambiguous information?

Often the test is evaluated relative to a **reference** test

- If reference test is **imperfect**: ref test result $\neq$ true underlying state
  - reference test result $\neq$ true underlying state
  - false positives and/or false negatives
- But, new test might also produce false results:
  - **Correlation** between these errors is fundamentally unknown

Why ambiguous information?

Often the test is evaluated relative to a **reference** test

- If reference test is **imperfect**: ref test result $\neq$ true underlying state
  - reference test result $\neq$ true underlying state
  - false positives and/or false negatives
- But, new test might also produce false results:
  - **Correlation** between these errors is fundamentally unknown
- Usual solution: Assume knowledge of correlation, e.g. *independence*

Why ambiguous information?

Often the test is evaluated relative to a **reference** test

- If reference test is **imperfect**: ref test result $\neq$ true underlying state
  - reference test result $\neq$ true underlying state
  - false positives and/or false negatives
- But, new test might also produce false results:
  - **Correlation** between these errors is fundamentally unknown
- Usual solution: Assume knowledge of correlation, e.g. *independence*
- Here: **Robust approach** to allow for all possible correlations
  - Possible interpretation: Ambiguity or group of disagreeing experts

Evaluating new tests

Policy maker might want know when new test is a dilation

Evaluating new tests

Policy maker might want know when new test is a dilation

Main question:

When is a binary classifier, evaluated with a reference, a dilation?

Evaluating new tests

Policy maker might want know when new test is a dilation

Main question:

When is a binary classifier, evaluated with a reference, a dilation?

Preview of results:

Evaluating new tests

Policy maker might want know when new test is a dilation

Main question:

When is a binary classifier, evaluated with a reference, a dilation?

Preview of results:

1. Ambiguous information when an imperfect test is reference, under full support assumption

Evaluating new tests

Policy maker might want know when new test is a dilation

Main question:

When is a binary classifier, evaluated with a reference, a dilation?

Preview of results:

1. Ambiguous information when an imperfect test is reference, under full support assumption
2. Full characterization of when a dilation occurs

Evaluating new tests

Policy maker might want know when new test is a dilation

Main question:

When is a binary classifier, evaluated with a reference, a dilation?

Preview of results:

1. Ambiguous information when an imperfect test is reference, under full support assumption
2. Full characterization of when a dilation occurs
3. Statistical inference for "$H0$: new test is a dilation" based on subvector inference

Evaluating new tests

Policy maker might want know when new test is a dilation

Main question:

When is a binary classifier, evaluated with a reference, a dilation?

Preview of results:

1. Ambiguous information when an imperfect test is reference, under full support assumption
2. Full characterization of when a dilation occurs
3. Statistical inference for "$H0$: new test is a dilation" based on subvector inference
4. Empirical application: Covid LFT tests ✓, but other procedures ?, ML prediction

More on Dilations

- Formally introduced by Seidenfeld and Wasserman (1993):

  "counterintuitive phenomenon" - after test one knows less than before

More on Dilations

- Formally introduced by Seidenfeld and Wasserman (1993):

  "counterintuitive phenomenon" - after test one knows less than before

- Since then mainly a theoretical curiosity
  - Some interesting applications to mechanism design/Bayesian persuasion

    Bose and Renou (2014), Beauchêne, Li, and Li (2019), Cheng (2021), Pahlke (2022)
  - Recently, also experimental evidence: in particular, Shishkin and Ortoleva (2023)

    Experiments on ambiguous information: Kellner, Le Quement, and Riener (2022), Kops and Pasichnichenko (2022), Liang (2022), Epstein and Halevy (2023)

More on Dilations

- Formally introduced by Seidenfeld and Wasserman (1993):

  "counterintuitive phenomenon" - after test one knows less than before

- Since then mainly a theoretical curiosity
  - Some interesting applications to mechanism design/Bayesian persuasion

    Bose and Renou (2014), Beauchêne, Li, and Li (2019), Cheng (2021), Pahlke (2022)
  - Recently, also experimental evidence: in particular, Shishkin and Ortoleva (2023)

    Experiments on ambiguous information: Kellner, Le Quement, and Riener (2022), Kops and Pasichnichenko (2022), Liang (2022), Epstein and Halevy (2023)

- This paper is the first illustration of this curious property "in the field"
  - But see Manski (2018)

Outline

1. Model
2. Illustrations & Applications
3. Statistical Inference (work in progress)
4. Extension and Conclusions

1 – Model

New test: Accuracy

Goal: Evaluate a new diagnostic test.

New test: Accuracy

Goal: Evaluate a new diagnostic test.

$t = 1$ when person has positive test, $t = 0$ otherwise

$+$ denote infection, $-$ otherwise

New test: Accuracy

Goal: Evaluate a new diagnostic test.

$t = 1$ when person has positive test, $t = 0$ otherwise

$+$ denote infection, $-$ otherwise

We will need:

$$\mathbb{P}(t = 1|+)$$

$$\mathbb{P}(t = 0|-)$$

New test: Accuracy

Goal: Evaluate a new diagnostic test.

$t = 1$ when person has positive test, $t = 0$ otherwise

$+$ denote infection, $-$ otherwise

We will need:

$$\mathbb{P}(t = 1|+) =: \theta_1 \quad \text{(Sensitivity)}$$

$$\mathbb{P}(t = 0|-)$$

New test: Accuracy

Goal: Evaluate a new diagnostic test.

$t = 1$ when person has positive test, $t = 0$ otherwise

$+$ denote infection, $-$ otherwise

We will need:

$$\mathbb{P}(t = 1|+) =: \theta_1 \quad \text{(Sensitivity)}$$

$$\mathbb{P}(t = 0|-) =: \theta_0 \quad \text{(Specificity)}$$

New test: Accuracy

Goal: Evaluate a new diagnostic test.

$t = 1$ when person has positive test, $t = 0$ otherwise

$+$ denote infection, $-$ otherwise

We will need:

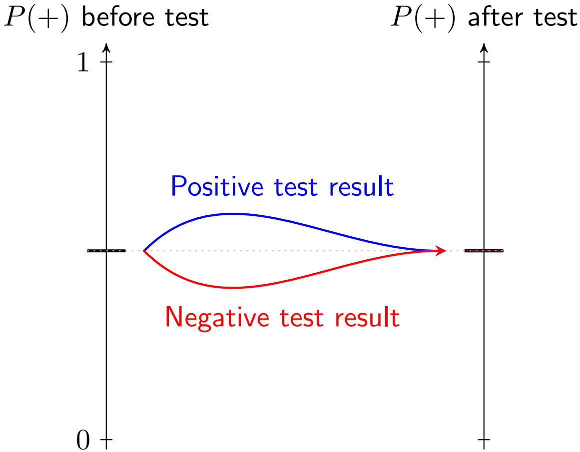
$$\mathbb{P}(t = 1|+) =: \theta_1 \quad \text{(Sensitivity)}$$

$$\mathbb{P}(t = 0|-) =: \theta_0 \quad \text{(Specificity)}$$

Assumption (Non-triviality)

The population satisfies $\mathbb{P}(+) \in (0, 1)$.

No Information

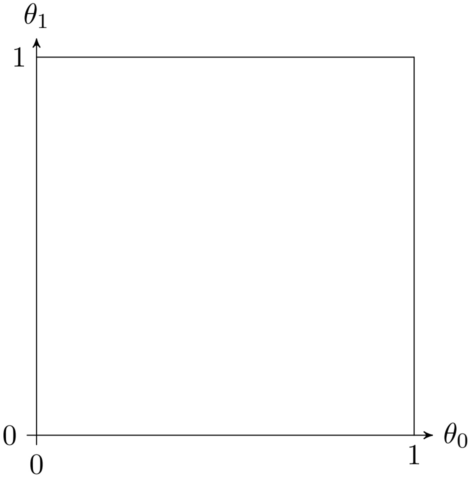


Observation

The new test provides no information (in Blackwell sense) if and only if

$$\theta_0 + \theta_1 = 1.$$

No Information

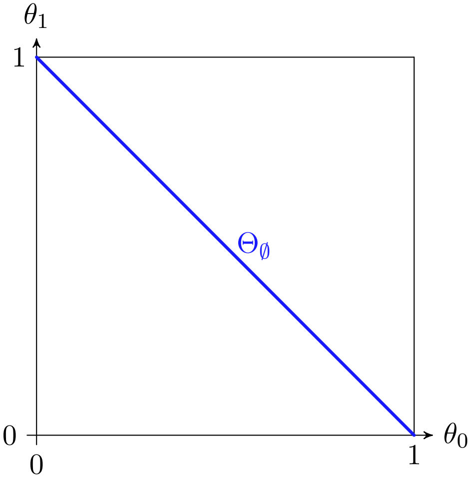


Observation

The new test provides no information (in Blackwell sense) if and only if

$$\theta_0 + \theta_1 = 1.$$

No Information



Observation

The new test provides no information (in Blackwell sense) if and only if

$$\theta_0 + \theta_1 = 1.$$

$$\Theta_\emptyset := \Big\{(\theta_0, \theta_1) : \theta_0 + \theta_1 = 1\Big\}$$

Gold Standards

To obtain $\theta = (\theta_0, \theta_1)$ true infection needs to be observed

Problem: Observing whether a person is truly infected is impossible/extremely costly

Gold Standards

To obtain $\theta = (\theta_0, \theta_1)$ true infection needs to be observed

Problem: Observing whether a person is truly infected is impossible/extremely costly

(Partial) Solution: Use an established test as reference. E.g. PCR test as "Gold Standard"

$r = 1$ when person has positive *reference* test, $r = 0$ otherwise

Gold Standards

To obtain $\theta = (\theta_0, \theta_1)$ true infection needs to be observed

Problem: Observing whether a person is truly infected is impossible/extremely costly

(Partial) Solution: Use an established test as reference. E.g. PCR test as "Gold Standard"

$r = 1$ when person has positive *reference* test, $r = 0$ otherwise

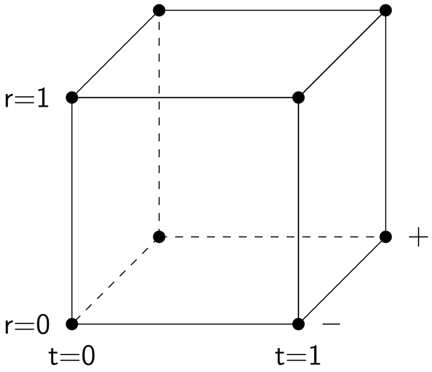
Data: $\mathbb{P}(r, t)$

Need: $\mathbb{P}(t|\pm)$ (or $\mathbb{P}(r, t, \pm)$)

Multiplicity and Ambiguous Information

Data: $\mathbb{P}(r,t)$

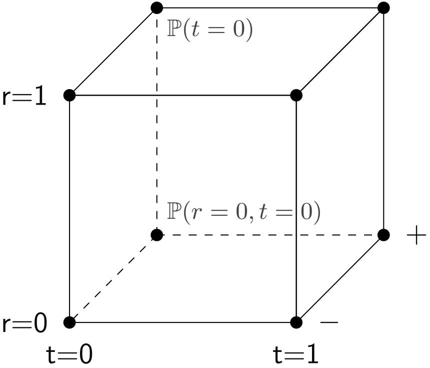
Need: $\mathbb{P}(r,t,\pm)$ or $\mathbb{P}(t|\pm)$



Multiplicity and Ambiguous Information

Data: $\mathbb{P}(r,t)$

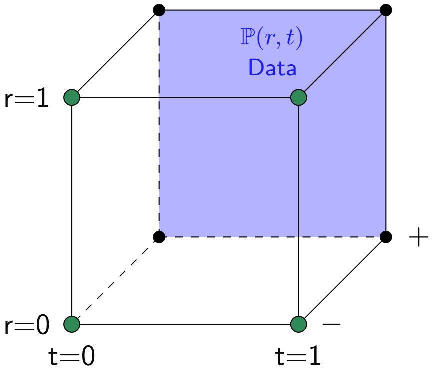
Need: $\mathbb{P}(r,t,\pm)$ or $\mathbb{P}(t|\pm)$



Multiplicity and Ambiguous Information

Data: $\mathbb{P}(r,t)$

Need: $\mathbb{P}(r,t,\pm)$ or $\mathbb{P}(t|\pm)$



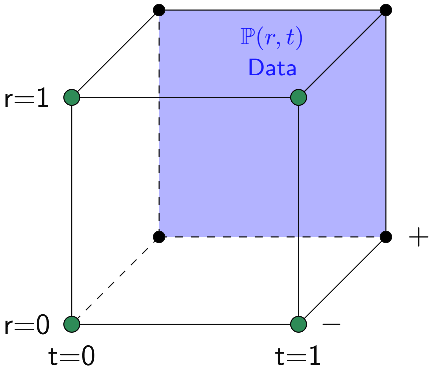
Multiplicity and Ambiguous Information

Data: $\mathbb{P}(r,t)$

Need: $\mathbb{P}(r,t,\pm)$ or $\mathbb{P}(t|\pm)$

The emergence of ambiguity:

- Multiple $\mathbb{P}(r,t,\pm) \in: \Pi$



Multiplicity and Ambiguous Information

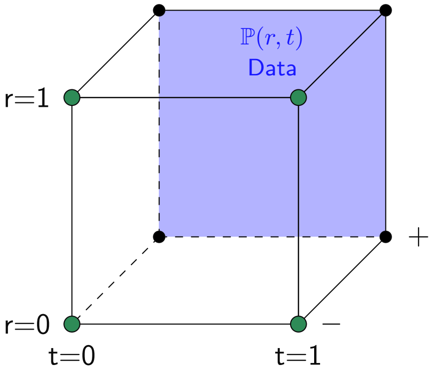
Data: $\mathbb{P}(r,t)$

Need: $\mathbb{P}(r,t,\pm)$ or $\mathbb{P}(t|\pm)$

The emergence of ambiguity:

- Multiple $\mathbb{P}(r,t,\pm) \in: \Pi$
- Multiple $\theta \in: \Theta$ possible
- One *ambiguous* experiment or

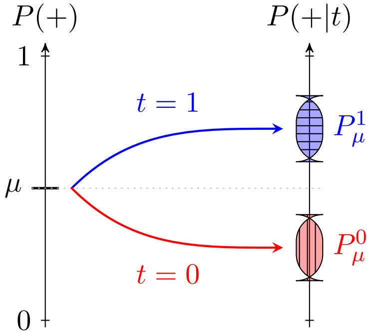
  a *set* of Blackwell experiments



Dilating test

For $\mu \in \Delta(\pm)$, define

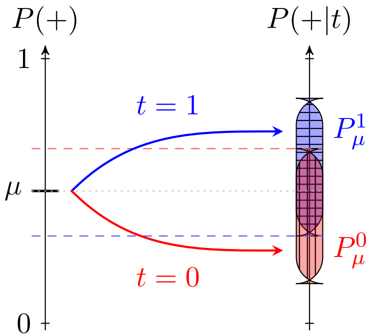
$$P_\mu^1 := \left\{ \frac{\mu\theta_1}{\mu\theta_1 + (1-\mu)(1-\theta_0)} : \theta \in \Theta \right\} \text{ and } P_\mu^0 := \left\{ \frac{\mu(1-\theta_1)}{\mu(1-\theta_1) + (1-\mu)\theta_0} : \theta \in \Theta \right\}$$



Dilating test

For $\mu \in \Delta(\pm)$, define

$$P^1_\mu := \left\{ \frac{\mu\theta_1}{\mu\theta_1 + (1-\mu)(1-\theta_0)} : \theta \in \Theta \right\} \text{ and } P^0_\mu := \left\{ \frac{\mu(1-\theta_1)}{\mu(1-\theta_1) + (1-\mu)\theta_0} : \theta \in \Theta \right\}$$



Definition

Fix a prior $\mu \in \Delta(\pm)$. The new test is a *μ-dilation* if $\{\mu\} \subsetneq P^0_\mu \cap P^1_\mu$.

It's a *dilation* if it's a μ-dilation for all $\mu \in \Delta(\pm)$.

Reference for what?

Observation

For any data $\mathbb{P}(r, t)$ and assuming non-triviality, the new test is a dilation.

Reference for what?

Observation

For any data $\mathbb{P}(r, t)$ and assuming non-triviality, the new test is a dilation.

Assumption (Knowledge of the reference test)

Reference test's sensitivity $\mathbb{P}(r = 1|+) =: s_1$ *and specificity* $\mathbb{P}(r = 0|-) =: s_0$ *are known.*

Reference for what?

Observation

For any data $\mathbb{P}(r,t)$ and assuming non-triviality, the new test is a dilation.

Assumption (Knowledge of the reference test)

Reference test's sensitivity $\mathbb{P}(r=1|+) =: s_1$ *and specificity* $\mathbb{P}(r=0|-) =: s_0$ *are known.*

Assumption (Informative reference test)

The reference test satisfies $s_1 + s_0 > 1$.

Characterization of the ambiguous experiment

Proposition

Fix data $\mathbb{P}(t, r)$. Under the assumptions the following hold:

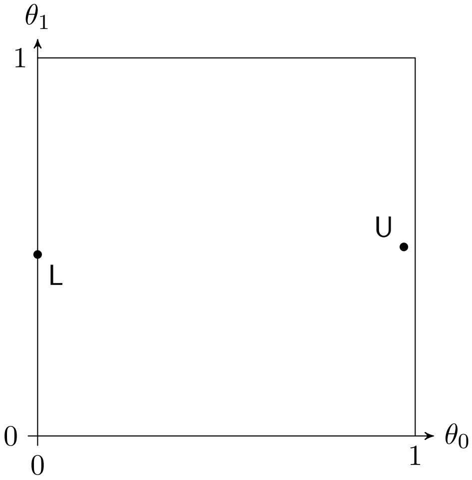
$$\Theta(s) := \Big\{(\theta_0, \theta_1) : \theta_1 \in [L_s, U_s], \mathbb{P}_s(-)\theta_0 - \mathbb{P}_s(+)\theta_1 = \mathbb{P}(t=0)\Big\}.$$

$$\mathbb{P}_s(+) = \frac{\mathbb{P}(r=1) + s_0 - 1}{s_1 + s_0 - 1} \qquad \mathbb{P}_s(-) = 1 - \mathbb{P}_s(+)$$

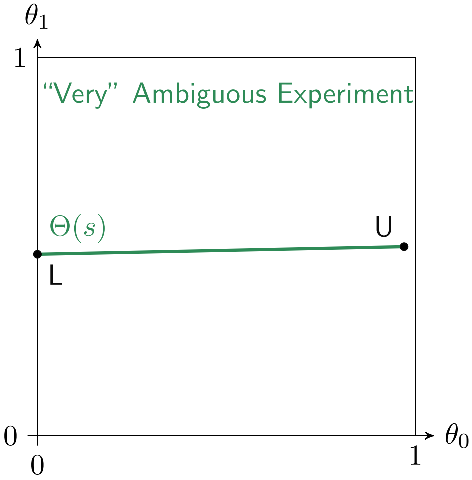
$$L_s = \frac{\big[\mathbb{P}(t=1, r=0) - s_0\mathbb{P}_s(-)\big]^+ + \big[\mathbb{P}(t=1, r=1) - (1-s_0)\mathbb{P}_s(-)\big]^+}{\mathbb{P}_s(+)}$$

$$U_s = \frac{\min\{\mathbb{P}(t=1, r=0), (1-s_1)\mathbb{P}_s(+)\} + \min\{\mathbb{P}(t=1, r=1), s_1\mathbb{P}_s(+)\}}{\mathbb{P}_s(+)}$$

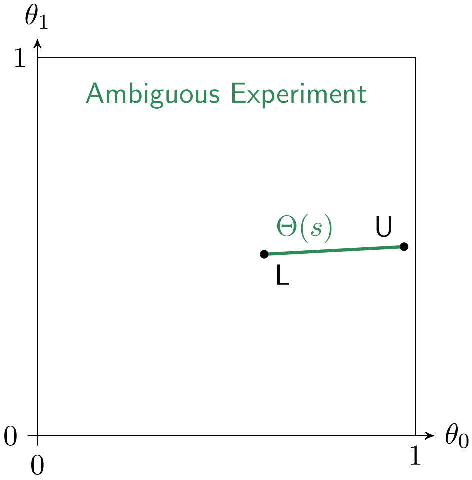
Visualizing the characterization



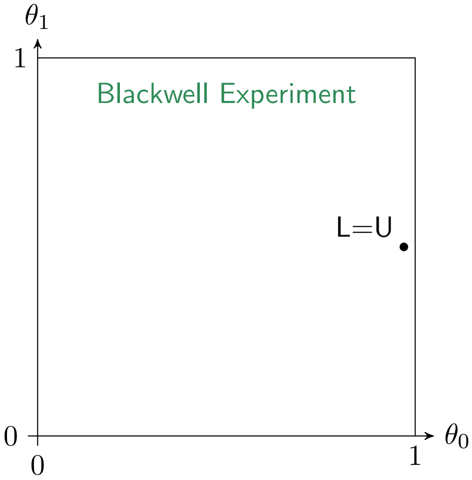
Visualizing the characterization



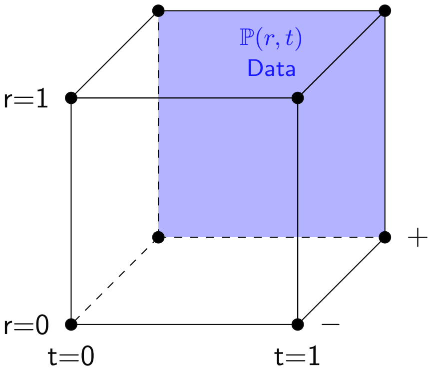
Visualizing the characterization



Visualizing the characterization

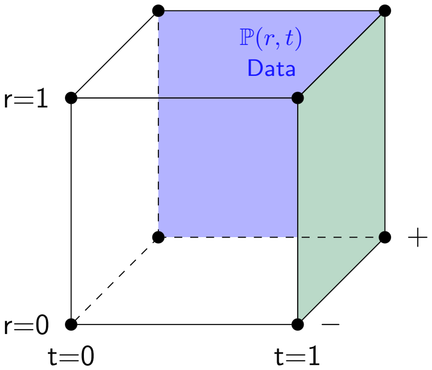


Proof



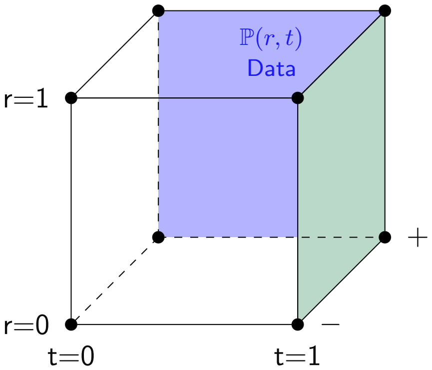
Proof

1. Informative reference $\implies \mathbb{P}(r, \pm)$



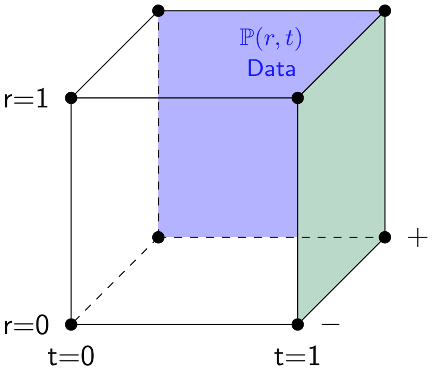
Proof

1. Informative reference $\implies$ $\mathbb{P}(r, \pm)$
2. Find $\mathbb{P}(t, r, \pm)$ with marginals $\mathbb{P}(r, \pm)$ and $\mathbb{P}(t, r)$



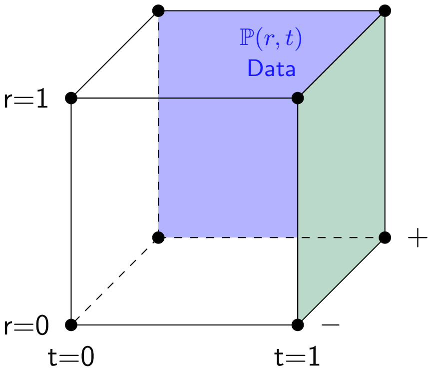
Proof

1. Informative reference $\implies \mathbb{P}(r, \pm)$
2. Find $\mathbb{P}(t, r, \pm)$ with marginals $\mathbb{P}(r, \pm)$ and $\mathbb{P}(t, r)$
3. A Copula problem:
   - Joe (1997, Theorem 3.10) gives set of 3-dim. distributions consistent with two 2-dim. marginals



Proof

1. Informative reference $\implies$ $\mathbb{P}(r, \pm)$
2. Find $\mathbb{P}(t, r, \pm)$ with marginals $\mathbb{P}(r, \pm)$ and $\mathbb{P}(t, r)$
3. A Copula problem:
   - Joe (1997, Theorem 3.10) gives set of 3-dim. distributions consistent with two 2-dim. marginals
4. U/L from most/least correlated distributions



The need for ambiguity

Recall: Definition of dilation required $\{\mu\} \subsetneq P_\mu^0 \cap P_\mu^1$

Observation

A Blackwell experiment (i.e. $|\Theta| = 1$) cannot be a dilation.

The need for ambiguity

Recall: Definition of dilation required $\{\mu\} \subsetneq P_\mu^0 \cap P_\mu^1$

Observation

A Blackwell experiment (i.e. $|\Theta| = 1$) cannot be a dilation.

Assumption (Everything goes)

Data is such that, for any $(j,k) \in \{0,1\}^2$, $\mathbb{P}(t = j, r = k) > 0$.

The need for ambiguity

Recall: Definition of dilation required $\{\mu\} \subsetneq P_\mu^0 \cap P_\mu^1$

Observation

A Blackwell experiment (i.e. $|\Theta| = 1$) cannot be a dilation.

Assumption (Everything goes)

Data is such that, for any $(j, k) \in \{0, 1\}^2$, $\mathbb{P}(t = j, r = k) > 0$.

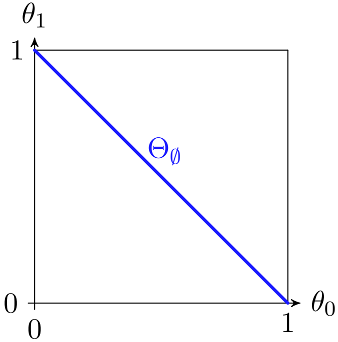
Lemma 4: Together with other assumptions, sufficient for proper ambiguity, *i.e.* $|\Theta| > 1$

Characterizing Dilations

Theorem

Under the mentioned assumptions,

t *is a dilation if and only if there exists* $\theta \in \Theta$ *such that* $\theta_1 + \theta_0 = 1$.

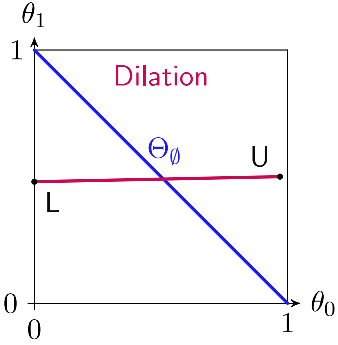


Characterizing Dilations

Theorem

Under the mentioned assumptions,

t *is a dilation if and only if there exists* $\theta \in \Theta$ *such that* $\theta_1 + \theta_0 = 1$.

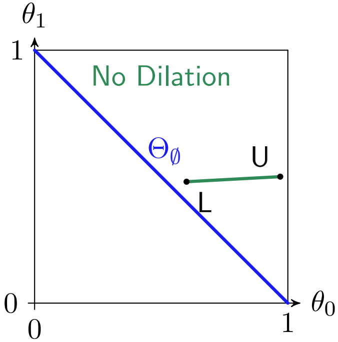


Characterizing Dilations

Theorem

Under the mentioned assumptions,

t is a dilation if and only if there exists $\theta \in \Theta$ such that $\theta_1 + \theta_0 = 1$.



Proof: Main steps

1. t is a μ-Dilation if and only if it is a dilation

Proof: Main steps

1. t is a μ-Dilation if and only if it is a dilation
2. dilation if and only if $\exists \theta, \theta' \in \Theta$ s.t. $\theta_1 + \theta_0 \geq 1$ and $\theta'_1 + \theta'_0 \leq 1$ with one inequality strict

   (*cf.* Seidenfeld and Wasserman, 1993, Theorem 2)

Proof: Main steps

1. t is a μ-Dilation if and only if it is a dilation
2. dilation if and only if $\exists \theta, \theta' \in \Theta$ s.t. $\theta_1 + \theta_0 \geq 1$ and $\theta'_1 + \theta'_0 \leq 1$ with one inequality strict

   (*cf.* Seidenfeld and Wasserman, 1993, Theorem 2)
3. Use path-connectedness and proper ambiguity to establish:

   t is a dilation if and only if there exists $\theta \in \Theta$ such that $\theta_1 + \theta_0 = 1$

Proof: Main steps

1. t is a μ-Dilation if and only if it is a dilation
2. dilation if and only if $\exists \theta, \theta' \in \Theta$ s.t. $\theta_1 + \theta_0 \geq 1$ and $\theta'_1 + \theta'_0 \leq 1$ with one inequality strict

   (*cf.* Seidenfeld and Wasserman, 1993, Theorem 2)

3. Use path-connectedness and proper ambiguity to establish:

   t is a dilation if and only if there exists $\theta \in \Theta$ such that $\theta_1 + \theta_0 = 1$

Corollary: Dilation if and only if $\mathbb{P}(t = 1) = \theta_1$ for some $\theta \in \Theta$.

That is, there is at most one intersection point with $\Theta_\emptyset$.

2 – Illustrative Applications

Application 1: Covid Antigen test (LFT) from Leber et al. (2021)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	179	38	217
$t = 1$	22	788	819
Sum	201	826	1027

- "Patients with ... flu-like symptoms attending a general practice network in an Austrian district" – Leber et al. (2021, p.1)
- Apparent Sensitivity: $\mathbb{P}(t = 1|r = 1) = 95.4\%$
- Apparent Specificity: $\mathbb{P}(t = 0|r = 0) = 89.1\%$

Application 1: Covid Antigen test (LFT) from Leber et al. (2021)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	179	38	217
$t = 1$	22	788	819
Sum	201	826	1027

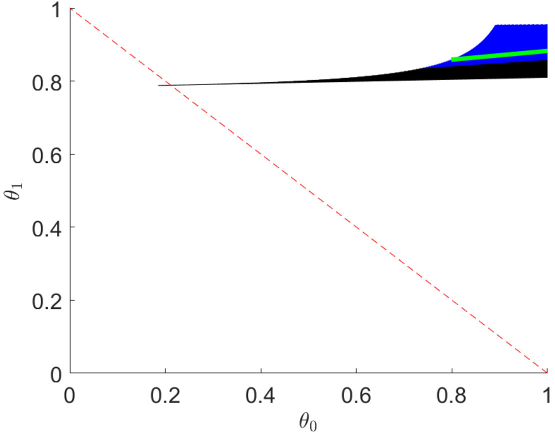
- "Patients with … flu-like symptoms attending a general practice network in an Austrian district" – Leber et al. (2021, p.1)
- Apparent Sensitivity: $\mathbb{P}(t = 1 | r = 1) = 95.4\%$
- Apparent Specificity: $\mathbb{P}(t = 0 | r = 0) = 89.1\%$
- Suppose $s_0 = 100\%$ and $s_1 \in$

  Green: $[90\%, 90\%]$

  Blue: $[87.5\%, 100\%]$

  Black: $[82.5\%, 100\%]$

Application 1: Covid Antigen test (LFT) from Leber et al. (2021)



- "Patients with ... flu-like symptoms attending a general practice network in an Austrian district" – Leber et al. (2021, p.1)
- Apparent Sensitivity: $\mathbb{P}(t = 1|r = 1) = 95.4\%$
- Apparent Specificity: $\mathbb{P}(t = 0|r = 0) = 89.1\%$
- Suppose $s_0 = 100\%$ and $s_1 \in$

  Green: $[90\%, 90\%]$

  Blue: $[87.5\%, 100\%]$

  Black: $[82.5\%, 100\%]$

Application 2: CT-Chest scan for Covid from Ai et al. (2020)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	105	21	126
$t = 1$	308	580	888
Sum	413	601	1014

- Apparent Sensitivity: $\mathbb{P}(t = 1 | r = 1) = 96.5\%$
- Apparent Specificity: $\mathbb{P}(t = 0 | r = 0) = 25.4\%$
- "Chest CT may be considered as a primary tool for the current COVID-19 detection"
  – Ai et al. (2020, p.1)

Application 2: CT-Chest scan for Covid from Ai et al. (2020)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	105	21	126
$t = 1$	308	580	888
Sum	413	601	1014

- Apparent Sensitivity: $\mathbb{P}(t = 1 | r = 1) = 96.5\%$
- Apparent Specificity: $\mathbb{P}(t = 0 | r = 0) = 25.4\%$
- "Chest CT may be considered as a primary tool for the current COVID-19 detection"

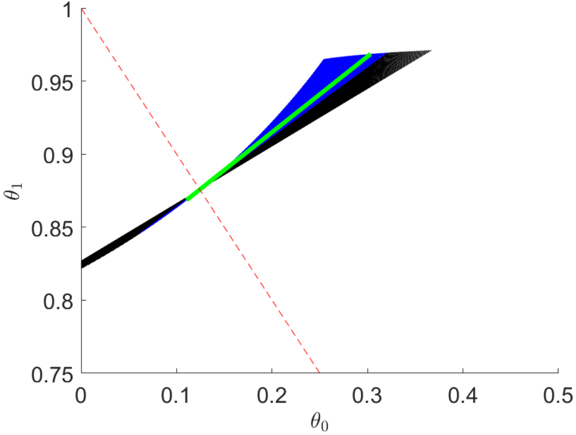
  – Ai et al. (2020, p.1)
- Suppose $s_0 = 100\%$ and $s_1 \in$

  Green: $[90\%, 90\%]$

  Blue: $[87.5\%, 100\%]$

  Black: $[82.5\%, 100\%]$

Application 2: CT-Chest scan for Covid from Ai et al. (2020)



- Apparent Sensitivity: $\mathbb{P}(t = 1|r = 1) = 96.5\%$
- Apparent Specificity: $\mathbb{P}(t = 0|r = 0) = 25.4\%$
- "Chest CT may be considered as a primary tool for the current COVID-19 detection"

  – Ai et al. (2020, p.1)
- Suppose $s_0 = 100\%$ and $s_1 \in$

  Green: $[90\%, 90\%]$

  Blue: $[87.5\%, 100\%]$

  Black: $[82.5\%, 100\%]$

3 – Statistical Inference

Statistical Test(s) for Dilations

Hypothesis: $H_0 =$ new test is a dilation ($\iff \exists(\theta_1, \theta_0) \in \Theta : \theta_1 + \theta_0 = 1$)

First, somewhat naive, attempt based on Goodman (1965)

new \ ref.	$r = 0$	$r = 1$
$t = 0$	p_{00}	p_{01}
$t = 1$	p_{10}	p_{11}

- Data $\sim$ multinomial distribution
- Goodman (1965): confidence set for parameters
- Get CI for $\theta_0 + \theta_1$, reject if 1 not in it

Statistical Test(s) for Dilations

Hypothesis: $H_0 =$ new test is a dilation ($\iff \exists(\theta_1, \theta_0) \in \Theta : \theta_1 + \theta_0 = 1$)

First, somewhat naive, attempt based on Goodman (1965)

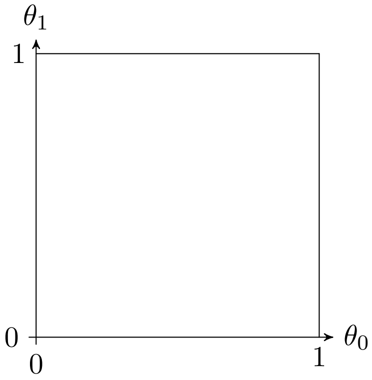
new \ ref.	$r = 0$	$r = 1$
$t = 0$	p_{00}	p_{01}
$t = 1$	p_{10}	p_{11}

- Data $\sim$ multinomial distribution
- Goodman (1965): confidence set for parameters
- Get CI for $\theta_0 + \theta_1$, reject if 1 not in it
- Issues:
  - Conservative
  - Low power
  - Uniformity problems

Statistical Test(s) for Dilations

Hypothesis: $H_0 =$ new test is a dilation ($\iff \exists(\theta_1, \theta_0) \in \Theta : \theta_1 + \theta_0 = 1$)

Key insight from Obradović (2024): Θ can be expressed via **moment inequalities**

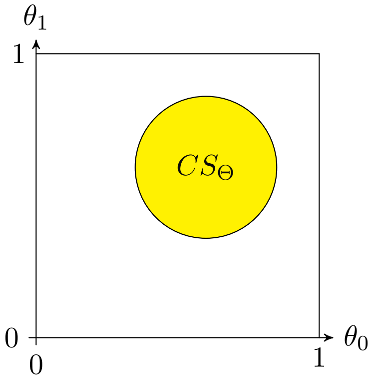


1. Based on Romano, Shaikh, and Wolf (2014):
   - Given data, search across all $\theta \in [0, 1]^2$

Statistical Test(s) for Dilations

Hypothesis: $H_0 =$ new test is a dilation ($\iff \exists(\theta_1, \theta_0) \in \Theta : \theta_1 + \theta_0 = 1$)

Key insight from Obradović (2024): Θ can be expressed via **moment inequalities**

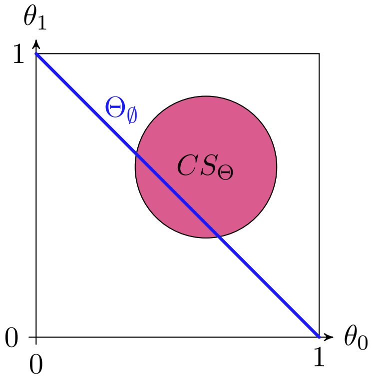


1. Based on Romano, Shaikh, and Wolf (2014):
   - Given data, search across all $\theta \in [0,1]^2$
   - Include $\theta \in CS_\Theta$ if all inequalities are satisfied

Statistical Test(s) for Dilations

Hypothesis: $H_0 =$ new test is a dilation ($\iff \exists(\theta_1, \theta_0) \in \Theta : \theta_1 + \theta_0 = 1$)

Key insight from Obradović (2024): Θ can be expressed via **moment inequalities**

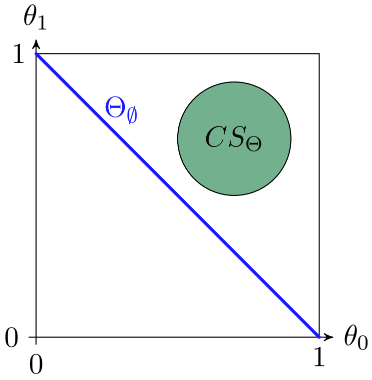


1. Based on Romano, Shaikh, and Wolf (2014):
   - Given data, search across all $\theta \in [0, 1]^2$
   - Include $\theta \in CS_\Theta$ if all inequalities are satisfied
   - Reject if $CS_\Theta \cap \Theta_\emptyset = \emptyset$

Statistical Test(s) for Dilations

Hypothesis: $H_0 =$ new test is a dilation ($\iff \exists(\theta_1, \theta_0) \in \Theta : \theta_1 + \theta_0 = 1$)

Key insight from Obradović (2024): Θ can be expressed via **moment inequalities**

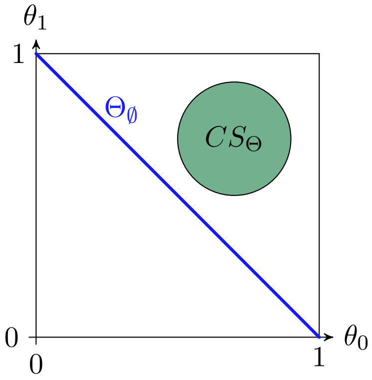


1. Based on Romano, Shaikh, and Wolf (2014):
   - Given data, search across all $\theta \in [0,1]^2$
   - Include $\theta \in CS_\Theta$ if all inequalities are satisfied
   - Reject if $CS_\Theta \cap \Theta_\emptyset = \emptyset$

Statistical Test(s) for Dilations

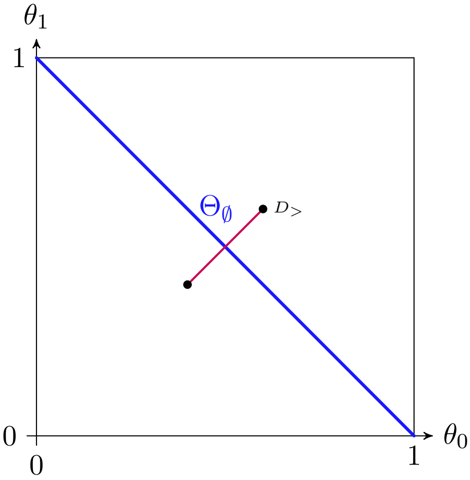
Hypothesis: $H_0 =$ new test is a dilation ($\iff \exists(\theta_1, \theta_0) \in \Theta : \theta_1 + \theta_0 = 1$)

Key insight from Obradović (2024): Θ can be expressed via **moment inequalities**



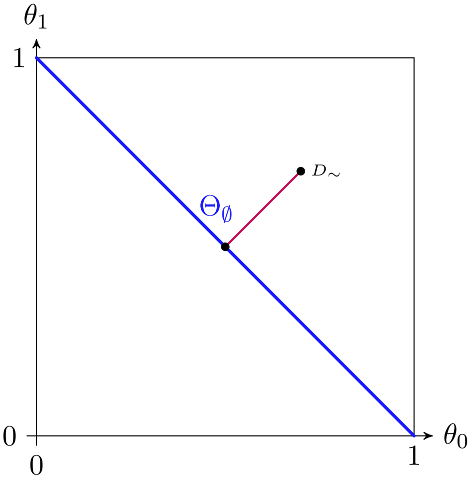
1. Based on Romano, Shaikh, and Wolf (2014):
   - Given data, search across all $\theta \in [0, 1]^2$
   - Include $\theta \in CS_\Theta$ if all inequalities are satisfied
   - Reject if $CS_\Theta \cap \Theta_\emptyset = \emptyset$
   - ... but only need to look across $\Theta_\emptyset$
2. Subvector inference:
   - based on Bugni, Canay, and Shi (2017)
   - better coverage and power

Simulations



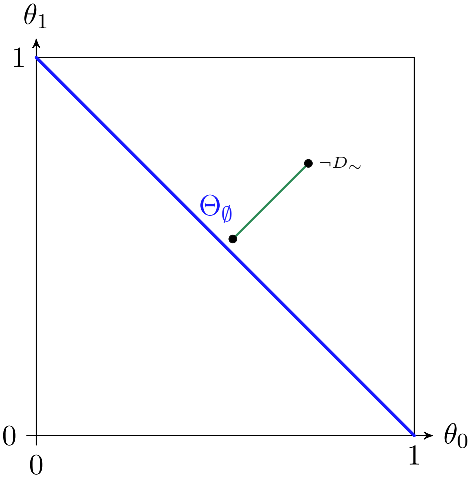
- 5 specifications with $s = (90\%, 90\%)$:
  - 2 dilations D

Simulations



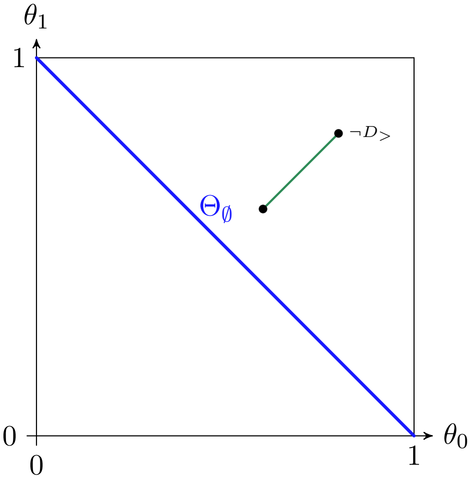
- 5 specifications with $s = (90\%, 90\%)$:
  - 2 dilations D

Simulations



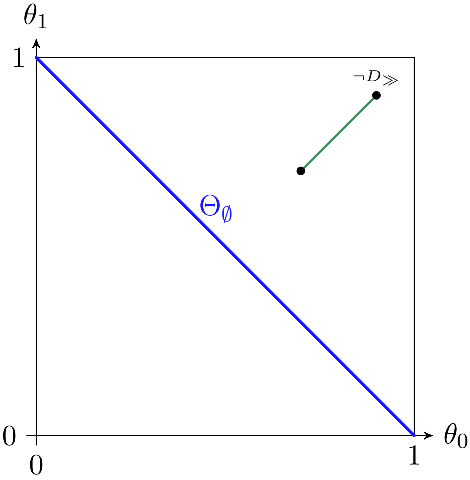
- 5 specifications with $s = (90\%, 90\%)$:
  - 2 dilations D
  - 3 non-dilations $\neg D$

Simulations



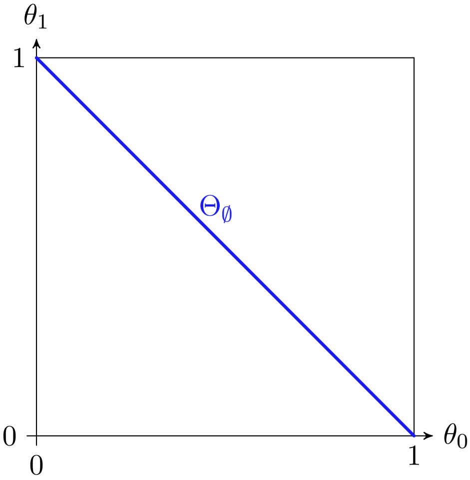
- 5 specifications with $s = (90\%, 90\%)$:
  - 2 dilations D
  - 3 non-dilations $\neg D$

Simulations



- 5 specifications with $s = (90\%, 90\%)$:
  - 2 dilations D
  - 3 non-dilations $\neg D$

Simulations



- 5 specifications with $s = (90\%, 90\%)$:
  - 2 dilations D
  - 3 non-dilations $\neg D$
- Sample Size:
  - $N \in \{50, 100, 500\}$
- Simulations replications: 1000

Simulations - Rejection Probabilities ($\alpha = 5\%$)

		$D_>$	$D_\sim$	$\neg D_\sim$	$\neg D_>$	$\neg D_\gg$
	G	0.9%	0.6%			
$N = 50$	RSW	0%	0%			
	BCS	0%	0.8%			
	G	0%	0%			
$N = 100$	RSW	0%	0%			
	BCS	0%	1%			
	G	0%	0%			
$N = 500$	RSW	0%	0%			
	BCS	0%	1.3%			

Simulations - Rejection Probabilities ($\alpha = 5\%$)

		$D_{>}$	$D_{\sim}$	$\neg D_{\sim}$	$\neg D_{>}$	$\neg D_{\gg}$
	G	0.9%	0.6%	0.2%	0.7%	4.3%
$N = 50$	RSW	0%	0%	0%	0.1%	4.3%
	BCS	0%	0.8%	3.9%	22%	76%
	G	0%	0%	0%	1%	30%
$N = 100$	RSW	0%	0%	0%	0.1%	19%
	BCS	0%	1%	3.1%	43%	100%
	G	0%	0%	0%	50%	100%
$N = 500$	RSW	0%	0%	0%	28%	100%
	BCS	0%	1.3%	9.2%	99%	100%

Moment (in)equalities

Fix $P = \mathbb{P}(r, t)$. Obradović (2024) shows that $\Theta(s)$ is can be rewritten as

$$\tilde{\Theta}_P(s) = \left\{\theta \in [0,1]^2 : \mathbb{E}_P[m_j(r,t;\theta,s)] \leq 0, j = 1, \ldots, 6 \text{ and } \mathbb{E}_P[m_j(r,t;\theta,s)] = 0\right\},$$

where $m_j : \{0,1\}^2 \times [0,1]^2 \times \mathcal{S} \to \mathbb{R}$ are *moment functions* for $j = 1, \ldots, 7$.

Moment (in)equalities

Fix $P = \mathbb{P}(r,t)$. Obradović (2024) shows that $\Theta(s)$ is can be rewritten as

$$\tilde{\Theta}_P(s) = \left\{\theta \in [0,1]^2 : \mathbb{E}_P[m_j(r,t;\theta,s)] \leq 0, j = 1,\ldots,6 \text{ and } \mathbb{E}_P[m_j(r,t;\theta,s)] = 0\right\},$$

where $m_j : \{0,1\}^2 \times [0,1]^2 \times \mathcal{S} \to \mathbb{R}$ are *moment functions* for $j = 1,\ldots,7$.

▷ **Moment Inequality Model**

Moment (in)equalities

Fix $P = \mathbb{P}(r, t)$. Obradović (2024) shows that $\Theta(s)$ is can be rewritten as

$$\tilde{\Theta}_P(s) = \left\{\theta \in [0,1]^2 : \mathbb{E}_P[m_j(r,t;\theta,s)] \leq 0, j = 1, \ldots, 6 \text{ and } \mathbb{E}_P[m_j(r,t;\theta,s)] = 0\right\},$$

where $m_j : \{0,1\}^2 \times [0,1]^2 \times \mathcal{S} \to \mathbb{R}$ are *moment functions* for $j = 1, \ldots, 7$.

- **Moment Inequality Model**

We are interested in testing the hypothesis that H_0: $\theta_1 + \theta_0 = 1$, i.e. based on a linear function of the parameters.

- **Subvector Inference**
- Bugni, Canay, and Shi (2017) propose bootstrap-based inference method

Assumptions

We consider a (relatively large) class of distributions $\mathcal{P} \subset \Delta^3 \subset \mathbb{R}^4$.

Assumptions

We consider a (relatively large) class of distributions $\mathcal{P} \subset \Delta^3 \subset \mathbb{R}^4$.

First, we impose standard random sampling:

Assumption

For every $P \in \mathcal{P}$, the study sample is a n-sequence of i.i.d. random vectors $W_i = (r_i, t_i) \sim P$.

Assumptions

We consider a (relatively large) class of distributions $\mathcal{P} \subset \Delta^3 \subset \mathbb{R}^4$.

First, we impose standard random sampling:

Assumption

For every $P \in \mathcal{P}$, the study sample is a n-sequence of i.i.d. random vectors $W_i = (r_i, t_i) \sim P$.

Then, strengthen identification assumptions to avoid (well-known) uniformity problems:

- Without: For any n, there might exist $P \in \mathcal{P}$ rendering confidence regions invalid.
- Often better finite sample performance if uniformity holds.
- Especially important for partially identified models.

Stronger Assumptions

Want the probability of infection to be strictly (and uniformly) bounded away from boundary:

Assumption

There exists $\varepsilon_y \in \left(0, \frac{1}{2}\right)$ *such that for all* $P \in \mathcal{P}$*, we have* $\mathbb{P}_s(+) \in [\varepsilon_y, 1 - \varepsilon_y]$.

Stronger Assumptions

Want the probability of infection to be strictly (and uniformly) bounded away from boundary:

Assumption

There exists $\varepsilon_y \in \left(0, \frac{1}{2}\right)$ *such that for all* $P \in \mathcal{P}$*, we have* $\mathbb{P}_s(+) \in [\varepsilon_y, 1 - \varepsilon_y]$.

Strengthen "Everything goes" assumption similarly:

Assumption

There exists an $\varepsilon_d \in \left(0, \frac{1}{4}\right)$ *s.t. for every* $P \in \mathcal{P}$ *and every* $(t, r) \in \{0, 1\}^2$, $P(t, r) \geq \varepsilon_d$ *holds.*

Asymptotic Size Control

Recall: Size is $\mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) = \mathbb{P}(\text{Type 1 error})$.

Asymptotic Size Control

Recall: Size is $\mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) = \mathbb{P}(\text{Type 1 error})$.

- We use the following estimating function to construct test proposed in Bugni et al (2017):

$$Q_P(\theta, s) = \sum_{j=1}^{6} \left[\max\left\{ 0, \frac{\mathbb{E}_P m_j(W, \theta, s)}{\sigma_{P,j}(\theta, s)} \right\} \right]^2 + \left[\frac{\mathbb{E}_P m_7(W, \theta, s)}{\sigma_{P,j}(\theta, s)} \right]^2$$

$$\sigma^2_{P,j}(\theta, s) := \mathbb{V}_P[m_7(W_i, \theta, s)],$$

with (scaled) plug-in estimators.

Asymptotic Size Control

Recall: Size is $\mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) = \mathbb{P}(\text{Type 1 error})$.

- We use the following estimating function to construct test proposed in Bugni et al (2017):

$$Q_P(\theta, s) = \sum_{j=1}^{6} \left[\max\left\{ 0, \frac{\mathbb{E}_P m_j(W, \theta, s)}{\sigma_{P,j}(\theta, s)} \right\} \right]^2 + \left[\frac{\mathbb{E}_P m_7(W, \theta, s)}{\sigma_{P,j}(\theta, s)} \right]^2$$

$$\sigma^2_{P,j}(\theta, s) := \mathbb{V}_P[m_7(W_i, \theta, s)],$$

  with (scaled) plug-in estimators.

- We show Bugni et al (2017, Thm 4.1) applies, so our test asympt. and unif. controls size:

$$\limsup_{n\to\infty} \sup_{(\theta,P):H_0 \text{ is true}} \mathbb{P}(\text{Type 1 error}) \leq \alpha.$$

4 – Applications

CT-Chest scan for Covid from Ai et al. (2020)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	105	21	126
$t = 1$	308	580	888
Sum	413	601	1014

- ▷ Suppose $s_0 = 100\%$ and $s_1 = 90\%$
- ▷ Set nominal significance level at $\alpha = 5\%$

CT-Chest scan for Covid from Ai et al. (2020)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	105	21	126
$t = 1$	308	580	888
Sum	413	601	1014

- Suppose $s_0 = 100\%$ and $s_1 = 90\%$
- Set nominal significance level at $\alpha = 5\%$
- Our proposed test:

  Test statistic $T_n = 1.2510^{-18}$

  Critical value $\hat{c}_n^{95\%} = 1.11$

CT-Chest scan for Covid from Ai et al. (2020)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	105	21	126
$t = 1$	308	580	888
Sum	413	601	1014

- Suppose $s_0 = 100\%$ and $s_1 = 90\%$
- Set nominal significance level at $\alpha = 5\%$
- Our proposed test:

  Test statistic $T_n = 1.2510^{-18}$

  Critical value $\hat{c}_n^{95\%} = 1.11$

  Cannot reject $H0$ of it being a dilation

  (p-value $> 99\%$)

Deep Neural Net to Predict Loan Approval (Abakarim et al. '18)

- Predict "good" ($t = 1$) vs. "bad" risk ($t = 0$)

Deep Neural Net to Predict Loan Approval (Abakarim et al. ’18)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	203	43	246
$t = 1$	97	657	754
Sum	300	700	1000

- Predict “good” $(t = 1)$ vs. “bad” risk $(t = 0)$
- Performance measure relative to historic data

  Hofmann (1994)

  known to contain errors

  and historic data $\neq$ ground truth

Deep Neural Net to Predict Loan Approval (Abakarim et al. '18)

new \ ref.	$r=0$	$r=1$	Sum
$t=0$	203	43	246
$t=1$	97	657	754
Sum	300	700	1000

- Predict "good" ($t=1$) vs. "bad" risk ($t=0$)
- Performance measure relative to historic data

  Hofmann (1994)

  known to contain errors

  and historic data $\neq$ ground truth
- Assume $\alpha = 5\%$ and $s = (0.9, 0.9)$

  Reject $H0$: $T_n = 35.96$ vs. $\hat{c}_n^{95\%} = 5.25$

Deep Neural Net to Predict Loan Approval (Abakarim et al. '18)

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	203	43	246
$t = 1$	97	657	754
Sum	300	700	1000

- Predict "good" ($t = 1$) vs. "bad" risk ($t = 0$)
- Performance measure relative to historic data

  Hofmann (1994)

  known to contain errors

  and historic data $\neq$ ground truth
- Assume $\alpha = 5\%$ and $s = (0.9, 0.9)$

  Reject $H0$: $T_n = 35.96$ vs. $\hat{c}_n^{95\%} = 5.25$

  Also reject with $s = (0.95, 0.95)$

5 – Extensions and Conclusion

Extensions: Partial Knowledge about Reference Test

- What if s is not known exactly?

Extensions: Partial Knowledge about Reference Test

- What if s is not known exactly?
- If *set* $\mathcal{S}$ is known and satisfies some regularity conditions:
  - Identification generalization

Extensions: Partial Knowledge about Reference Test

- What if s is not known exactly?
- If *set* $\mathcal{S}$ is known and satisfies some regularity conditions:
  - Identification generalization
  - Moment inequality/subvector inference too

Extensions: Partial Knowledge about Reference Test

- What if s is not known exactly?
- If *set* $\mathcal{S}$ is known and satisfies some regularity conditions:
  - Identification generalization
  - Moment inequality/subvector inference too
  - However, we were not able (yet) to prove size control

Extensions: No Knowledge about Reference Test

- What if $\mathcal{S}$ is not known at all?

Extensions: No Knowledge about Reference Test

- What if $\mathcal{S}$ is not known at all?
- Given data $\rightarrow$ find $s = (s_0, s_1)$ for dilation

Extensions: No Knowledge about Reference Test

- What if $\mathcal{S}$ is not known at all?
- Given data $\rightarrow$ find $s = (s_0, s_1)$ for dilation
- Thinking of $\Theta(s)$ as correspondence: $\mathcal{D} := \Theta^-(\Theta_\emptyset)$
- Again, a moment inequality approach is applicable

Illustration 1: Independent tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	25%	25%	50%
$t = 1$	25%	25%	50%
Sum	50%	50%	

Illustration 1: Independent tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	25%	25%	50%
$t = 1$	25%	25%	50%
Sum	50%	50%	

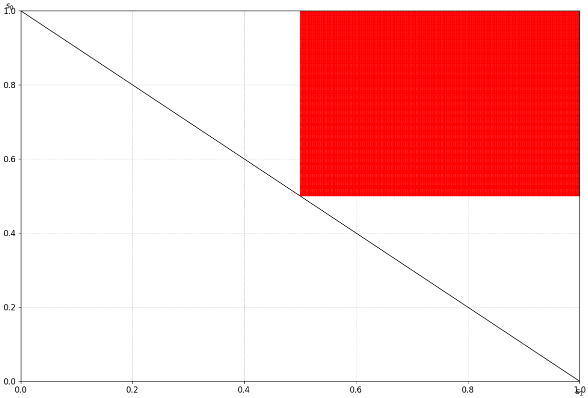


Illustration 2: Weakly correlated tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	39%	11%	50%
$t = 1$	11%	39%	50%
Sum	50%	50%	

Illustration 2: Weakly correlated tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	39%	11%	50%
$t = 1$	11%	39%	50%
Sum	50%	50%	

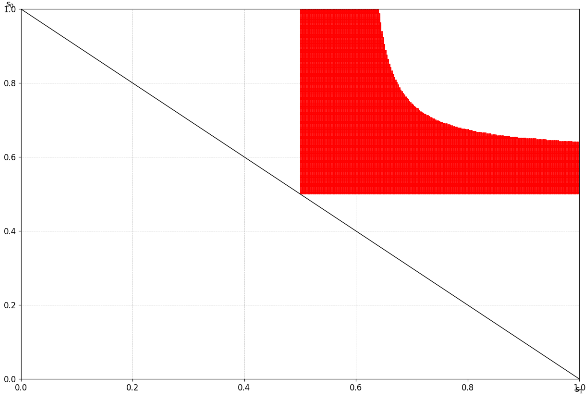
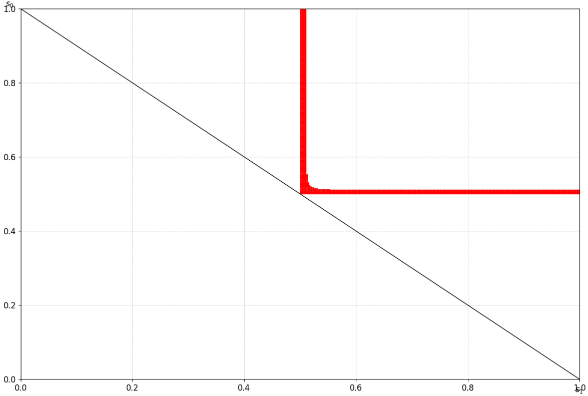


Illustration 3: Highly correlated tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	49%	1%	50%
$t = 1$	1%	49%	50%
Sum	50%	50%	

Illustration 3: Highly correlated tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	49%	1%	50%
$t = 1$	1%	49%	50%
Sum	50%	50%	



Policy implication: Regulation of medical tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	388	120	508
$t = 1$	12	480	492
Sum	400	600	1000

- If reference test is perfect:
  - 97% specificity and
  - 80% sensitivity

Policy implication: Regulation of medical tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	388	120	508
$t = 1$	12	480	492
Sum	400	600	1000

- If reference test is perfect:
  - 97% specificity and
  - 80% sensitivity
- **BUT:** even if perfect specificity of reference

Policy implication: Regulation of medical tests

new \ ref.	$r = 0$	$r = 1$	Sum
$t = 0$	388	120	508
$t = 1$	12	480	492
Sum	400	600	1000

- If reference test is perfect:
  - 97% specificity and
  - 80% sensitivity
- **BUT:** even if perfect specificity of reference

  if s_1 close to 62%

  we have a dilation

Further research

- Partial knowledge of correlation, *e.g.* local robustness around independence

Further research

- Partial knowledge of correlation, *e.g.* local robustness around independence
- Beyond Binary Tests
    - Some tests are not just binary
    - Copula idea generalizes, closed-form expressions messier

Further research

- Partial knowledge of correlation, *e.g.* local robustness around independence
- Beyond Binary Tests
  - Some tests are not just binary
  - Copula idea generalizes, closed-form expressions messier
- Sequential testing? I.e. take both tests

Further research

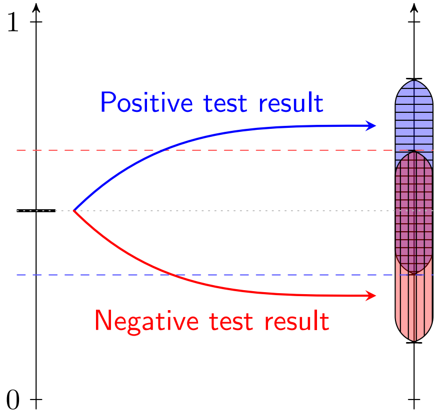
- Partial knowledge of correlation, *e.g.* local robustness around independence
- Beyond Binary Tests
  - Some tests are not just binary
  - Copula idea generalizes, closed-form expressions messier
- Sequential testing? I.e. take both tests
- Would be interesting to run experiment with medical professionals, for example, with test similar to CT-chest scan

Further research

- Partial knowledge of correlation, *e.g.* local robustness around independence
- Beyond Binary Tests
  - Some tests are not just binary
  - Copula idea generalizes, closed-form expressions messier
- Sequential testing? I.e. take both tests
- Would be interesting to run experiment with medical professionals, for example, with test similar to CT-chest scan
- Theory of ambiguous information

Summary

Do we encounter dilations in (existing?) classifiers?



Thanks! – Gabriel Ziegler, U Edinburgh, ziegler@ed.ac.uk