Credal Two-Sample Tests

Hypothesis Testing under Dataset Uncertainty

Krikamol Muandet | OIST Machine Learning Workshop | March 3-5, 2025

Contents

Dataset Uncertainty Example #1: Distribution Shift

≠

New York, USA

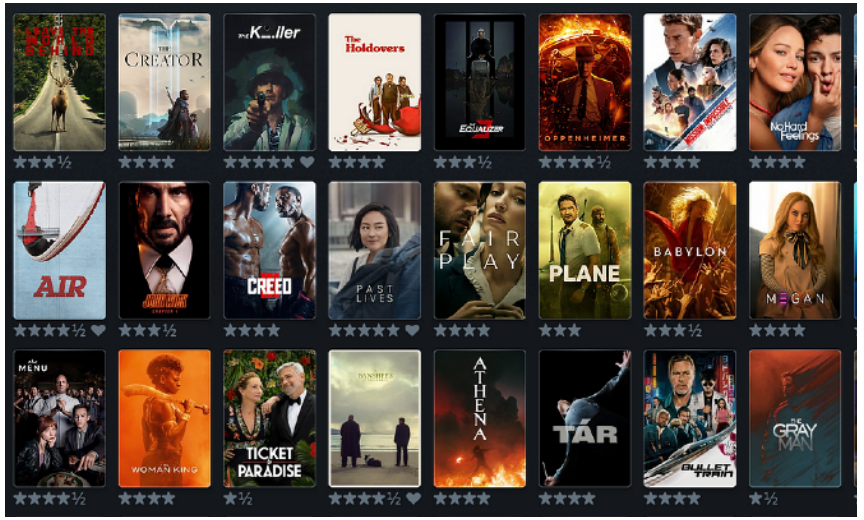
Bangkok, Thailand

Example #2: Multiple Relevant Datasets acting as Proxies

Example #2: Multiple Relevant Datasets acting as Proxies

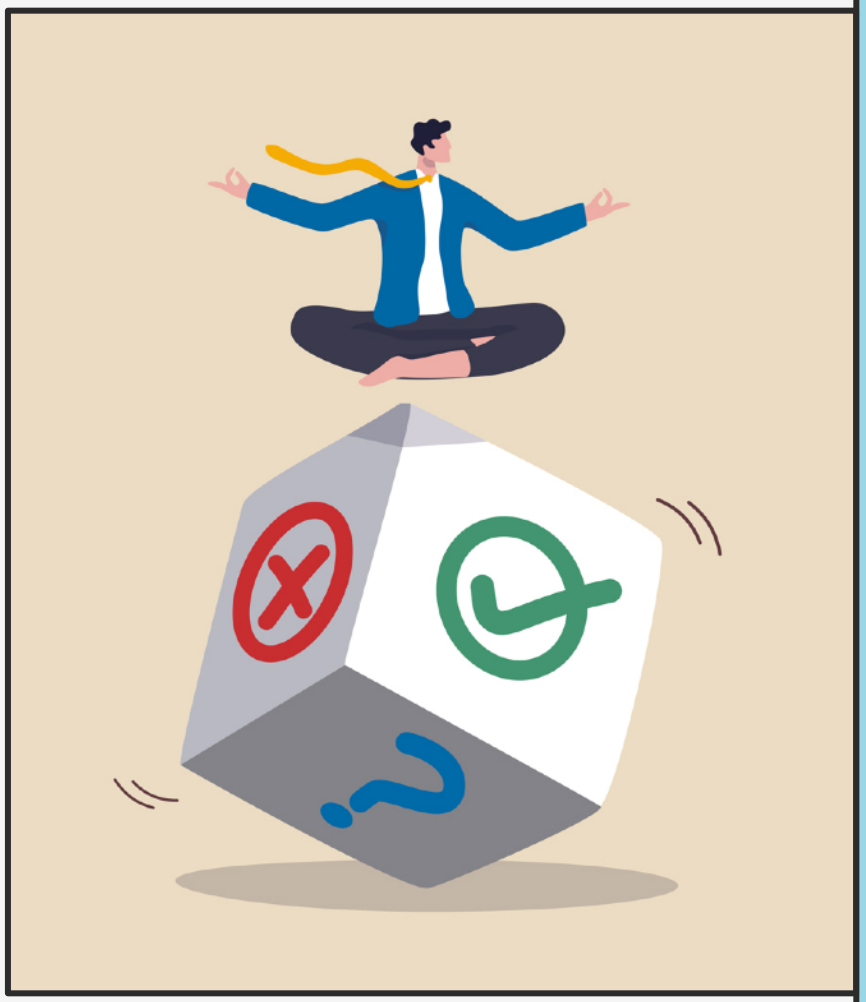
Example #3: Collective Decision-Making

What should we watch tonight?



Rational preferences over items induces distribution (Savage 1971)

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. JASA



How should we test hypothesis under dataset uncertainty?

- *Review on why hypothesis testing is important.*
- *How dataset uncertainty can be part of the hypothesis?*
- *How to actually conduct the test?*

Credal Two-Sample Tests of Epistemic Uncertainty

Siu Lun Chau
CISPA —> NTU

Antonin Schrab
Gatsby Unit, UCL

Arthur Gretton
Gatsby Unit, UCL

Dino Sejdinovic
University of Adelaide

Krikamol Muandet
CISPA

Credal Two-Sample Tests of Epistemic Uncertainty

Siu Lun Chau[1] **Antonin Schrab**[2,3] **Arthur Gretton**[3] **Dino Sejdinovic**[4] **Krikamol Muandet**[1]

[1]Rational Intelligence Lab, CISPA Helmholtz Center for Information Security, 66123 Saarbrücken, Germany
[2]Centre for Artificial Intelligence, University College London & Inria London, London, WC1V 6LJ, United Kingdom
[3]Gatsby Computational Neuroscience Unit, University College London, London, W1T 4JG, United Kingdom
[4]School of Computer and Mathematical Sciences & AIML, University of Adelaide, Adelaide, SA 5005, Australia

Abstract

We introduce credal two-sample testing, a new hypothesis testing framework for com-

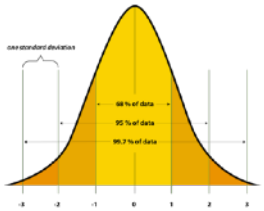
uncertainty (AU), which refers to inherent variability, and *epistemic uncertainty* (EU), arising from limited information such as finite data or model assumptions (Hora, 1996). These uncertainties often overlap, as

What is Statistical Hypothesis Testing?

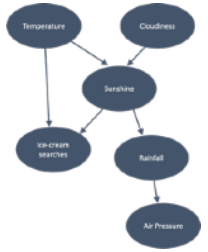
Given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, we wish to evaluate whether observations provide **sufficient evidence** to reject a default assumption about P.

Examples

A/B Testing



Goodness-of-fit test



Causal Discovery

Treatment Effect Test

What is Statistical Hypothesis Testing?

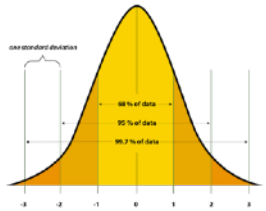
Given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, we wish to evaluate whether observations provide **sufficient evidence** to reject a default assumption about P.

$$\mathscr{H}_0 : P \in \mathscr{P}_0 \subseteq \mathscr{P} \quad \text{v.s.} \quad \mathscr{H}_1 : P \in \mathscr{P}_1 \subseteq \mathscr{P}$$

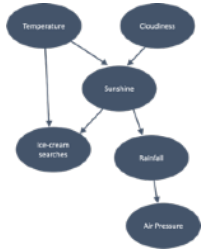
Null Hypothesis — Alternative Hypothesis

Examples

A/B Testing



Goodness-of-fit test



Causal Discovery

Treatment Effect Test

What is Statistical Hypothesis Testing?

Given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, we wish to evaluate whether observations provide **sufficient evidence** to reject a default assumption about P.

$$\mathscr{P}_0 \cap \mathscr{P}_1 = \varnothing$$

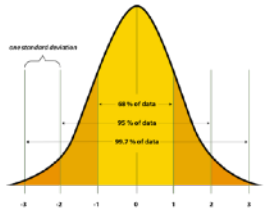
$\mathscr{H}_0 : P \in \mathscr{P}_0 \subseteq \mathscr{P}$ v.s. $\mathscr{H}_1 : P \in \mathscr{P}_1 \subseteq \mathscr{P}$

Null Hypothesis

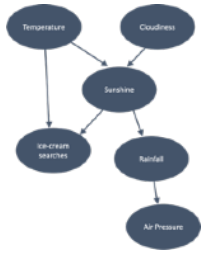
Alternative Hypothesis

Examples

A/B Testing



Goodness-of-fit test



Causal Discovery

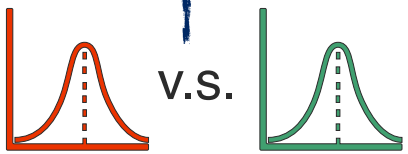
Treatment Effect Test

What is Statistical Hypothesis Testing?

Given $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, we wish to evaluate whether observations provide **sufficient evidence** to reject a default assumption about P.

Typical Statistical Testing Pipeline

1) Formulate **hypothesis**



2) Compute **test statistic**

$T(X_{1:n})$ should be small when the null holds

3) Return **p-value**

$\mathbb{P}(T(\mathbf{X}) \geq T(X_{1:n}) \mid H_0)$

4) Reject or fail to reject the null

What is Statistical Hypothesis Testing?

An ***ideal*** testing procedure should yield small

False Positives (Type I error) and False Negatives (Type II error)

What is Statistical Hypothesis Testing?

An ***ideal*** testing procedure should yield small

False Positives (Type I error) **and** **False Negatives (Type II error)**

$$\mathbb{P}(\text{Reject } \mathscr{H}_0 \mid \mathscr{H}_0)$$

What is Statistical Hypothesis Testing?

An ***ideal*** testing procedure should yield small

False Positives (Type I error) and False Negatives (Type II error)

$\mathbb{P}(\text{Reject } \mathscr{H}_0 \mid \mathscr{H}_0)$

$\mathbb{P}(\text{Fail to reject } \mathscr{H}_0 \mid \mathscr{H}_1)$

What is Statistical Hypothesis Testing?

An *ideal* testing procedure should yield small

False Positives (Type I error) and False Negatives (Type II error)

$$\mathbb{P}(\text{Fail to reject } \mathscr{H}_0 \mid \mathscr{H}_1)$$

- A **valid** test has a controlled **Type I** error
- A **consistent** test has vanishing **Type II** error asymptotically.

What is Statistical Hypothesis Testing?

An *ideal* testing procedure should yield small

False Positives (Type I error) and False Negatives (Type II error)

$\mathbb{P}(\text{Reject } \mathscr{H}_0 \mid \mathscr{H}_0) \leq \alpha$

$\mathbb{P}(\text{Fail to reject } \mathscr{H}_0 \mid \mathscr{H}_1)$

- A **valid** test has a controlled **Type I** error
- A **consistent** test has vanishing **Type II** error asymptotically.

What is Two-Sample Testing?

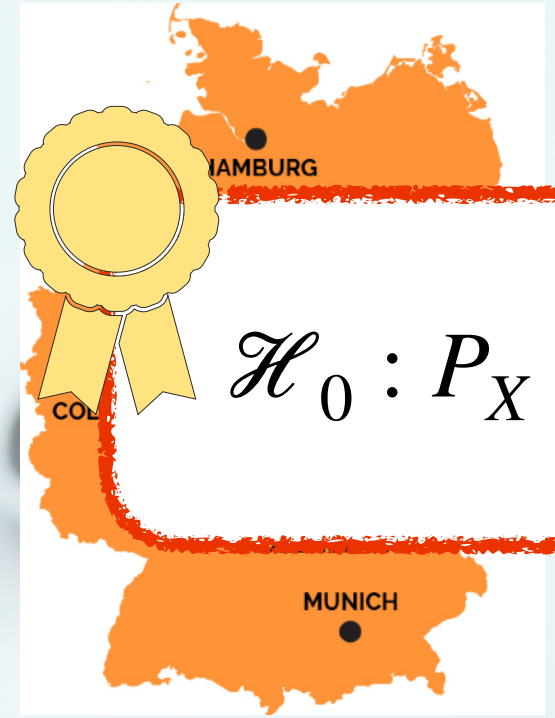
What is Two-Sample Testing?

$$X_1, \ldots, X_n \overset{iid}{\sim} P_X$$

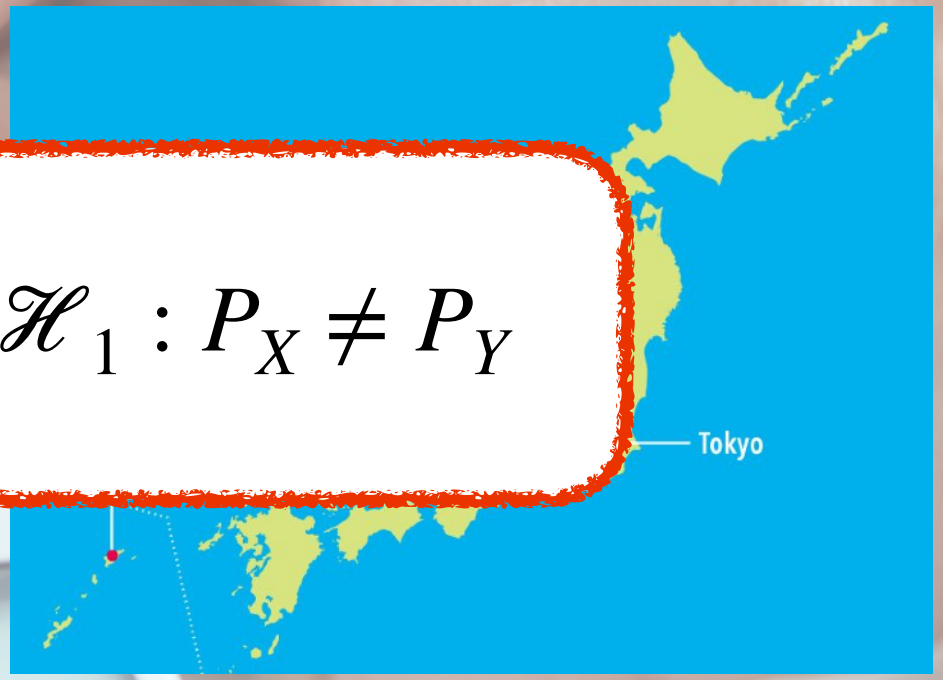
What is Two-Sample Testing?

$$X_1, \ldots, X_n \overset{iid}{\sim} P_X \qquad Y_1, \ldots, Y_m \overset{iid}{\sim} P_Y$$

What is Two-Sample Testing?

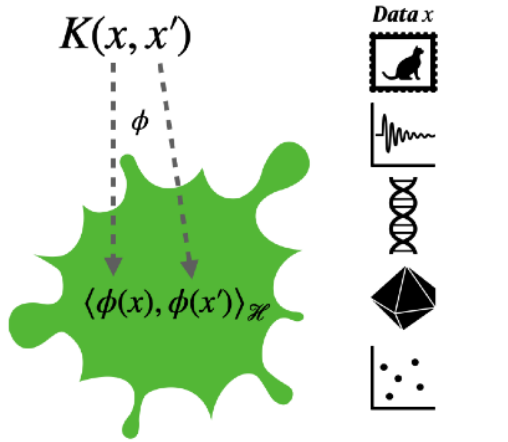


$$\mathcal{H}_0 : P_X = P_Y \quad \text{v.s.} \quad \mathcal{H}_1 : P_X \neq P_Y$$



$$X_1, \ldots, X_n \overset{iid}{\sim} P_X \qquad Y_1, \ldots, Y_m \overset{iid}{\sim} P_Y$$

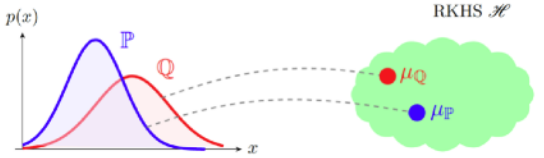
Kernel Two-Sample Testing (Gretton et al. 2012)



Reproducing kernel Hilbert space (RKHS)

Kernel Mean Embedding of Distributions

$$\mathbb{P} \mapsto \int K(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \in \mathscr{H}$$



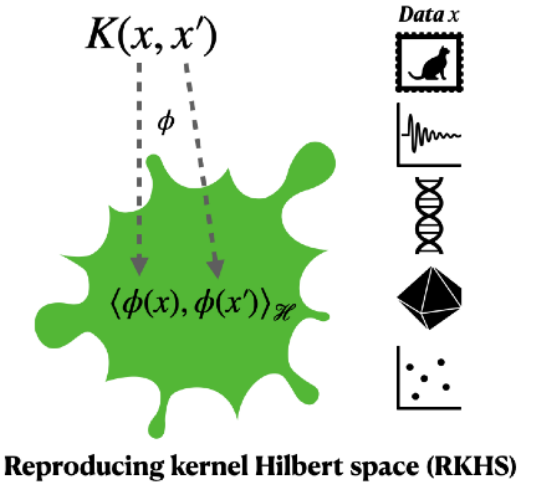
Operations on Distributions

1. Expectation: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathscr{H}}$
2. Distance metric: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}}$
3. Sampling: $\{x_1, x_2, \ldots, x_n\} \sim \mu_{\mathbb{P}}$

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

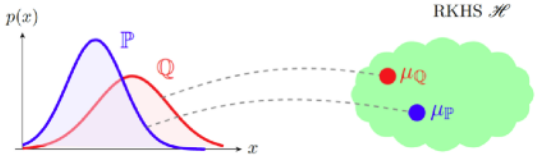
Kernel Two-Sample Testing (Gretton et al. 2012)

- For a space $\mathcal{X}$ where a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is well defined, **kernel mean embedding (KME)** is defined as



Kernel Mean Embedding of Distributions

$$\mathbb{P} \mapsto \int K(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \in \mathscr{H}$$



Operations on Distributions

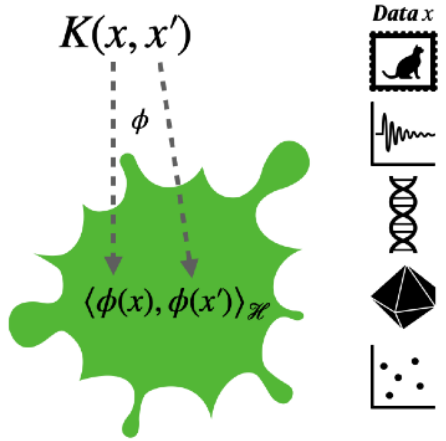
1. Expectation: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathscr{H}}$
2. Distance metric: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}}$
3. Sampling: $\{x_1, x_2, \ldots, x_n\} \sim \mu_{\mathbb{P}}$

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-Sample Testing (Gretton et al. 2012)

- For a space $\mathcal{X}$ where a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is well defined, **kernel mean embedding (KME)** is defined as

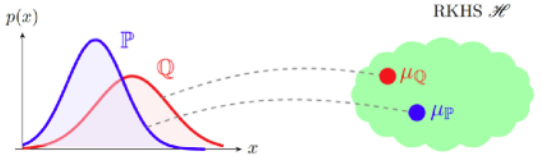
$$\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}, \qquad \mu_{\mathbb{P}} = \int_{\mathcal{X}} k(X, \cdot)\, d\mathbb{P}(X)$$



Reproducing kernel Hilbert space (RKHS)

Kernel Mean Embedding of Distributions

$$\mathbb{P} \mapsto \int K(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}$$



Operations on Distributions

1. Expectation: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$
2. Distance metric: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$
3. Sampling: $\{x_1, x_2, \ldots, x_n\} \sim \mu_{\mathbb{P}}$

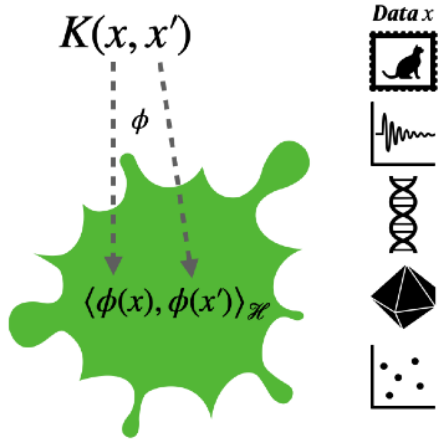
Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-Sample Testing (Gretton et al. 2012)

- For a space $\mathcal{X}$ where a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is well defined, **kernel mean embedding (KME)** is defined as

$$\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}, \qquad \mu_{\mathbb{P}} = \int_{\mathcal{X}} k(X, \cdot)\, d\mathbb{P}(X)$$

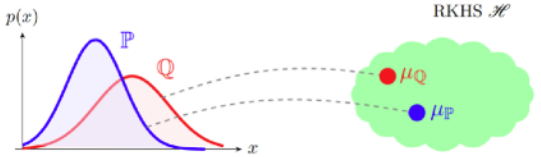
- The **maximum mean discrepancy (MMD)** between probability measures $\mathbb{P}, \mathbb{Q}$ defined on $\mathcal{X}$ is



Reproducing kernel Hilbert space (RKHS)

Kernel Mean Embedding of Distributions

$$\mathbb{P} \mapsto \int K(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}$$



Operations on Distributions

1. Expectation: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$
2. Distance metric: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$
3. Sampling: $\{x_1, x_2, \ldots, x_n\} \sim \mu_{\mathbb{P}}$

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

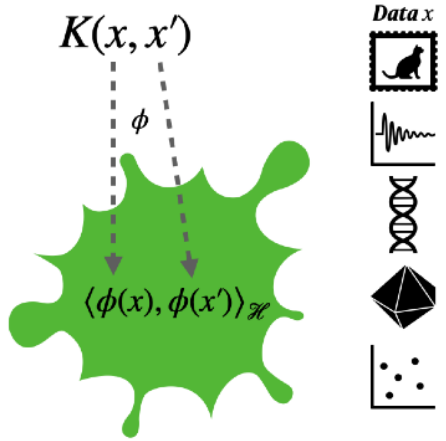
Kernel Two-Sample Testing (Gretton et al. 2012)

- For a space $\mathcal{X}$ where a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is well defined, **kernel mean embedding (KME)** is defined as

$$\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}, \qquad \mu_{\mathbb{P}} = \int_{\mathcal{X}} k(X, \cdot)\, d\mathbb{P}(X)$$

- The **maximum mean discrepancy (MMD)** between probability measures $\mathbb{P}, \mathbb{Q}$ defined on $\mathcal{X}$ is

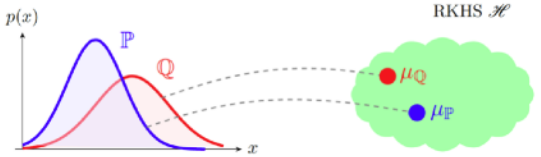
$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k}$$



Reproducing kernel Hilbert space (RKHS)

Kernel Mean Embedding of Distributions

$$\mathbb{P} \mapsto \int K(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}$$



Operations on Distributions

1. Expectation: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$
2. Distance metric: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$
3. Sampling: $\{x_1, x_2, \ldots, x_n\} \sim \mu_{\mathbb{P}}$

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-Sample Testing (Gretton et al. 2012)

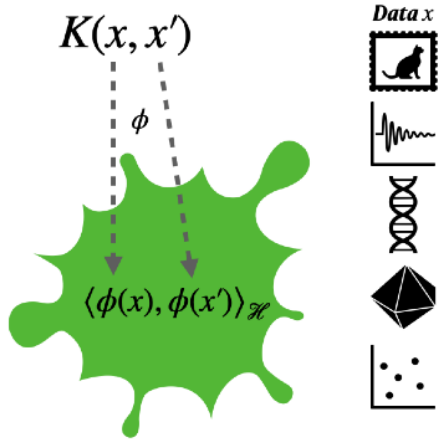
- For a space $\mathcal{X}$ where a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is well defined, **kernel mean embedding (KME)** is defined as

$$\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}, \qquad \mu_{\mathbb{P}} = \int_{\mathcal{X}} k(X, \cdot)\, d\mathbb{P}(X)$$

- The **maximum mean discrepancy (MMD)** between probability measures $\mathbb{P}, \mathbb{Q}$ defined on $\mathcal{X}$ is

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k}$$

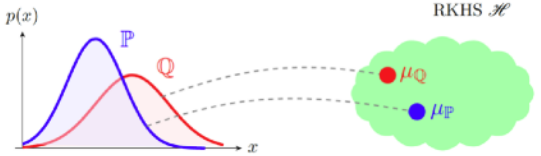
- The **empirical MMD** can be estimated through the *U*-statistic



Reproducing kernel Hilbert space (RKHS)

Kernel Mean Embedding of Distributions

$$\mathbb{P} \mapsto \int K(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}$$



Operations on Distributions

1. Expectation: $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$
2. Distance metric: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$
3. Sampling: $\{x_1, x_2, \ldots, x_n\} \sim \mu_{\mathbb{P}}$

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-Sample Testing (Gretton et al. 2012)

- For a space $\mathcal{X}$ where a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is well defined, **kernel mean embedding (KME)** is defined as

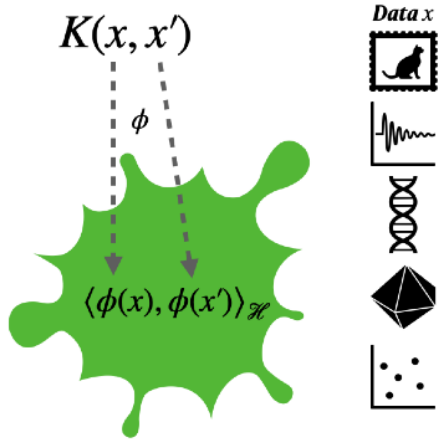
$$\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}, \qquad \mu_{\mathbb{P}} = \int_{\mathcal{X}} k(X, \cdot)\, d\mathbb{P}(X)$$

- The **maximum mean discrepancy (MMD)** between probability measures $\mathbb{P}, \mathbb{Q}$ defined on $\mathcal{X}$ is

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k}$$

- The **empirical MMD** can be estimated through the *U*-statistic

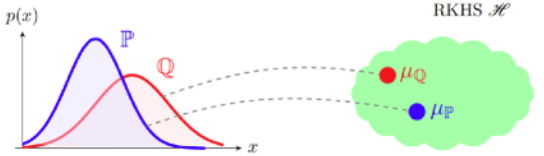
$$\widehat{\mathrm{MMD}^2}(\mathbb{P}, \mathbb{Q}) = \sum_{j=1}^{m} \sum_{i=1, i\neq j}^{n} k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$$



Reproducing kernel Hilbert space (RKHS)

Kernel Mean Embedding of Distributions

$$\mathbb{P} \mapsto \int K(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}$$



Operations on Distributions

1. Expectation: $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$
2. Distance metric: $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$
3. Sampling: $\{x_1, x_2, \ldots, x_n\} \sim \mu_{\mathbb{P}}$

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-sample Testing (Gretton et al. 2012)

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-sample Testing (Gretton et al. 2012)

- Under the null $H_0 : \mathbb{P} = \mathbb{Q}$, we have $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$, but $\widehat{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) \neq 0$.

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-sample Testing (Gretton et al. 2012)

- Under the null $H_0 : \mathbb{P} = \mathbb{Q}$, we have $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$, but $\widehat{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) \neq 0$.
- The limiting null distribution of $\widehat{\mathrm{MMD}}^2(\mathbb{P}, \mathbb{Q})$ can quantify **the surprise** given a specific test statistic value:

Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-sample Testing (Gretton et al. 2012)

- Under the null $H_0 : \mathbb{P} = \mathbb{Q}$, we have $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$, but $\widehat{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) \neq 0$.
- The limiting null distribution of $\widehat{\mathrm{MMD}}^2(\mathbb{P}, \mathbb{Q})$ can quantify **the surprise** given a specific test statistic value:

$$\lim_{n \to \infty} \Pr\left(\widehat{\mathrm{MMD}}^2(\mathbb{P}, \mathbb{Q}) \geq t \mid H_0\right)$$

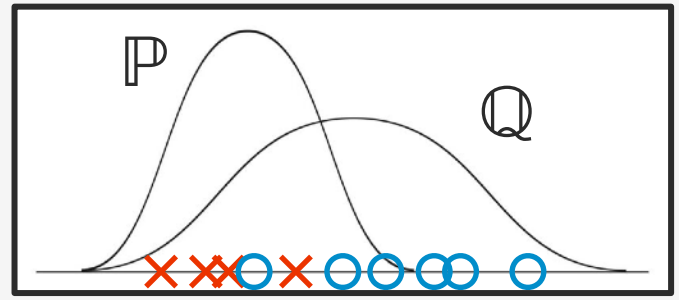
Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

Kernel Two-sample Testing (Gretton et al. 2012)

- Under the null $H_0 : \mathbb{P} = \mathbb{Q}$, we have $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$, but $\widehat{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) \neq 0$.
- The limiting null distribution of $\widehat{\mathrm{MMD}}^2(\mathbb{P}, \mathbb{Q})$ can quantify **the surprise** given a specific test statistic value:

$$\lim_{n \to \infty} \Pr\left(\widehat{\mathrm{MMD}}^2(\mathbb{P}, \mathbb{Q}) \geq t \mid H_0\right)$$

- For $t = 1, \ldots, B$
  - Reshuffle samples between $\mathbb{P}$ and $\mathbb{Q}$
  - Compute the test statistic
- Plot a histogram and use the $1 - \alpha$ level quantile as the critical value.
- Reject the test if the original test statistic exceeds the critical value.



Gretton, A. Borgwardt, K. M. Rasch M. J. Schölkopf, B. Smola, A (2012). A Kernel Two-sample Test. JMLR

What if We Face Dataset Uncertainties?

What if We Face Dataset Uncertainties?

$$X_{1:n}^{(1)} \overset{iid}{\sim} P_X^{(1)}$$

$$X_{1:n}^{(2)} \overset{iid}{\sim} P_X^{(2)}$$

$$X_{1:n}^{(3)} \overset{iid}{\sim} P_X^{(3)}$$

$$X_{1:n}^{(4)} \overset{iid}{\sim} P_X^{(4)}$$

What if We Face Dataset Uncertainties?

$$X^{(1)}_{1:n} \overset{iid}{\sim} P^{(1)}_X$$

$$X^{(2)}_{1:n} \overset{iid}{\sim} P^{(2)}_X$$

$$X^{(3)}_{1:n} \overset{iid}{\sim} P^{(3)}_X$$

$$X^{(4)}_{1:n} \overset{iid}{\sim} P^{(4)}_X$$

$$Y^{(1)}_{1:m} \overset{iid}{\sim} P^{(1)}_Y$$

$$Y^{(2)}_{1:m} \overset{iid}{\sim} P^{(2)}_Y$$

$$Y^{(3)}_{1:m} \overset{iid}{\sim} P^{(3)}_Y$$

What if We Face Dataset Uncertainties?

Proxies for P_X

$X_{1:n}^{(3)} \overset{iid}{\sim} P_X^{(3)}$

$X_{1:n}^{(4)} \overset{iid}{\sim} P_X^{(4)}$

$X_{1:n}^{(1)} \overset{iid}{\sim} P_X^{(1)}$

$X_{1:n}^{(2)} \overset{iid}{\sim} P_X^{(2)}$

Proxies for P_Y

$Y_{1:m}^{(1)} \overset{iid}{\sim} P_Y^{(1)}$

$Y_{1:m}^{(2)} \overset{iid}{\sim} P_Y^{(2)}$

$Y_{1:m}^{(3)} \overset{iid}{\sim} P_Y^{(3)}$

Scientist

What if We Face Dataset Uncertainties?

Proxies for P_X

$X_{1:n}^{(3)} \overset{iid}{\sim} P_X^{(3)}$

$X_{1:n}^{(4)} \overset{iid}{\sim} P_X^{(4)}$

$X_{1:n}^{(1)} \overset{iid}{\sim} P_X^{(1)}$

$X_{1:n}^{(2)} \overset{iid}{\sim} P_X^{(2)}$

Proxies for P_Y

$Y_{1:m}^{(1)} \overset{iid}{\sim} P_Y^{(1)}$

$Y_{1:m}^{(2)} \overset{iid}{\sim} P_Y^{(2)}$

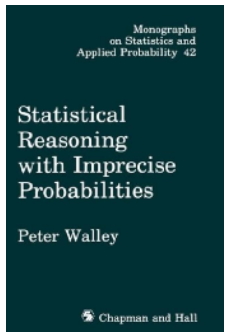
$Y_{1:m}^{(3)} \overset{iid}{\sim} P_Y^{(3)}$

Scientist

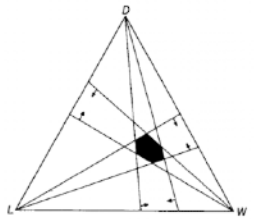
Dataset uncertainty ↦ representation?

Dataset Uncertainty Representation

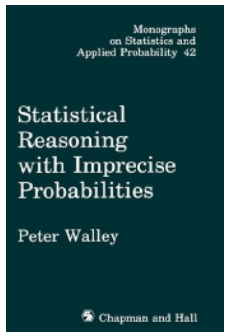
Dataset Uncertainty Representation



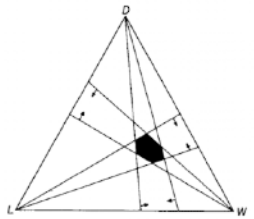
Subjective assessments of lower and upper probabilities leads to closed convex sets of probabilities.



Dataset Uncertainty Representation



Subjective assessments of lower and upper probabilities leads to closed convex sets of probabilities.



Under rationality axioms akin to Savage, partial ordering over items induces a closed convex sets of probabilities.

Quasi-Bayesian behaviour: a more realistic approach to decision making?

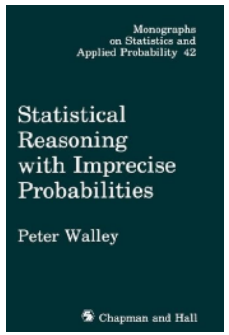
Francisco Javier Girón; Sixto Ríos

Trabajos de Estadística e Investigación Operativa (1980)

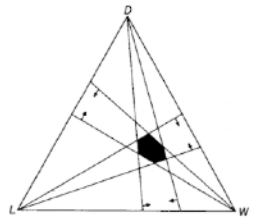
Volume: 31, Issue: 1, page 17-38

ISSN: 0041-0241

Dataset Uncertainty Representation



Subjective assessments of lower and upper probabilities leads to closed convex sets of probabilities.



Under rationality axioms akin to Savage, partial ordering over items induces a closed convex sets of probabilities.

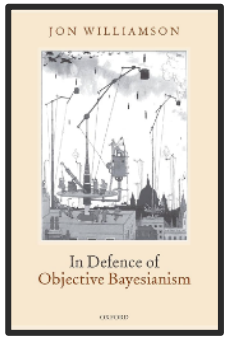
Quasi-Bayesian behaviour: a more realistic approach to decision making?

Francisco Javier Girón; Sixto Ríos

Trabajos de Estadística e Investigación Operativa (1980)

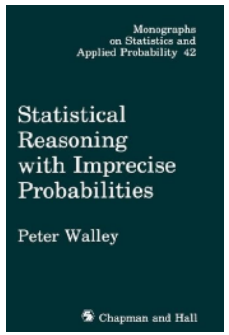
Volume: 31, Issue: 1, page 17-38

ISSN: 0041-0241

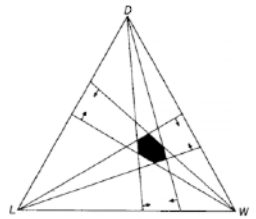


Chance Calibration If, according to E, the current chance function P^* lies within some set $\mathbb{P}^*$ of probability functions, $P^* \in \mathbb{P}^*$, then one's belief function P_E should lie within the convex hull $\langle \mathbb{P}^* \rangle$ of that set, $P_E \in \langle \mathbb{P}^* \rangle$.

Dataset Uncertainty Representation



Subjective assessments of lower and upper probabilities leads to closed convex sets of probabilities.



Under rational

items induces

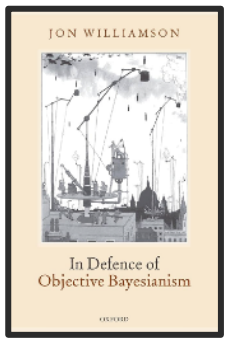
$C \subseteq \mathscr{P}$ is called Credal Set

behaviour: a more realistic
cision making?

Sixto Ríos

e Investigación Operativa (1980)

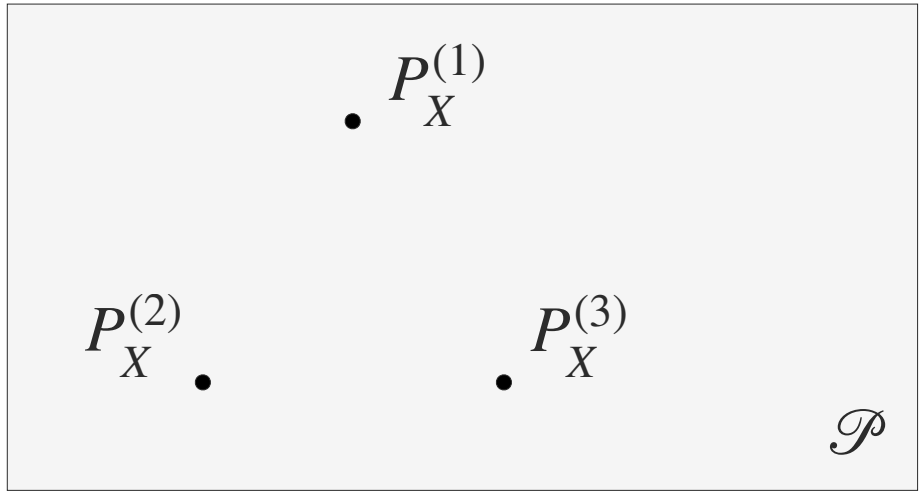
38



Chance Calibration If, according to E, the current chance function P^* lies within some set $\mathbb{P}^*$ of probability functions, $P^* \in \mathbb{P}^*$, then one's belief function P_E should lie within the convex hull $\langle \mathbb{P}^* \rangle$ of that set, $P_E \in \langle \mathbb{P}^* \rangle$.

From Physical to Belief Probabilities with Credal Sets

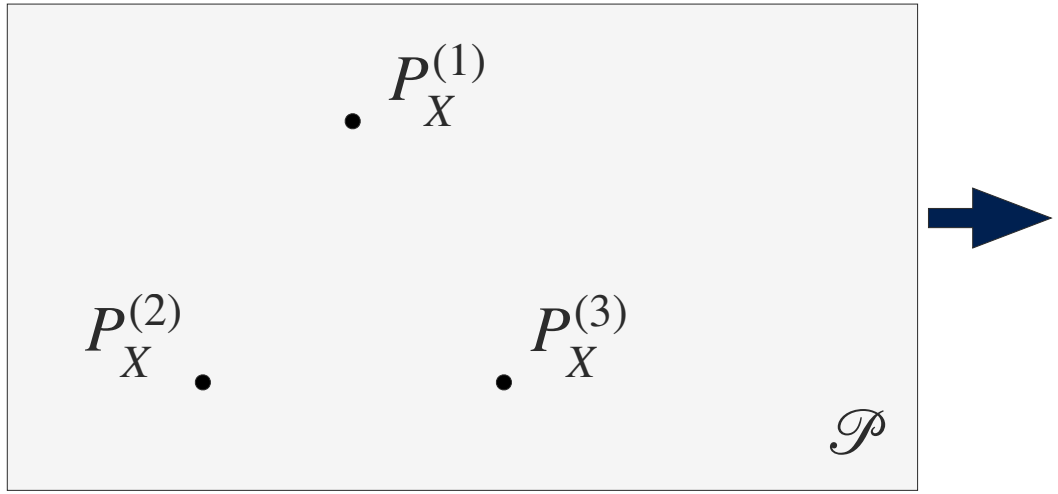
Evidences/Physical/Aleatoric Probabilities



$$\bar{C}_X = \{P_X^{(1)}, \ldots, P_X^{(m)}\}$$

From Physical to Belief Probabilities with Credal Sets

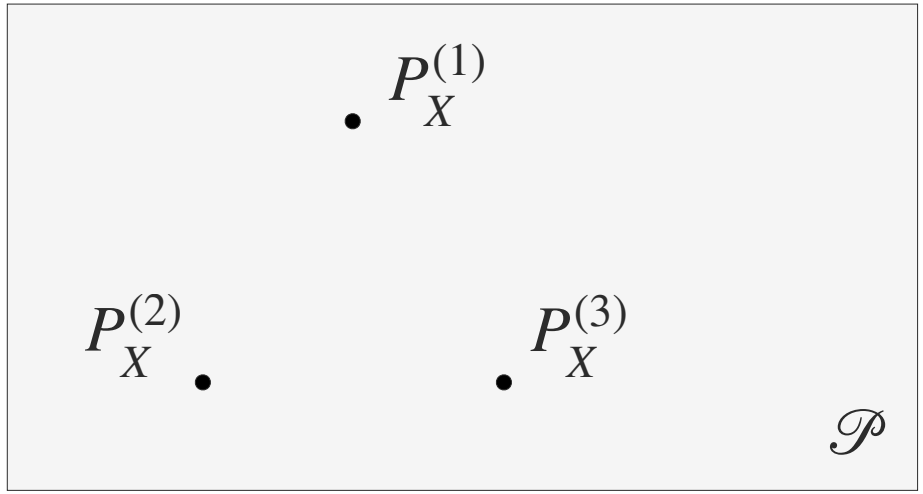
Evidences/Physical/Aleatoric Probabilities



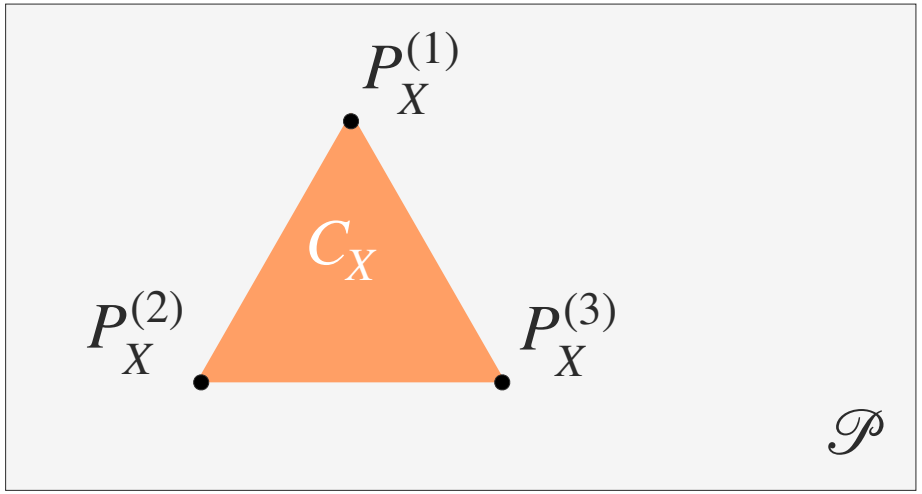
$$\bar{C}_X = \{P_X^{(1)}, \ldots, P_X^{(m)}\}$$

From Physical to Belief Probabilities with Credal Sets

Evidences/Physical/Aleatoric Probabilities



Subjective/Belief/Epistemic Probabilities

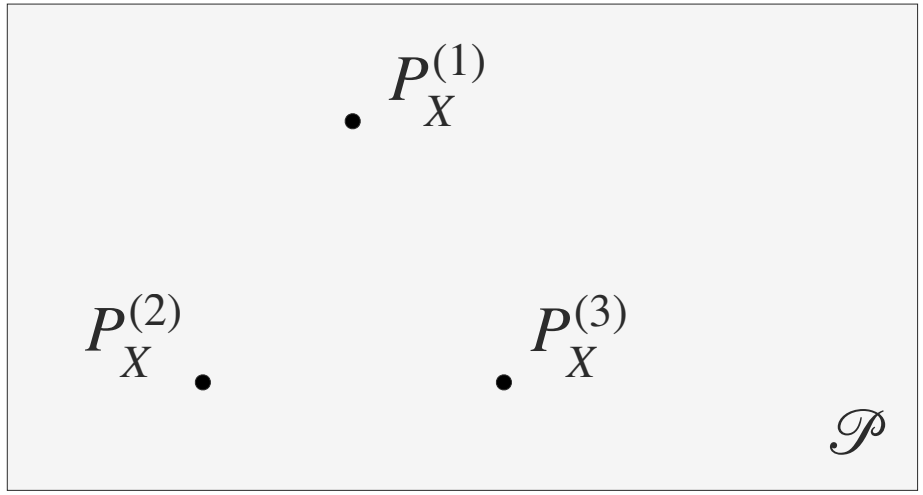


$$\bar{C}_X = \{P_X^{(1)}, \ldots, P_X^{(m)}\}$$

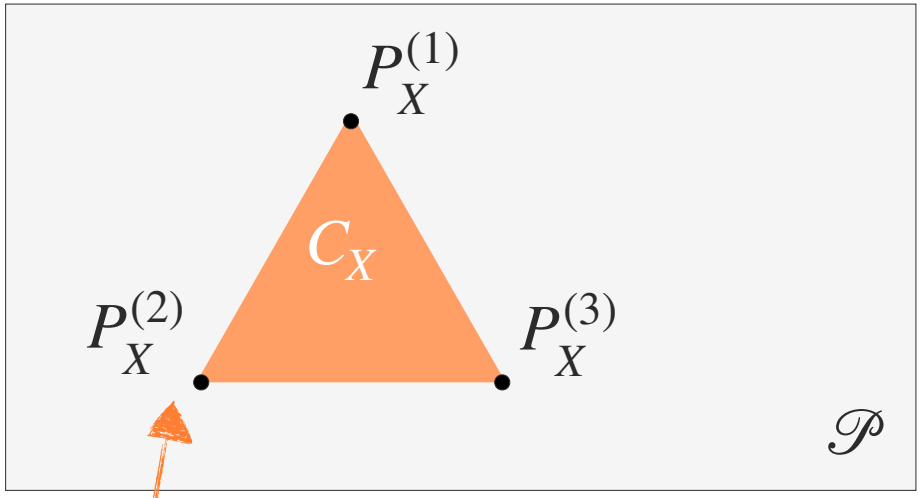
$$C_X = \mathrm{CH}\left(\{P_X^{(1)}, \ldots, P_X^{(m)}\}\right)$$
$$= \{\lambda^\top \mathbf{P}_X \mid \lambda \in \Delta_{m-1}\}$$

From Physical to Belief Probabilities with Credal Sets

Evidences/Physical/Aleatoric Probabilities



Subjective/Belief/Epistemic Probabilities



$$\bar{C}_X = \{P_X^{(1)}, \ldots, P_X^{(m)}\}$$

$$C_X = \mathrm{CH}\left(\{P_X^{(1)}, \ldots, P_X^{(m)}\}\right) = \{\lambda^\top \mathbf{P}_X \mid \lambda \in \Delta_{m-1}\}$$

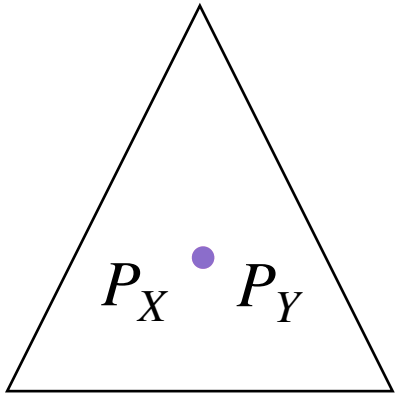
Scientist

Can we test with our rational beliefs instead?

Credal Null Hypothesis for Comparing Epistemic Uncertainties

Credal Null Hypothesis for Comparing Epistemic Uncertainties

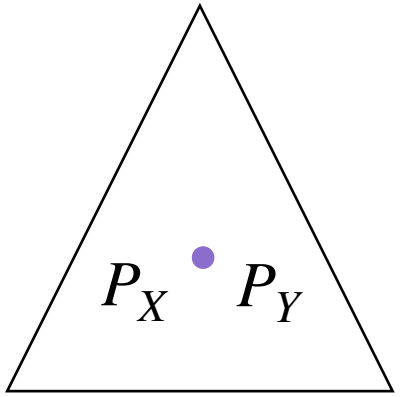
Null Hypothesis for Two-sample Test



Equality H_0

Credal Null Hypothesis for Comparing Epistemic Uncertainties

Null Hypothesis for Two-sample Test

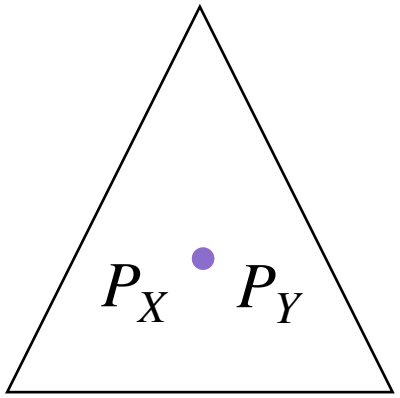


Equality H_0

Null Hypotheses Available for Credal Two-sample Tests

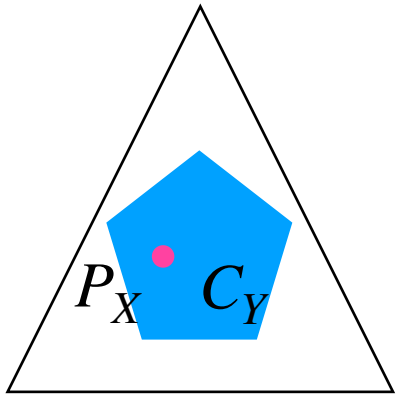
Credal Null Hypothesis for Comparing Epistemic Uncertainties

Null Hypothesis for Two-sample Test



Equality H_0

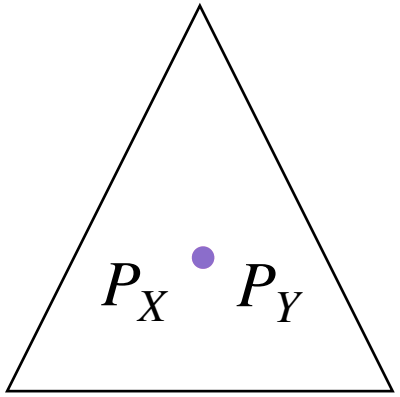
Null Hypotheses Available for Credal Two-sample Tests



Specification $H_{0,\in}$

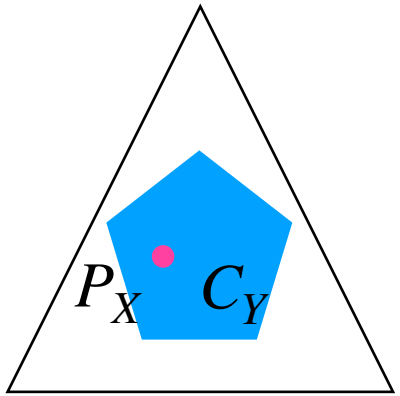
Credal Null Hypothesis for Comparing Epistemic Uncertainties

Null Hypothesis for Two-sample Test

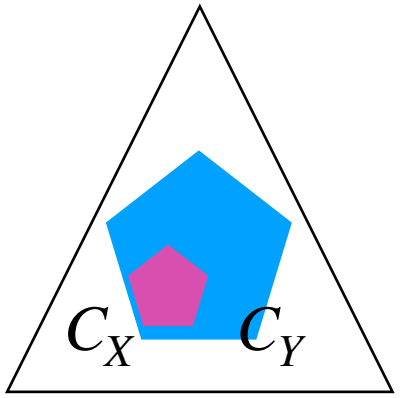


Equality H_0

Null Hypotheses Available for Credal Two-sample Tests



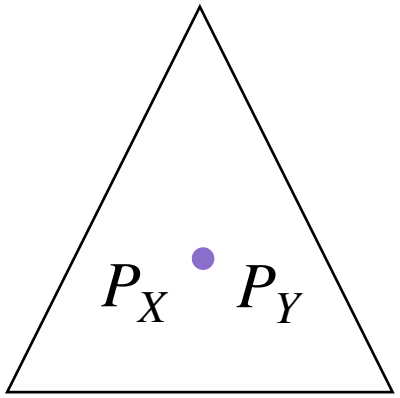
Specification $H_{0,\in}$



Inclusion $H_{0,\subseteq}$

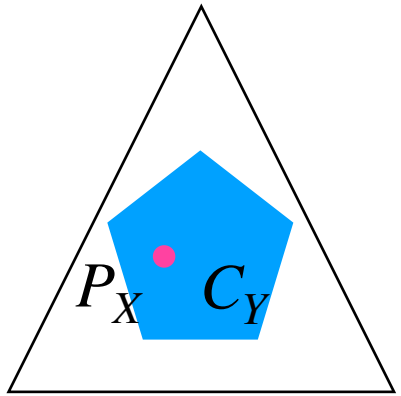
Credal Null Hypothesis for Comparing Epistemic Uncertainties

Null Hypothesis for Two-sample Test

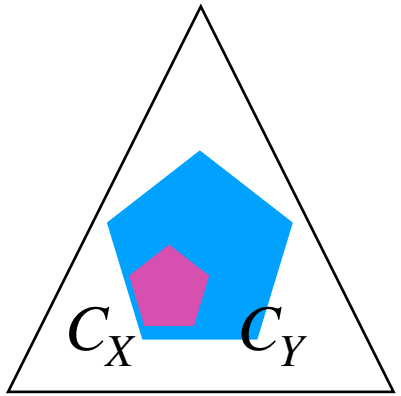


Equality H_0

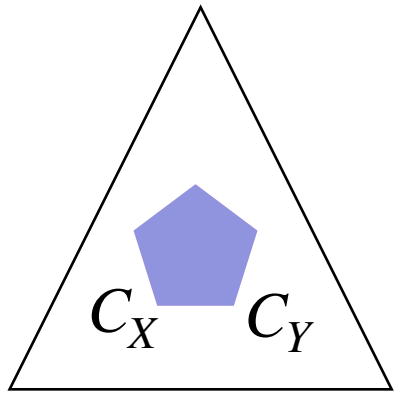
Null Hypotheses Available for Credal Two-sample Tests



Specification $H_{0,\in}$



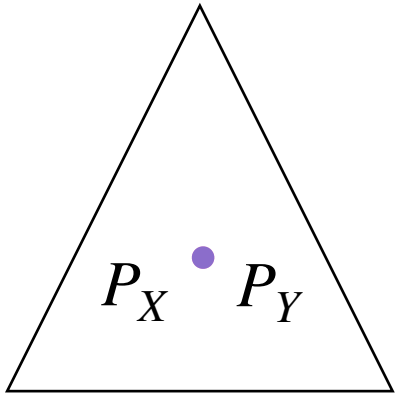
Inclusion $H_{0,\subseteq}$



Equality $H_{0,=}$

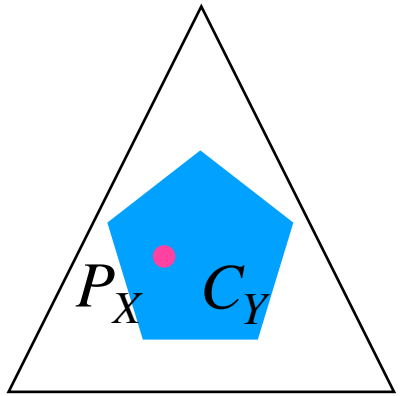
Credal Null Hypothesis for Comparing Epistemic Uncertainties

Null Hypothesis for Two-sample Test

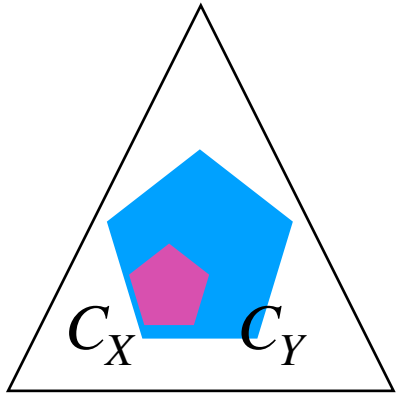


Equality H_0

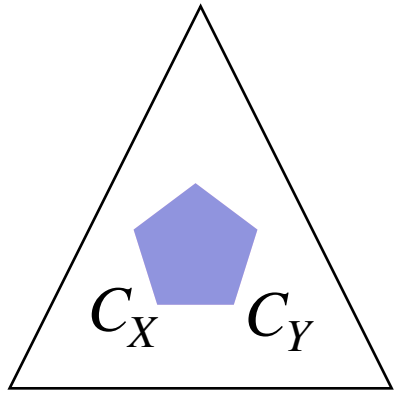
Null Hypotheses Available for Credal Two-sample Tests



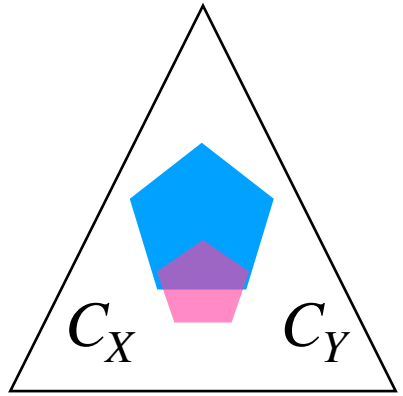
Specification $H_{0,\in}$



Inclusion $H_{0,\subseteq}$



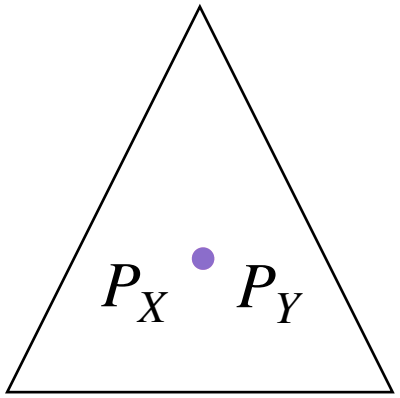
Equality $H_{0,=}$



Plausibility $H_{0,\cap}$

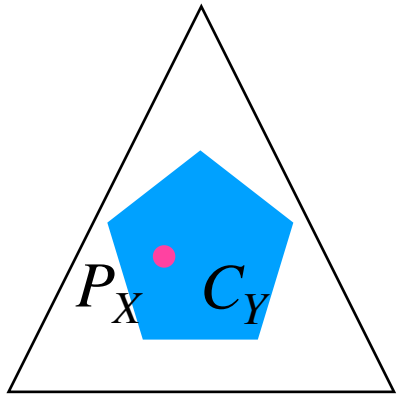
Credal Null Hypothesis for Comparing Epistemic Uncertainties

Null Hypothesis for Two-sample Test

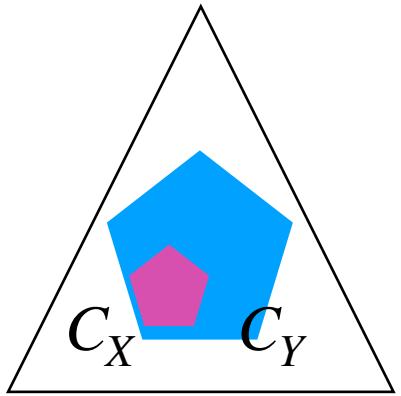


Equality H_0

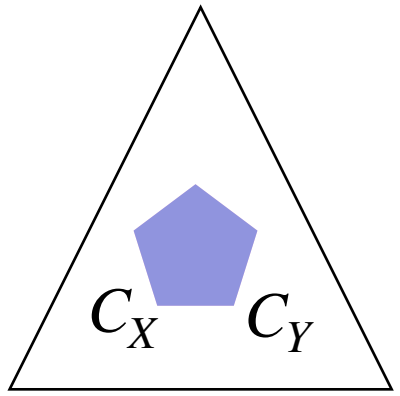
Null Hypotheses Available for Credal Two-sample Tests



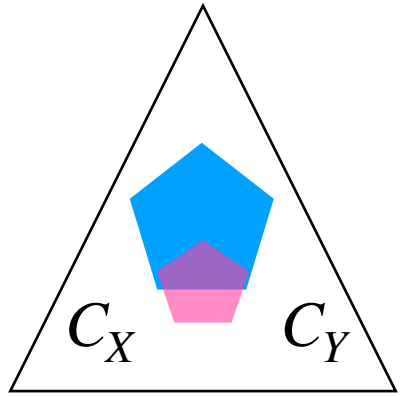
Specification $H_{0,\in}$



Inclusion $H_{0,\subseteq}$



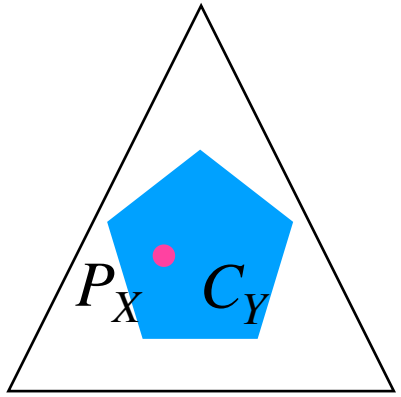
Equality $H_{0,=}$



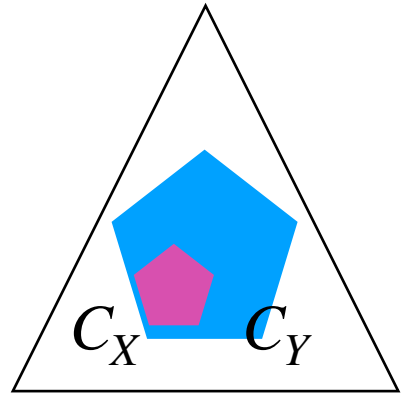
Plausibility $H_{0,\cap}$

Table 1: Different hypotheses to compare credal sets.

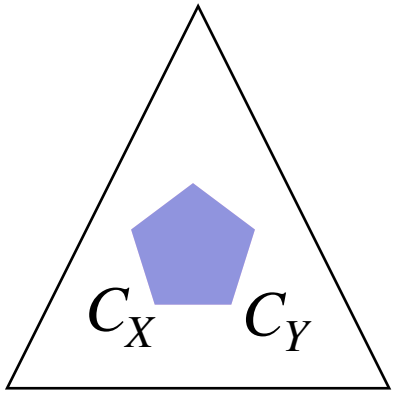
Specification	Inclusion	Equality	Plausibility
$H_{0,\in} : P_X \in \mathcal{C}_Y$	$H_{0,\subseteq} : \mathcal{C}_X \subseteq \mathcal{C}_Y$	$H_{0,=} : \mathcal{C}_X = \mathcal{C}_Y$	$H_{0,\cap} : \mathcal{C}_X \cap \mathcal{C}_Y \neq \emptyset$
$H_{A,\in} : P_X \notin \mathcal{C}_Y$	$H_{A,\subseteq} : \mathcal{C}_X \not\subseteq \mathcal{C}_Y$	$H_{A,=} : \mathcal{C}_X \neq \mathcal{C}_Y$	$H_{A,\cap} : \mathcal{C}_X \cap \mathcal{C}_Y = \emptyset$



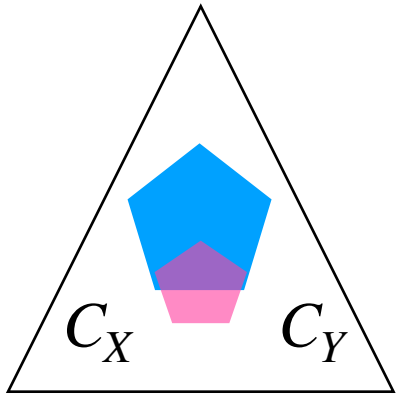
Specification $H_{0,\in} : P_X \in C_Y$



Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$



Equality $H_{0,=} : C_X = C_Y$

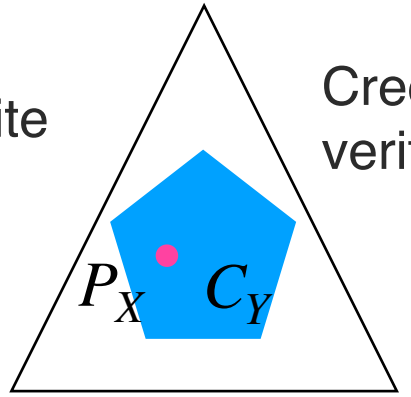


Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$

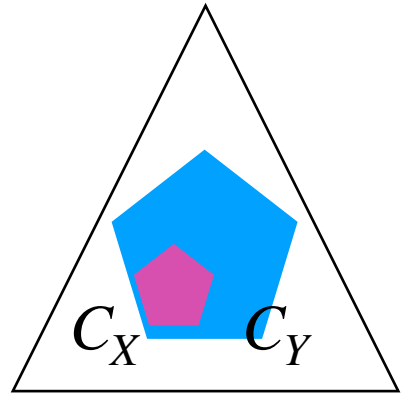
OOD Detection

Testing finite mixtures

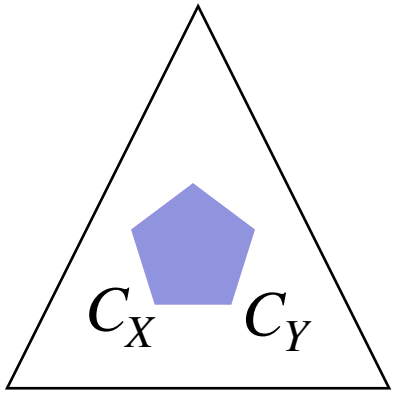
Credal predictor verification



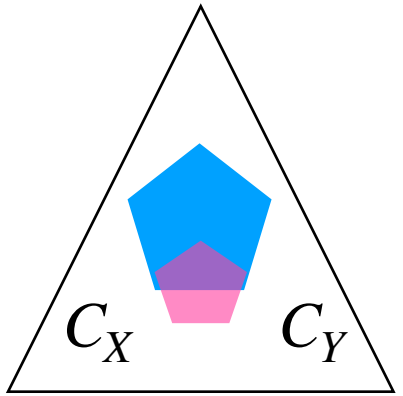
Specification $H_{0,\in} : P_X \in C_Y$



Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$



Equality $H_{0,=} : C_X = C_Y$

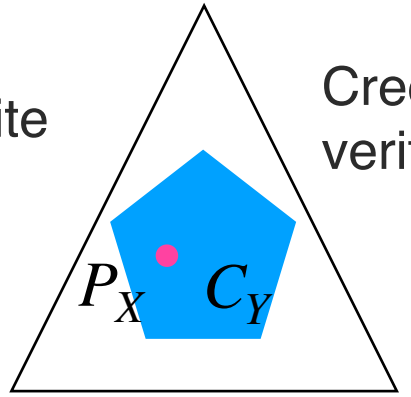


Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$

OOD Detection

Testing finite mixtures

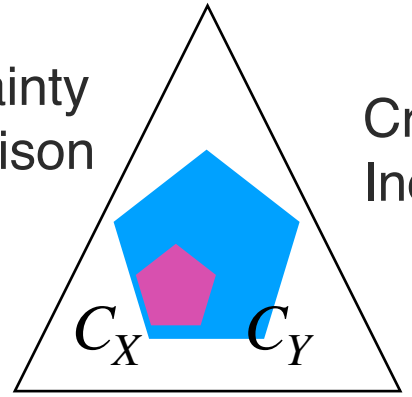
Credal predictor verification



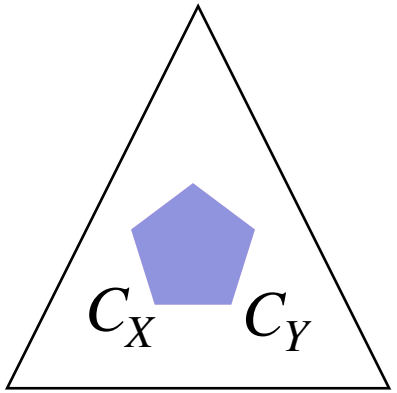
Specification $H_{0,\in} : P_X \in C_Y$

Uncertainty comparison

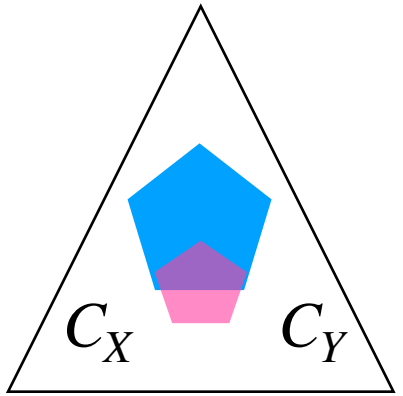
Credal Independence



Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$



Equality $H_{0,=} : C_X = C_Y$

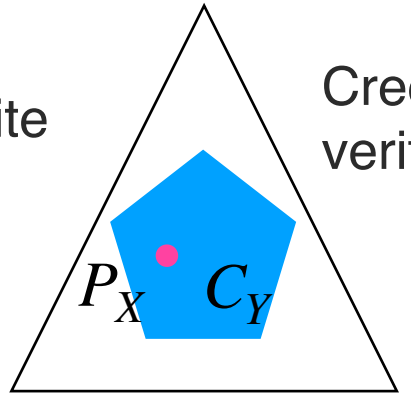


Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$

OOD Detection

Testing finite mixtures

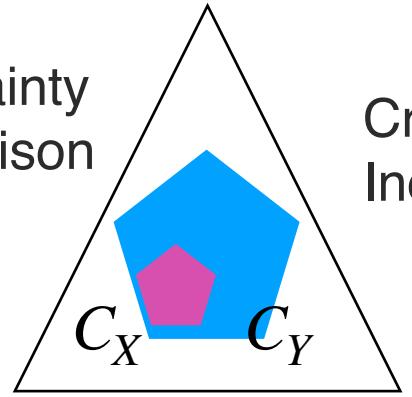
Credal predictor verification



Specification $H_{0,\in} : P_X \in C_Y$

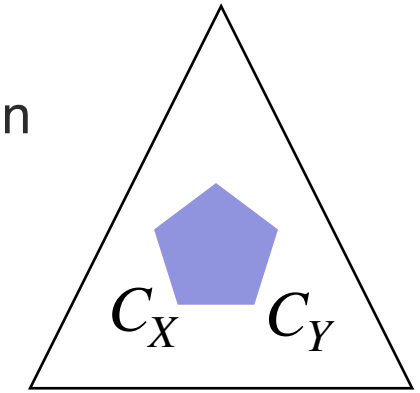
Uncertainty comparison

Credal Independence

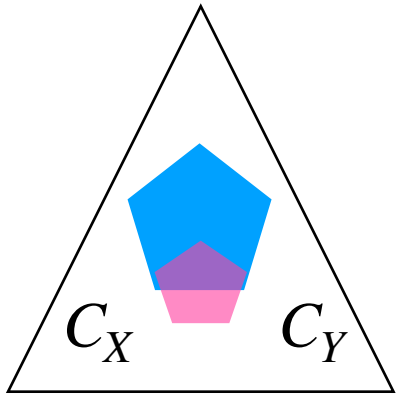


Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$

Strict generalisation of two-sample test



Equality $H_{0,=} : C_X = C_Y$

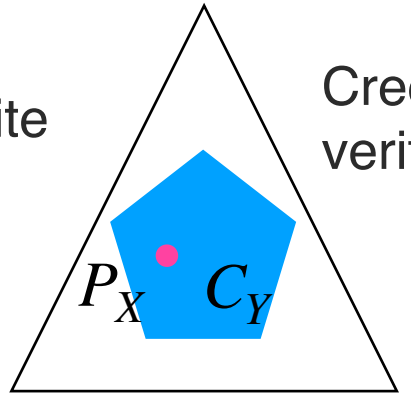


Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$

OOD Detection

Testing finite mixtures

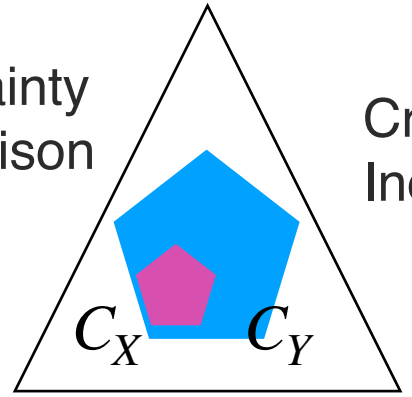
Credal predictor verification



Specification $H_{0,\in} : P_X \in C_Y$

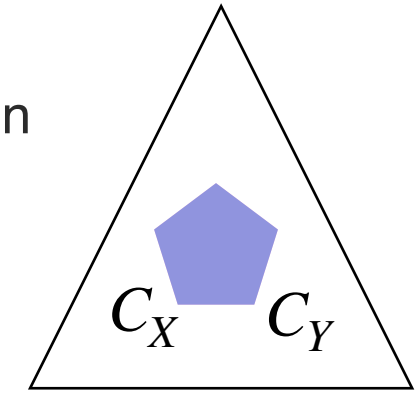
Uncertainty comparison

Credal Independence



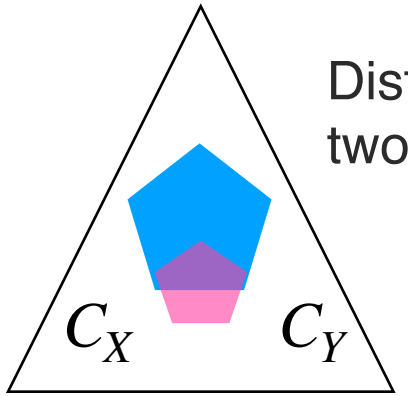
Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$

Strict generalisation of two-sample test

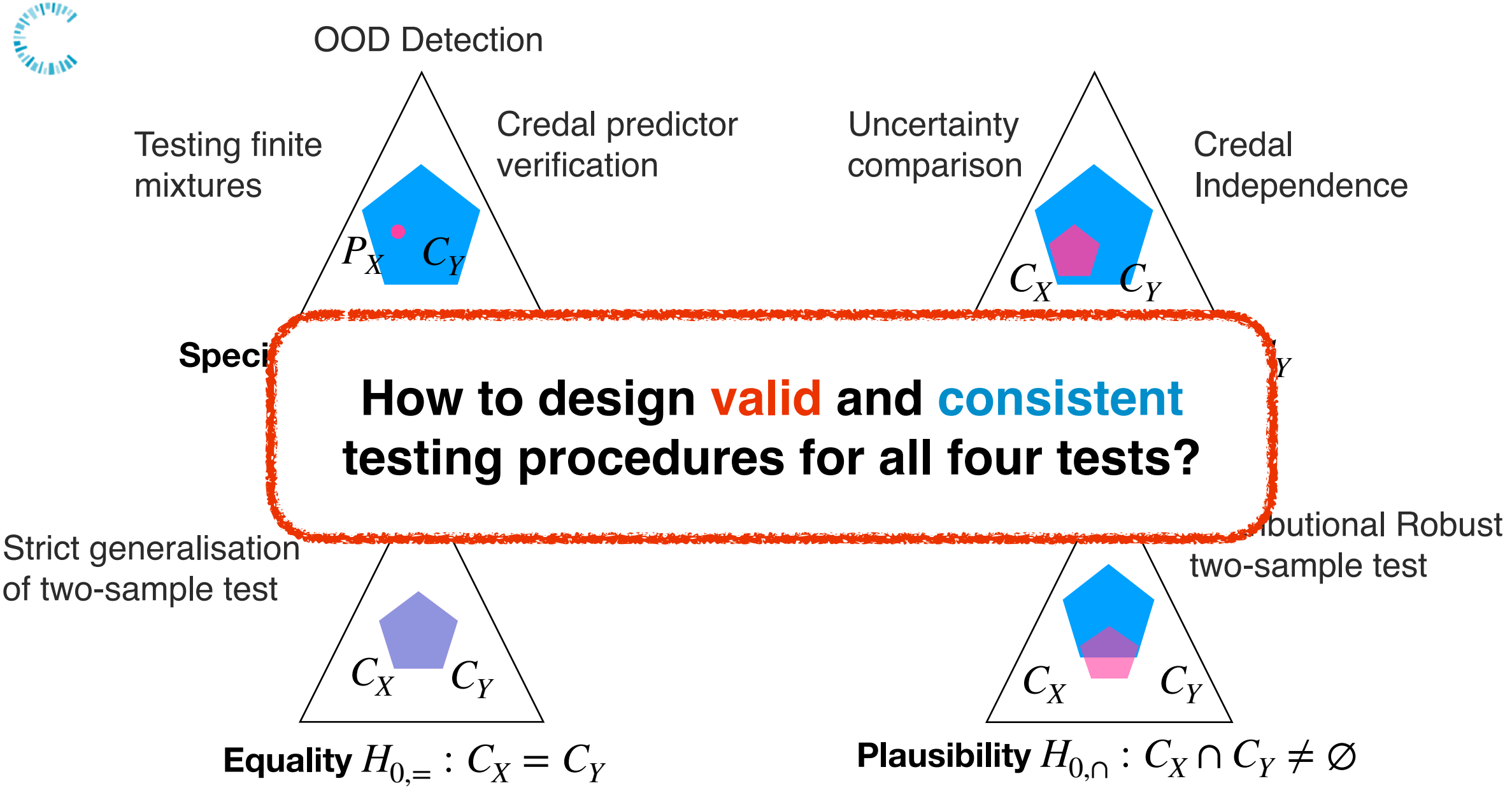


Equality $H_{0,=} : C_X = C_Y$

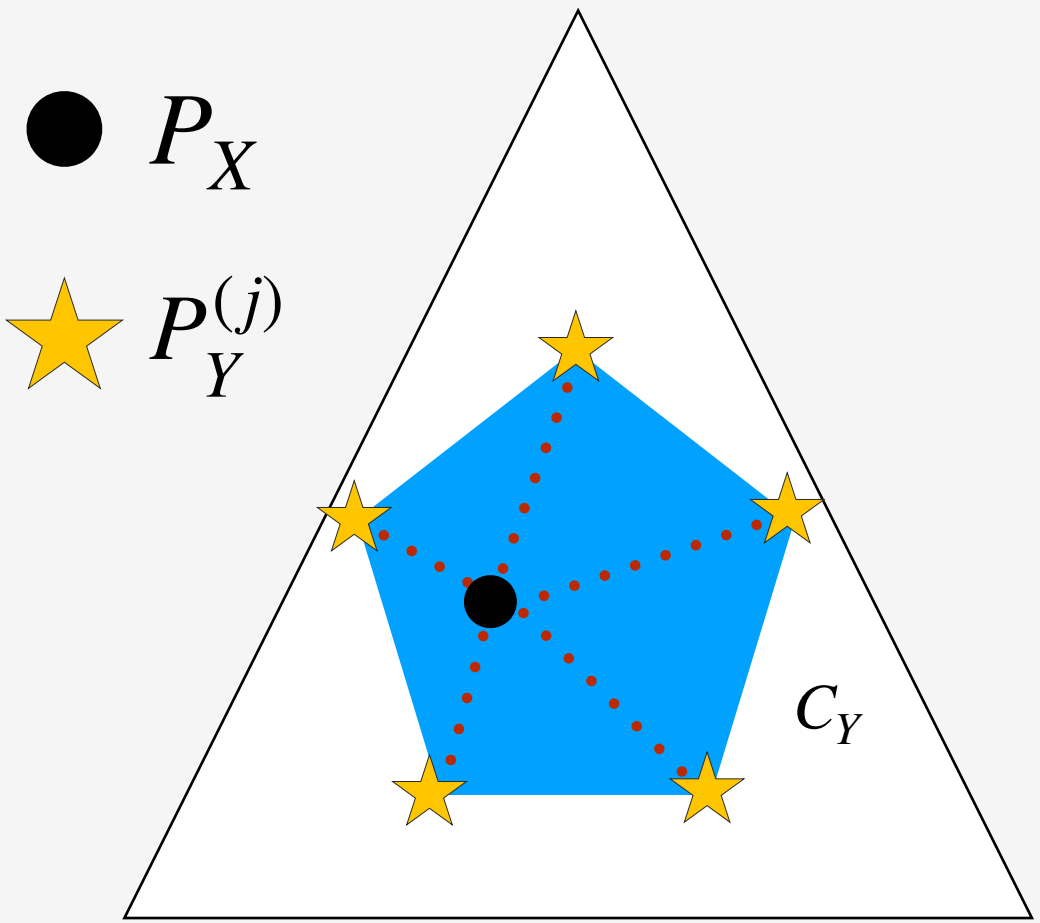
Distributional Robust two-sample test



Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$


OOD Detection
Testing finite mixtures
Credal predictor verification
Uncertainty comparison
Credal Independence
P_X
C_Y
C_X
C_Y
Speci
Y
How to design valid and consistent testing procedures for all four tests?
Strict generalisation of two-sample test
ibutional Robust two-sample test
C_X
C_Y
C_X
C_Y
Equality $H_{0,=} : C_X = C_Y$
Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$

Credal Specification Test

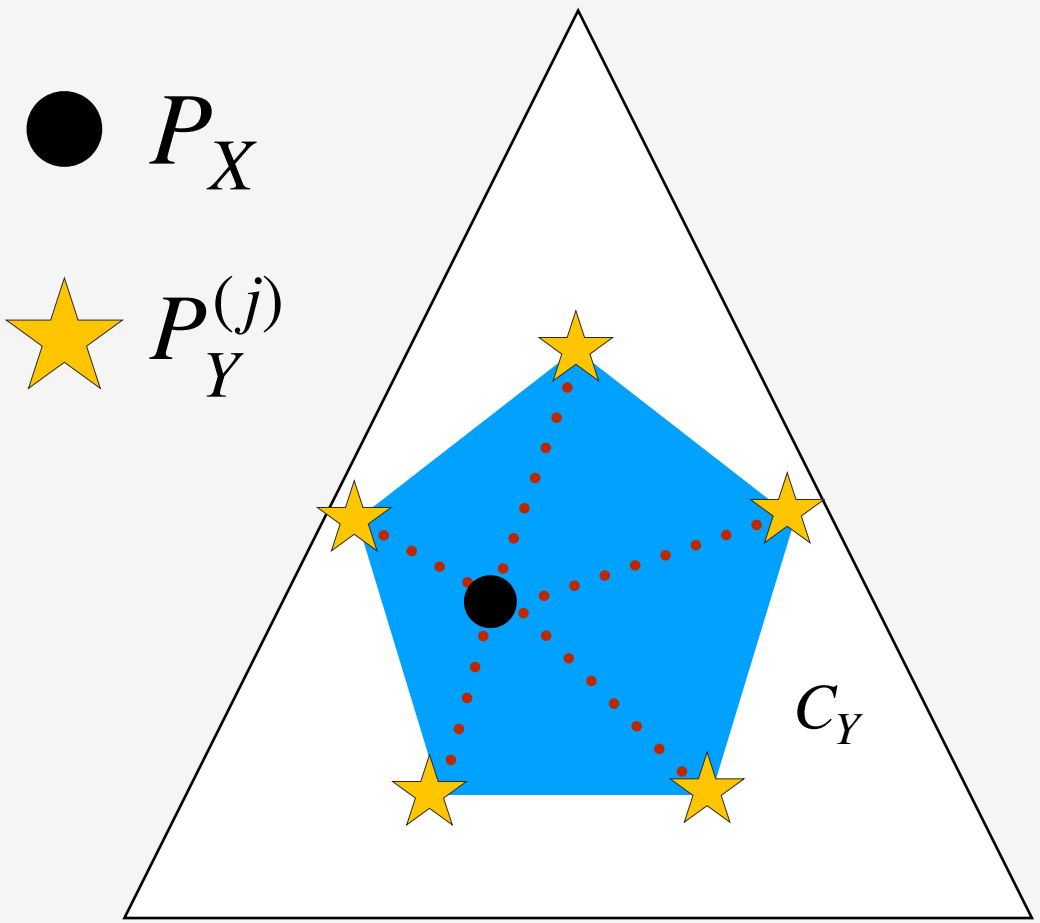


Specification $H_{0,\in} : P_X \in C_Y$

Problem statement

- Observe $X_{1:n} \overset{iid}{\sim} P_X$, $Y_{1:m}^{(j)} \overset{iid}{\sim} P_Y^{(j)}$ for $j = 1, \ldots, r$
- Can we find evidence to reject the null $\mathscr{H}_{0,\in}$?

Credal Specification Test



Specification $H_{0,\in} : P_X \in C_Y$

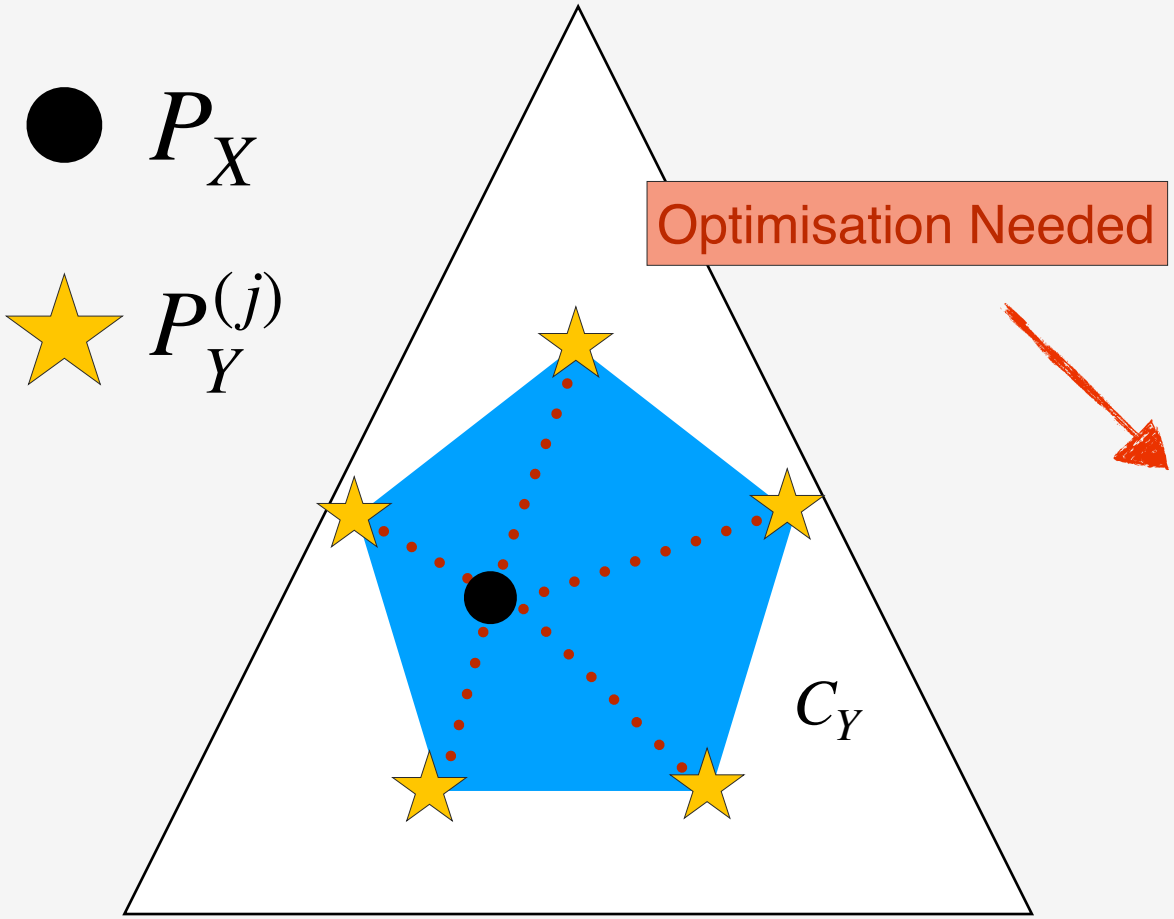
Problem statement

- Observe $X_{1:n} \overset{iid}{\sim} P_X$, $Y_{1:m}^{(j)} \overset{iid}{\sim} P_Y^{(j)}$ for $j = 1, \ldots, r$
- Can we find evidence to reject the null $\mathscr{H}_{0,\in}$?

High-level idea

- $P_X \in C_Y \implies P_X = \lambda^\top \mathbf{P}_Y$ for some $\lambda \in \Delta_{r-1}$
- Can we find this λ based on samples?

Credal Specification Test



Specification $H_{0,\in} : P_X \in C_Y$

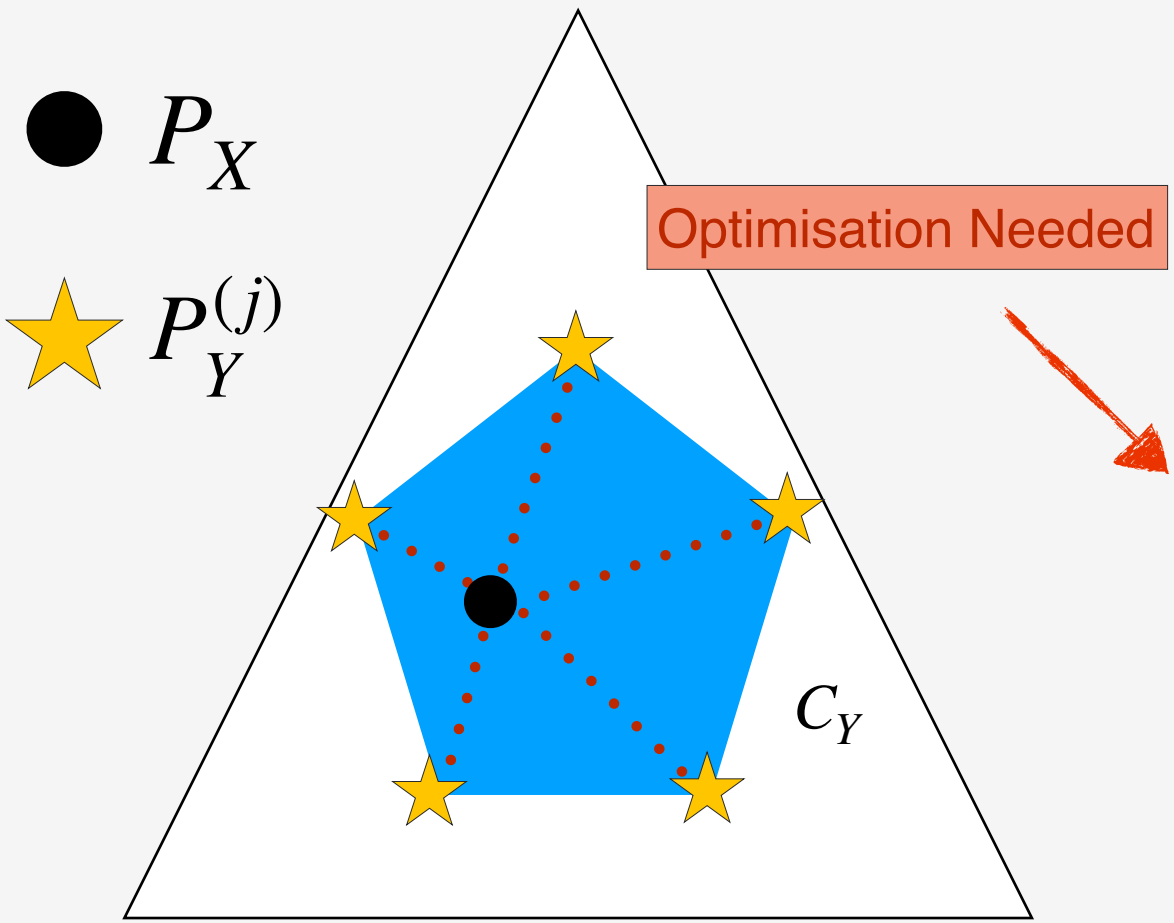
Problem statement

- Observe $X_{1:n} \overset{iid}{\sim} P_X$, $Y_{1:m}^{(j)} \overset{iid}{\sim} P_Y^{(j)}$ for $j = 1, \ldots, r$
- Can we find evidence to reject the null $\mathscr{H}_{0,\in}$?

High-level idea

- $P_X \in C_Y \implies P_X = \lambda^\top \mathbf{P}_Y$ for some $\lambda \in \Delta_{r-1}$
- Can we find this λ based on samples?

Credal Specification Test



Specification $H_{0,\in} : P_X \in C_Y$

Problem statement

- Observe $X_{1:n} \overset{iid}{\sim} P_X$, $Y^{(j)}_{1:m} \overset{iid}{\sim} P_Y^{(j)}$ for $j = 1, \ldots, r$
- Can we find evidence to reject the null $\mathscr{H}_{0,\in}$?

High-level idea

- $P_X \in C_Y \implies P_X = \lambda^\top \mathbf{P}_Y$ for some $\lambda \in \Delta_{r-1}$
- Can we find this λ based on samples?

Natural two-stage procedure

I. Simulate pseudo samples from $\hat{\lambda}^\top \mathbf{P}_Y$

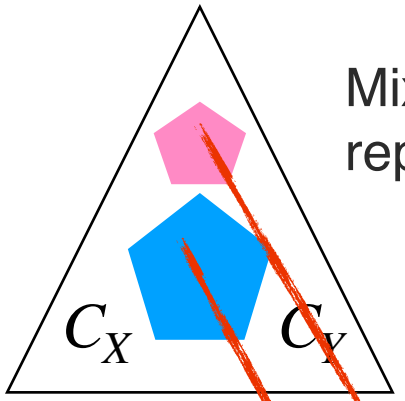
II. Conduct two-sample test with samples from P_X

Kernel Credal Discrepancies (KCD)

$$\mathrm{KCD}(\eta, \lambda) = \mathrm{MMD}(\eta^\top \mathbf{P}_X, \lambda^\top \mathbf{P}_Y) = \|\mu_{\eta^\top \mathbf{P}_X} - \mu_{\lambda^\top \mathbf{P}_Y}\|_{\mathscr{H}_k}$$

Muandet, K. Fukumizu, K. Sriperumbudur, B. Schölkopf, B. (2017), **Kernel Mean Embedding of Distributions: A Review and Beyond**, FnT

Kernel Credal Discrepancies (KCD)

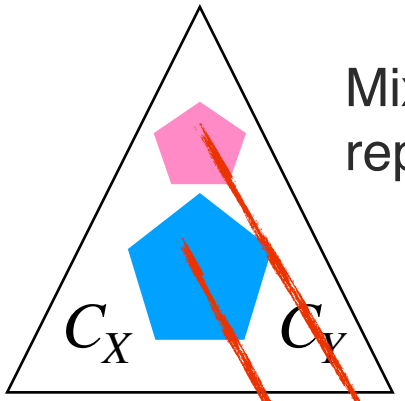


Mixture representation

$$\mathrm{KCD}(\eta, \lambda) = \mathrm{MMD}(\eta^\top \mathbf{P}_X, \lambda^\top \mathbf{P}_Y) = \|\mu_{\eta^\top \mathbf{P}_X} - \mu_{\lambda^\top \mathbf{P}_Y}\|_{\mathscr{H}_k}$$

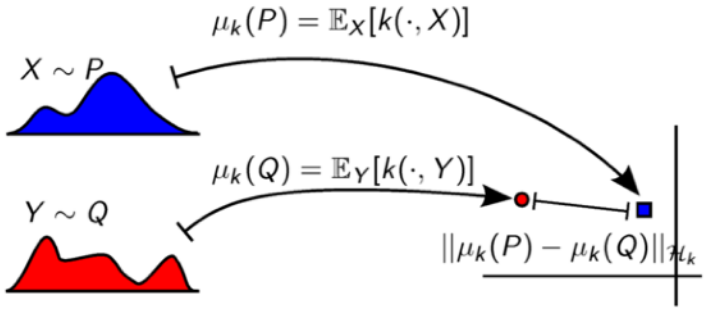
Muandet, K. Fukumizu, K. Sriperumbudur, B. Schölkopf, B. (2017), **Kernel Mean Embedding of Distributions: A Review and Beyond**, FnT

Kernel Credal Discrepancies (KCD)



Mixture representation

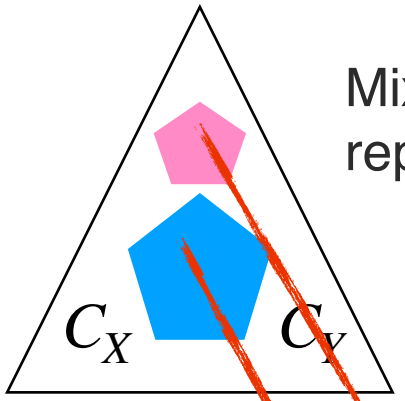
Maximum Mean Discrepancy



$$\mathrm{KCD}(\eta, \lambda) = \mathrm{MMD}(\eta^\top \mathbf{P}_X, \lambda^\top \mathbf{P}_Y) = \|\mu_{\eta^\top \mathbf{P}_X} - \mu_{\lambda^\top \mathbf{P}_Y}\|_{\mathscr{H}_k}$$

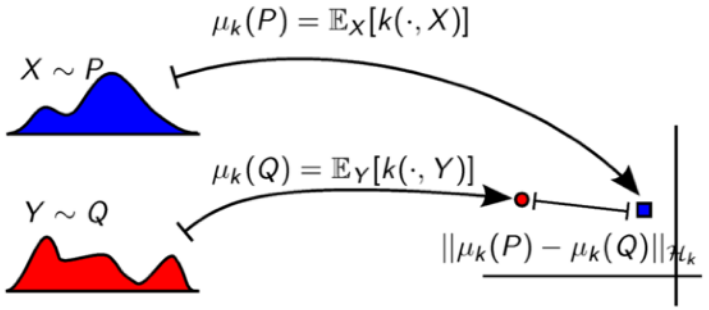
Muandet, K. Fukumizu, K. Sriperumbudur, B. Schölkopf, B. (2017), **Kernel Mean Embedding of Distributions: A Review and Beyond**, FnT

Kernel Credal Discrepancies (KCD)



Mixture representation

Maximum Mean Discrepancy



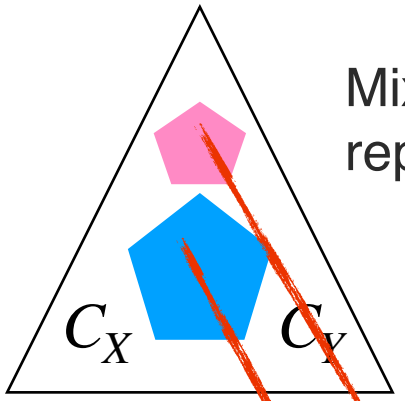
$$\mathrm{KCD}(\eta, \lambda) = \mathrm{MMD}(\eta^\top \mathbf{P}_X, \lambda^\top \mathbf{P}_Y) = \|\mu_{\eta^\top \mathbf{P}_X} - \mu_{\lambda^\top \mathbf{P}_Y}\|_{\mathscr{H}_k}$$

Kernel mean embedding

$$\mu : P_X \mapsto \int k(X, \cdot)\, dP_X(X)$$

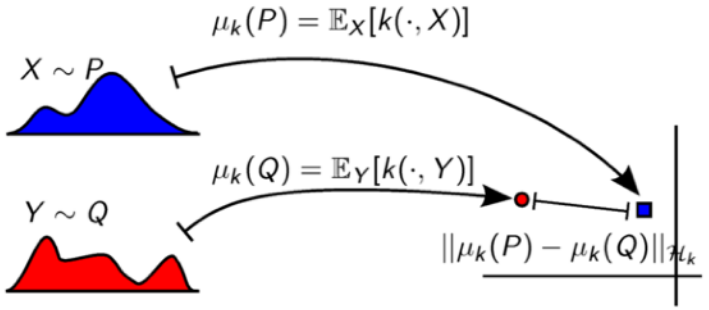
Muandet, K. Fukumizu, K. Sriperumbudur, B. Schölkopf, B. (2017), **Kernel Mean Embedding of Distributions: A Review and Beyond**, FnT

Kernel Credal Discrepancies (KCD)

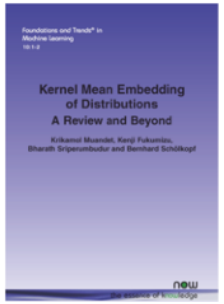


Mixture representation

Maximum Mean Discrepancy



$$\mathrm{KCD}(\eta, \lambda) = \mathrm{MMD}(\eta^\top \mathbf{P}_X, \lambda^\top \mathbf{P}_Y) = \|\mu_{\eta^\top \mathbf{P}_X} - \mu_{\lambda^\top \mathbf{P}_Y}\|_{\mathscr{H}_k}$$



$\sqrt{n}$-consistent estimator
(no parametric assumption needed)

$$\hat{\mu}_{P_X} = \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot)$$

Kernel mean embedding

$$\mu : P_X \mapsto \int k(X, \cdot)\, dP_X(X)$$

Muandet, K. Fukumizu, K. Sriperumbudur, B. Schölkopf, B. (2017), **Kernel Mean Embedding of Distributions: A Review and Beyond**, FnT

Kernel Credal Discrepancies (KCD)

$$\mathrm{KCD}(\eta, \lambda)^2 = \eta^\top \mathbf{M}_{XX} \eta - 2\eta^\top \mathbf{M}_{XY} \lambda + \lambda^\top \mathbf{M}_{YY} \lambda$$

Kernel Credal Discrepancies (KCD)

Proposition 4. *Under Assumption 2, L_{n_e} and $\mathcal{L}_{n_t}$ converges uniformly to L at $O(1/\sqrt{n_e})$ and $O(1/\sqrt{n_t})$.*

A convex/biconvex objective! That is also $\sqrt{n}$-consistent!

$$\mathrm{KCD}(\eta, \lambda)^2 = \eta^\top \mathbf{M}_{XX} \eta - 2\eta^\top \mathbf{M}_{XY} \lambda + \lambda^\top \mathbf{M}_{YY} \lambda$$

Kernel Credal Discrepancies (KCD)

Proposition 4. *Under Assumption 2, L_{n_e} and $\mathcal{L}_{n_t}$ converges uniformly to L at $O(1/\sqrt{n_e})$ and $O(1/\sqrt{n_t})$.*

A convex/biconvex objective! That is also $\sqrt{n}$-consistent!

$$\mathrm{KCD}(\eta, \lambda)^2 = \eta^\top \mathbf{M}_{XX} \eta - 2\eta^\top \mathbf{M}_{XY} \lambda + \lambda^\top \mathbf{M}_{YY} \lambda$$

Measures distribution similarity

$$\mathbf{M}_{XY_{i,j}} = \langle \mu_{P_X^{(i)}}, \mu_{P_Y^{(j)}} \rangle$$

Kernel Credal Discrepancies (KCD)

Proposition 4. *Under Assumption 2, L_{n_e} and $\mathcal{L}_{n_t}$ converges uniformly to L at $O(1/\sqrt{n_e})$ and $O(1/\sqrt{n_t})$.*

A convex/biconvex objective! That is also $\sqrt{n}$-consistent!

$$\mathrm{KCD}(\eta,\lambda)^2 = \eta^\top \mathbf{M}_{XX}\eta - 2\eta^\top \mathbf{M}_{XY}\lambda + \lambda^\top \mathbf{M}_{YY}\lambda$$

A valid objective for credal hypotheses! e.g.

$$P_X \in C_Y \iff \inf_{\lambda\in\Delta_{r-1}} \mathrm{KCD}(1,\lambda) = 0$$

Measures distribution similarity

$$\mathbf{M}_{XY_{i,j}} = \langle \mu_{P_X^{(i)}}, \mu_{P_Y^{(j)}} \rangle$$

Proposition 9. *Under Assumption 1, 2, and under the null $H_{0,\in} : P_X \in C_Y$, η^e, the minimiser of the empirical KCD $L_{n_e}(1,\eta)$, converges to η_0, the minimiser of the population KCD $L(1,\eta)$, at the rate of $O(1/\sqrt{n_e})$.*

Recall in the Two-Stage Approach

Problem statement

- Observe $X_{1:n} \overset{iid}{\sim} P_X$, $Y^{(j)}_{1:m} \overset{iid}{\sim} P^{(j)}_Y$ for $j = 1, \ldots, r$
- Can we find evidence to reject the null $\mathscr{H}_{0,\in}$?

Recall in the Two-Stage Approach

Problem statement

- Observe $X_{1:n} \overset{iid}{\sim} P_X$, $Y^{(j)}_{1:m} \overset{iid}{\sim} P^{(j)}_Y$ for $j = 1, \ldots, r$
- Can we find evidence to reject the null $\mathscr{H}_{0,\in}$?

(1) Epistemic Alignment
= Optimisation with KCD

(2) Conduct Testing
with pseudo samples

Recall in the Two-Stage Approach

Problem statement

- Observe $X_{1:n} \overset{iid}{\sim} P_X$, $Y^{(j)}_{1:m} \overset{iid}{\sim} P^{(j)}_Y$ for $j = 1, \ldots, r$
- Can we find evidence to reject the null $\mathscr{H}_{0,\in}$?

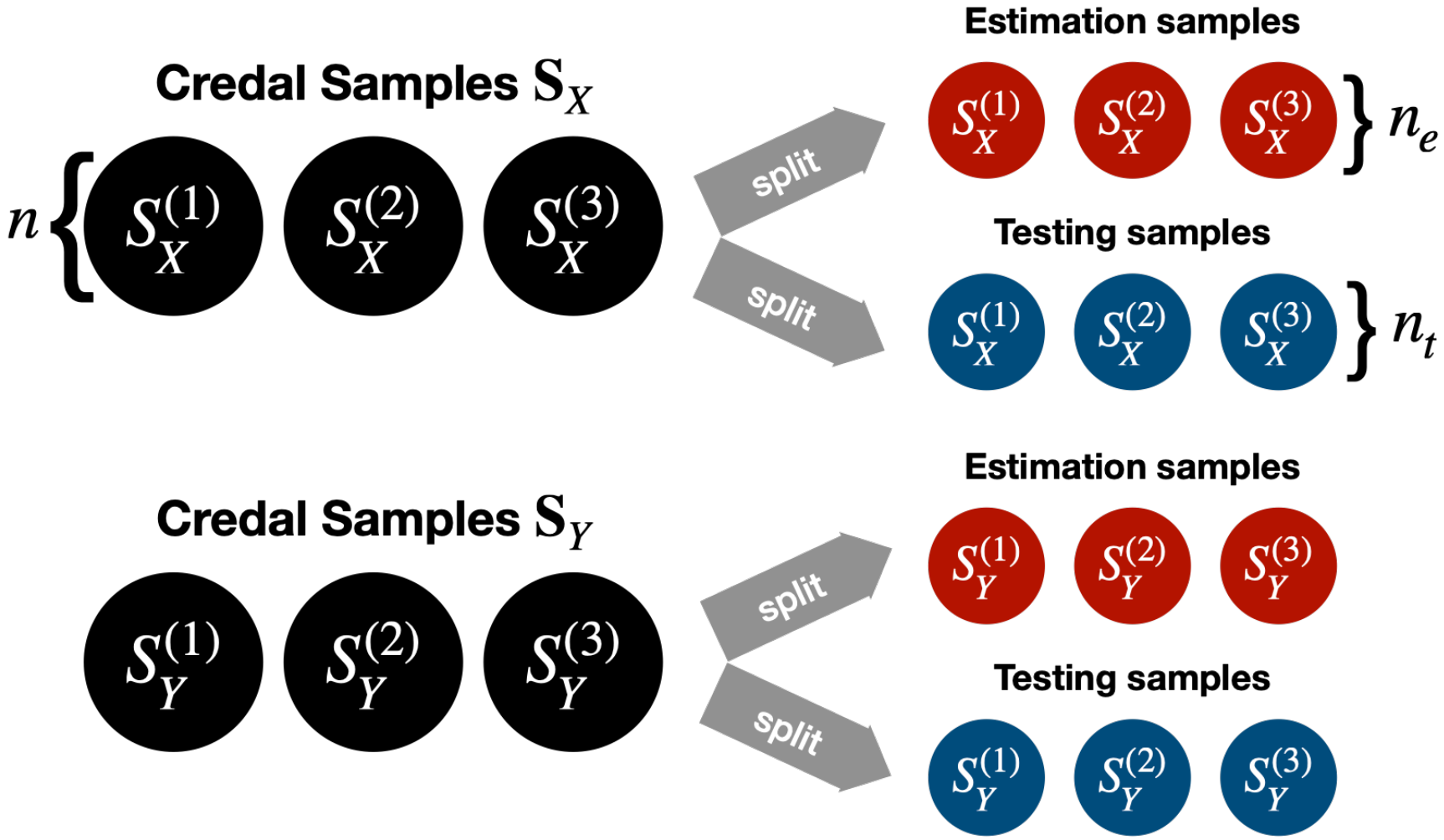
(1) Epistemic Alignment
= Optimisation with KCD

(2) Conduct Testing
with pseudo samples

To avoid unwanted dependencies, we need sample splitting, but how? 50:50, 30:70, 20:80?

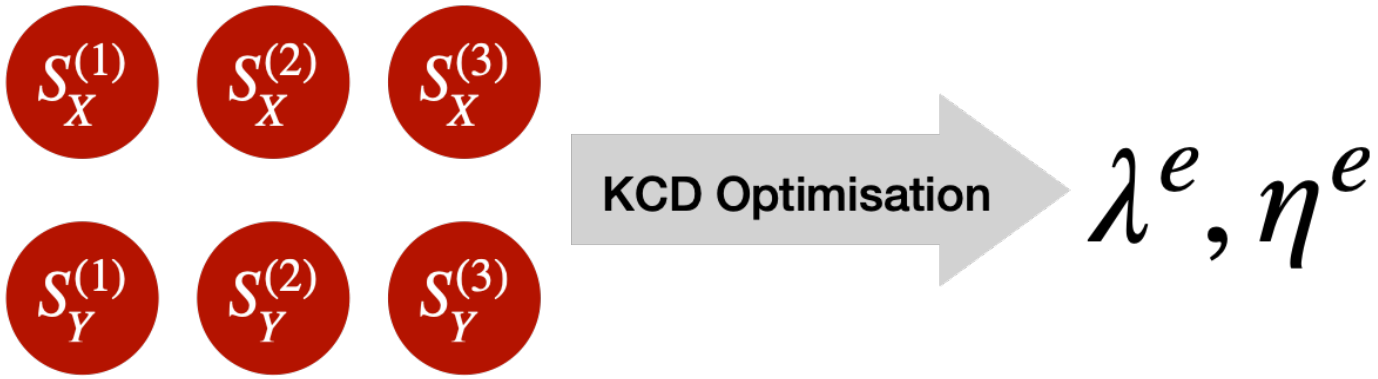
Scientist

Overall Strategy

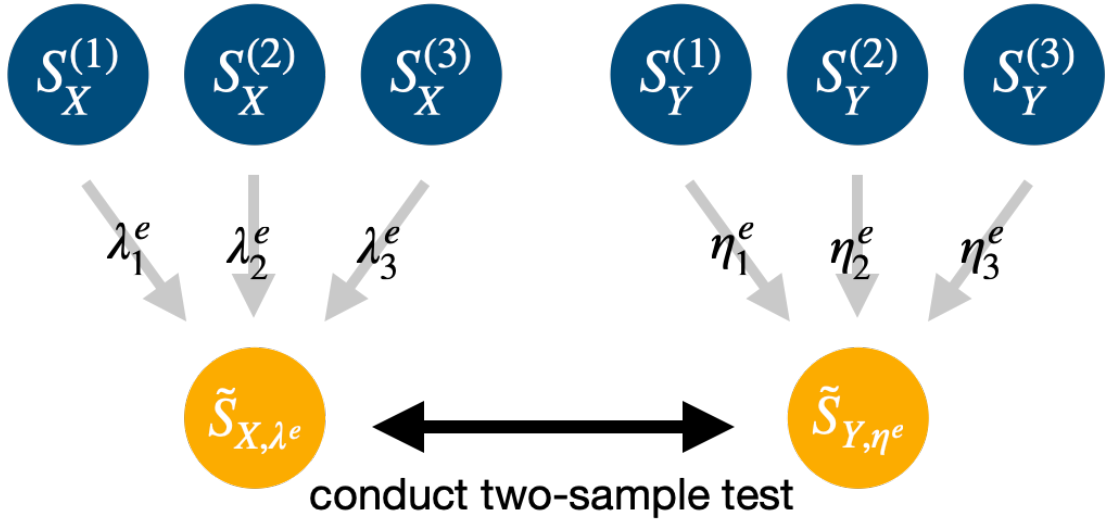


Step (1) Sample Splitting

Estimate and Test



Step (2) Estimation



Step (3+4) Simulate and test

Conventional Wisdom of Fixed Splitting Fails

Scientist

Split samples (at a fixed ratio) to avoid unwanted dependencies

Samples → Estimation + Testing

Conventional Wisdom of Fixed Splitting Fails

Scientist

Split samples (at a fixed ratio) to avoid unwanted dependencies

Samples → Estimation + Testing

BUT….

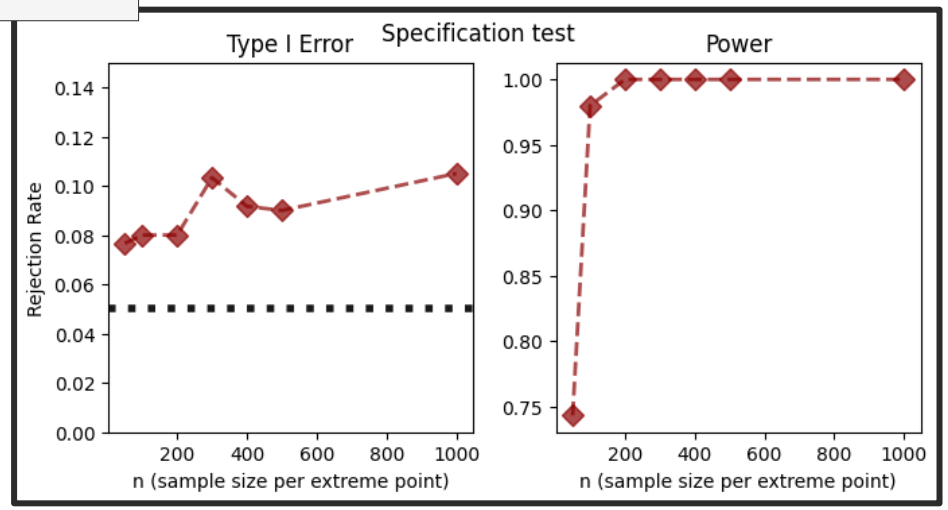
Conventional Wisdom of Fixed Splitting Fails

Scientist

Split samples (at a fixed ratio) to avoid unwanted dependencies

Samples → Estimation + Testing

BUT….



Fixed sample splitting —> **Inflated** Type I control
(**Invalid** testing procedure)

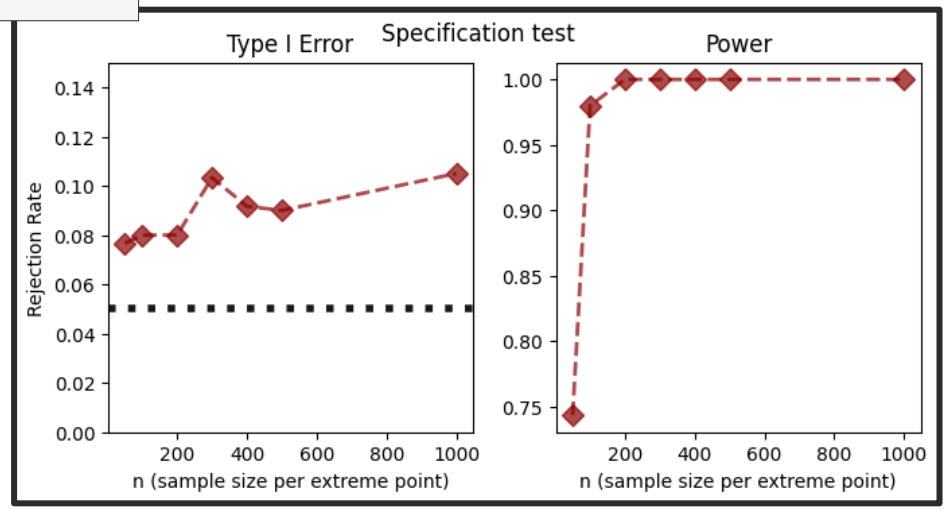
Conventional Wisdom of Fixed Splitting Fails

Scientist

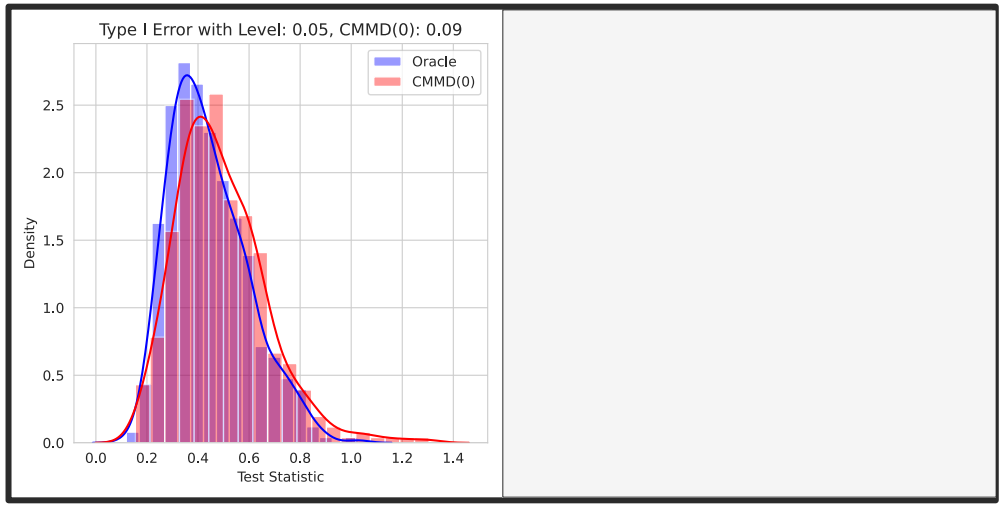
Split samples (at a fixed ratio) to avoid unwanted dependencies

Samples → Estimation + Testing

BUT….



Fixed sample splitting —> **Inflated** Type I control
(**Invalid** testing procedure)



Fixed sample splitting —> shift in the null distribution

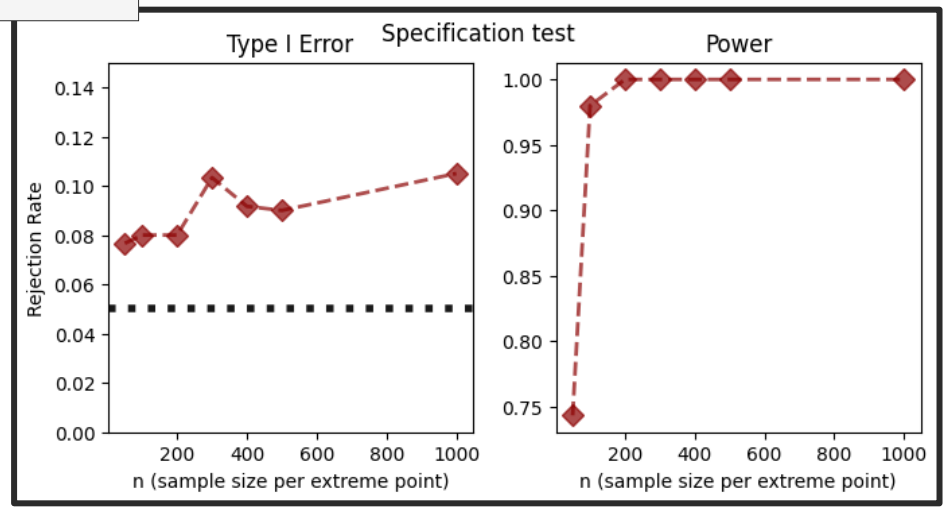
Conventional Wisdom of Fixed Splitting Fails

Scientist

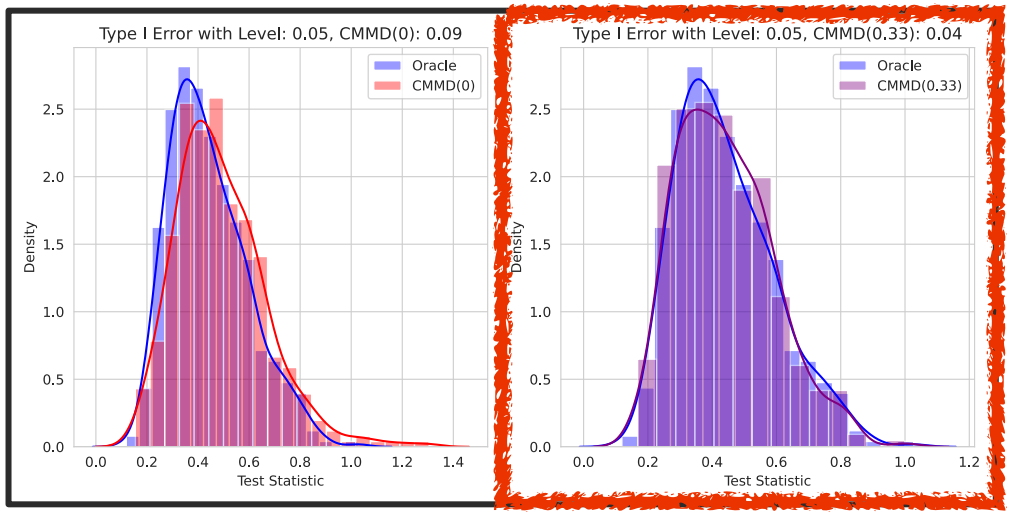
Split samples (at a fixed ratio) to avoid unwanted dependencies

Samples → Estimation + Testing

BUT….



Fixed sample splitting —> **Inflated** Type I control
(**Invalid** testing procedure)



Fixed sample splitting —> shift in the null distribution

Why Adaptive Splitting Works?

Theorem

Under $H_{0,\in}$ and Assumptions 1,2,3, there exists some n_0, such that for $n_t > n_0$,

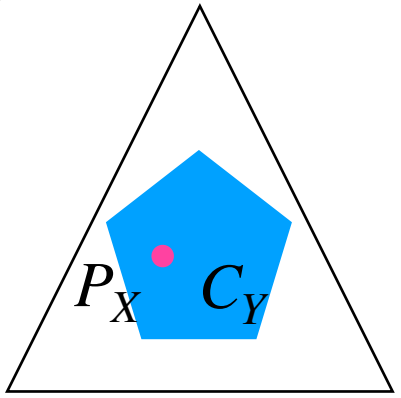
KCD errors $$|n_t \mathcal{L}_{n_t}(1, \eta^e) - n_t \mathcal{L}_{n_t}(1, \eta_0)| = O\left(\sqrt{\frac{n_t}{n_e}}\right).$$

Furthermore, if splitting ratio ρ is chosen adaptively such that $n_t/n_e \to 0$ as $n \to \infty$, then

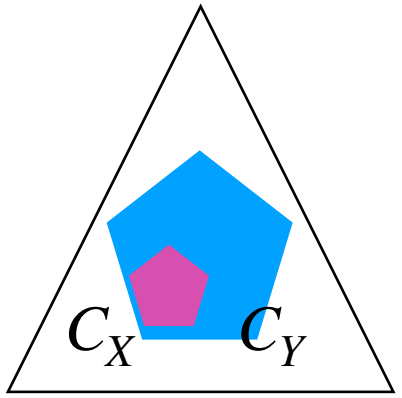
$$n_t \mathcal{L}_{n_t}(1, \eta^e) \xrightarrow{D} \sum_{i=1}^{\infty} \zeta_i Z_i^2,$$

where $\{Z_i\}_{i\geq 1} \overset{i.i.d.}{\sim} N(0,1)$ and $\{\zeta_i\}_{i\geq 1}$ are certain eigenvalues depending on the choice of kernel and P_X, with $\sum_{i=1}^{\infty} \zeta_i < \infty$. Furthermore, under $H_{A,\in}$, $n_t \mathcal{L}_{n_t}(1, \eta^e) \to \infty$ as $n \to \infty$.

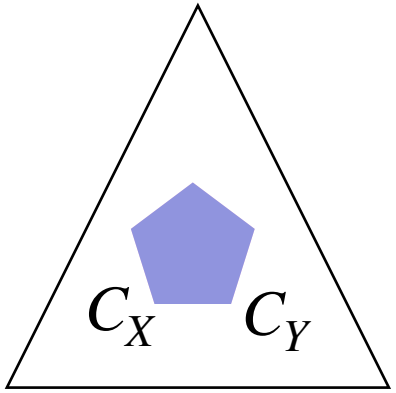
- If $n_t/n_e = c$, the estimation error will not converge.
- Adaptive splitting ratio leads to the same limiting distribution as standard two-sample test.
- It can reject any fixed alternative given enough samples too



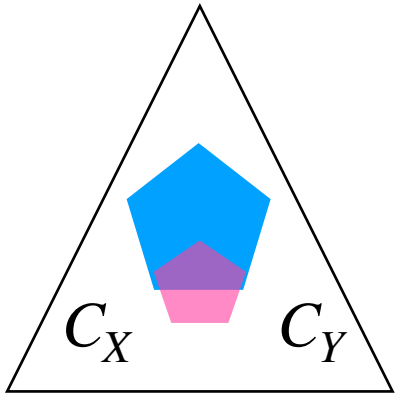
Specification $H_{0,\in} : P_X \in C_Y$



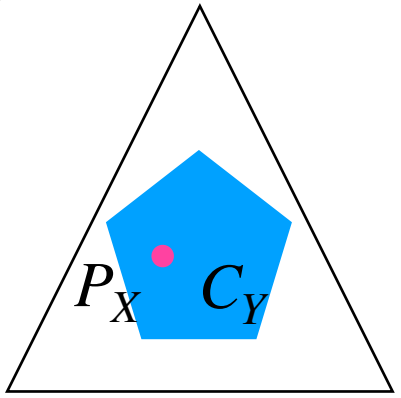
Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$



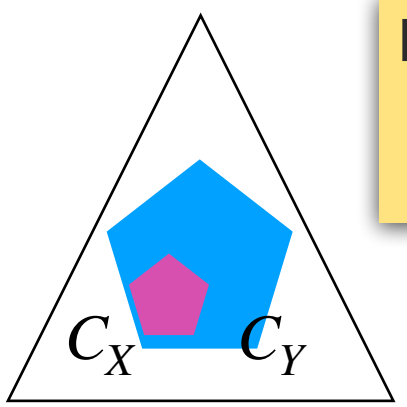
Equality $H_{0,=} : C_X = C_Y$



Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$

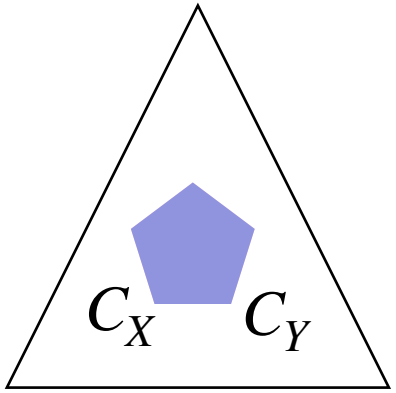


Specification $H_{0,\in} : P_X \in C_Y$

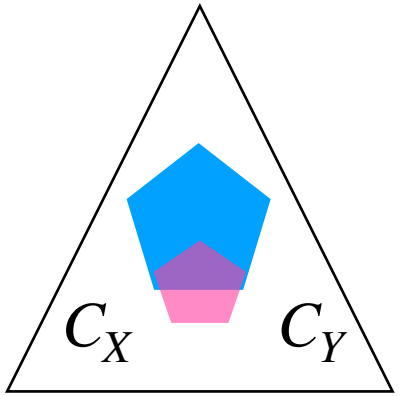


Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$

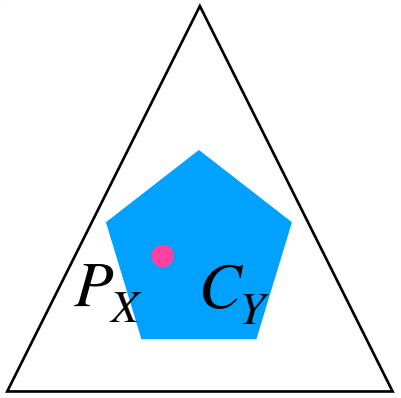
Multiple specification tests
+
Bonferroni correction



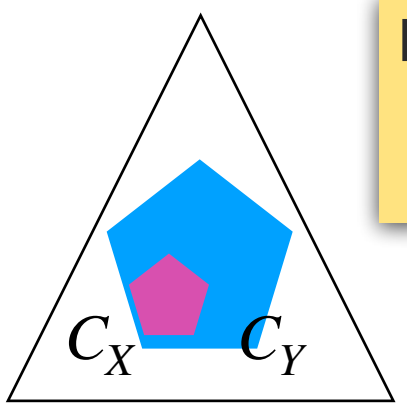
Equality $H_{0,=} : C_X = C_Y$



Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$



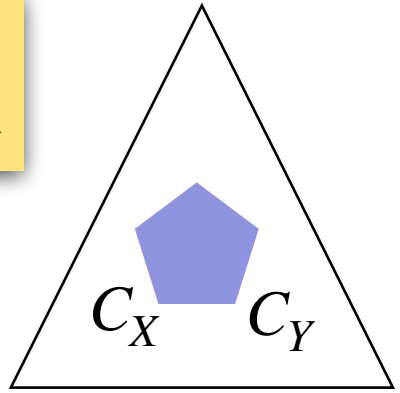
Specification $H_{0,\in} : P_X \in C_Y$



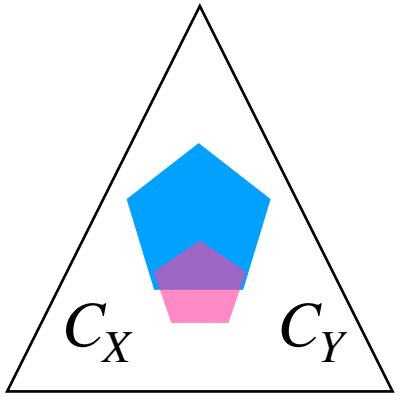
Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$

Multiple specification tests
+
Bonferroni correction

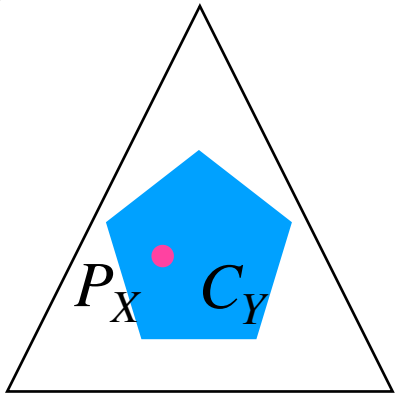
Verify that both
$C_X \subseteq C_Y$ and $C_Y \subseteq C_X$



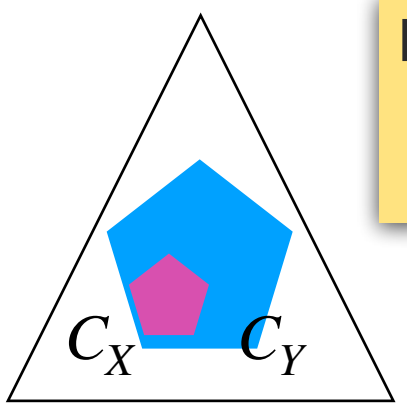
Equality $H_{0,=} : C_X = C_Y$



Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$



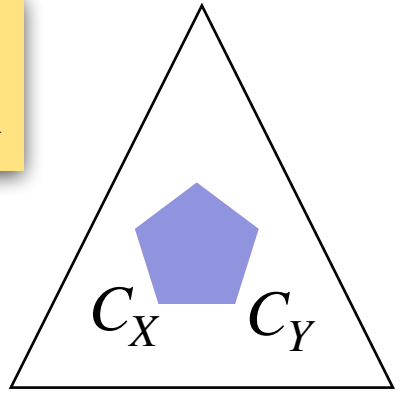
Specification $H_{0,\in} : P_X \in C_Y$



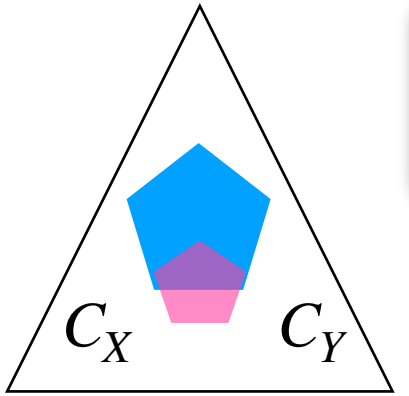
Inclusion $H_{0,\subseteq} : C_X \subseteq C_Y$

Multiple specification tests
+
Bonferroni correction

Verify that both
$C_X \subseteq C_Y$ and $C_Y \subseteq C_X$



Equality $H_{0,=} : C_X = C_Y$



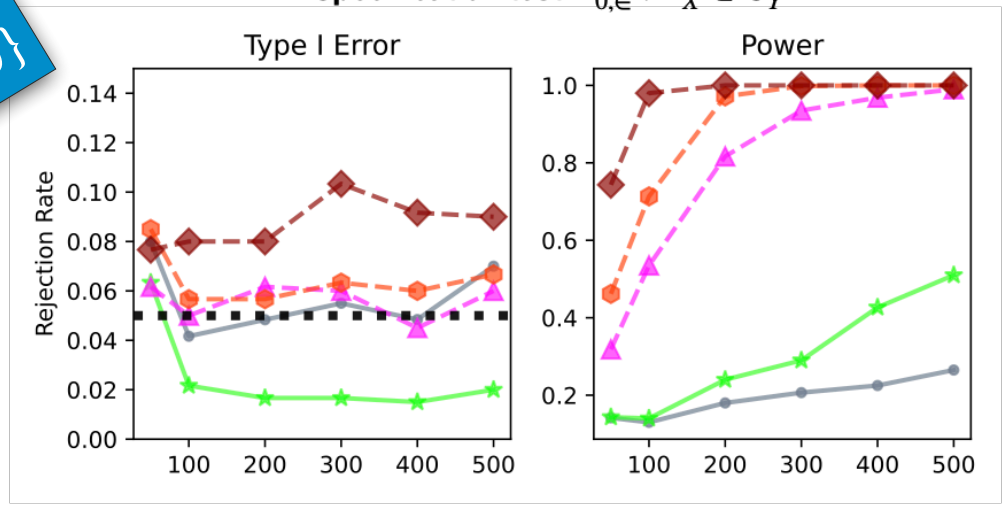
Plausibility $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$

There exists λ and η:
$\lambda^\top \mathbf{P}_X = \eta^\top \mathbf{P}_Y$

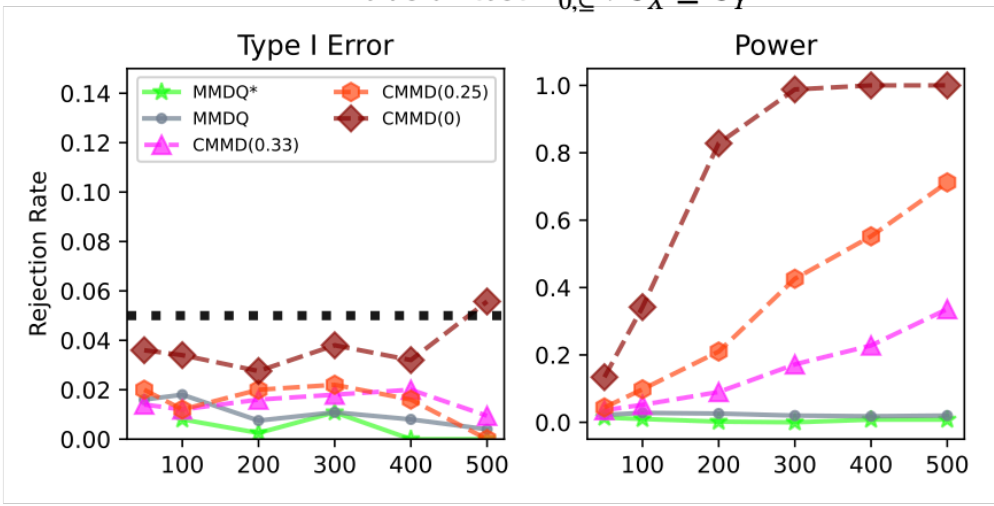
Experiments (Mix of Gaussians v.s. Mix of Students)

$n_t/n_e = n_e^{-\beta}$
$\beta \in \{1/3, 1/4, 0\}$

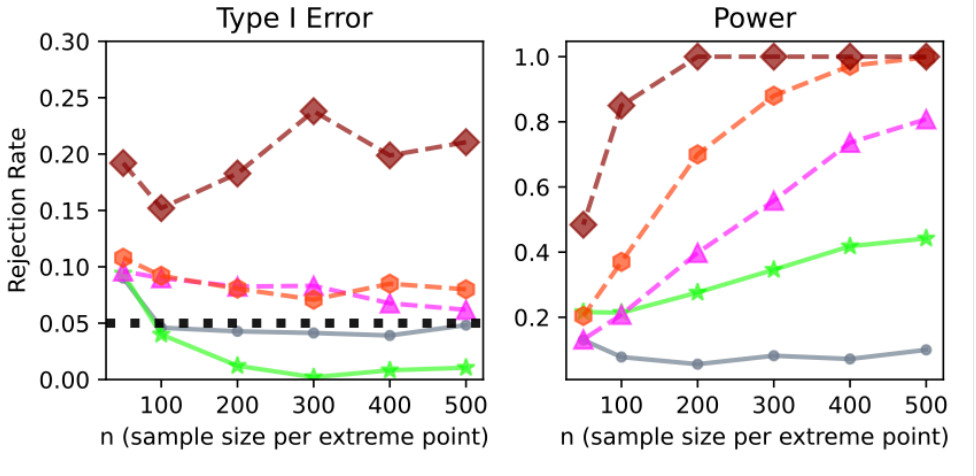
Specification test $H_{0,\in} : P_X \in C_Y$



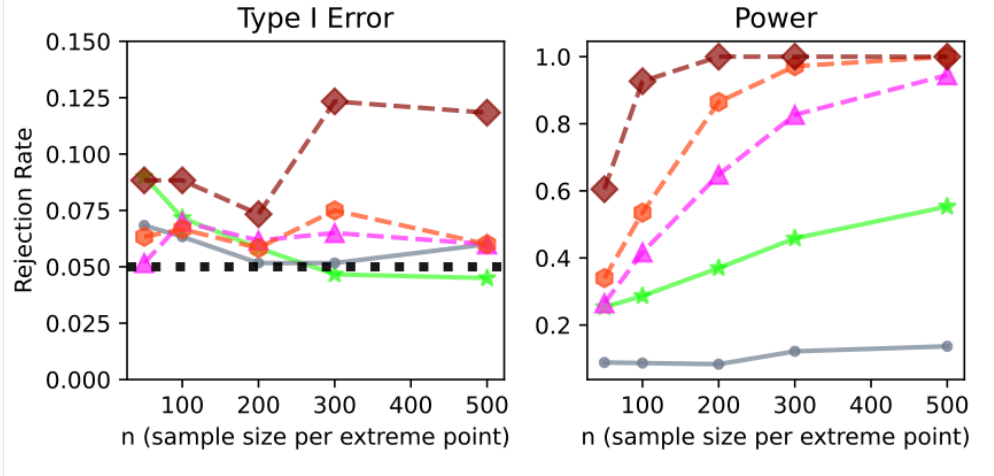
Inclusion test $H_{0,\subseteq} : C_X \subseteq C_Y$



Equality test $H_{0,=} : C_X = C_Y$



Plausibility test $H_{0,\cap} : C_X \cap C_Y \neq \emptyset$



Experiments (MNIST)

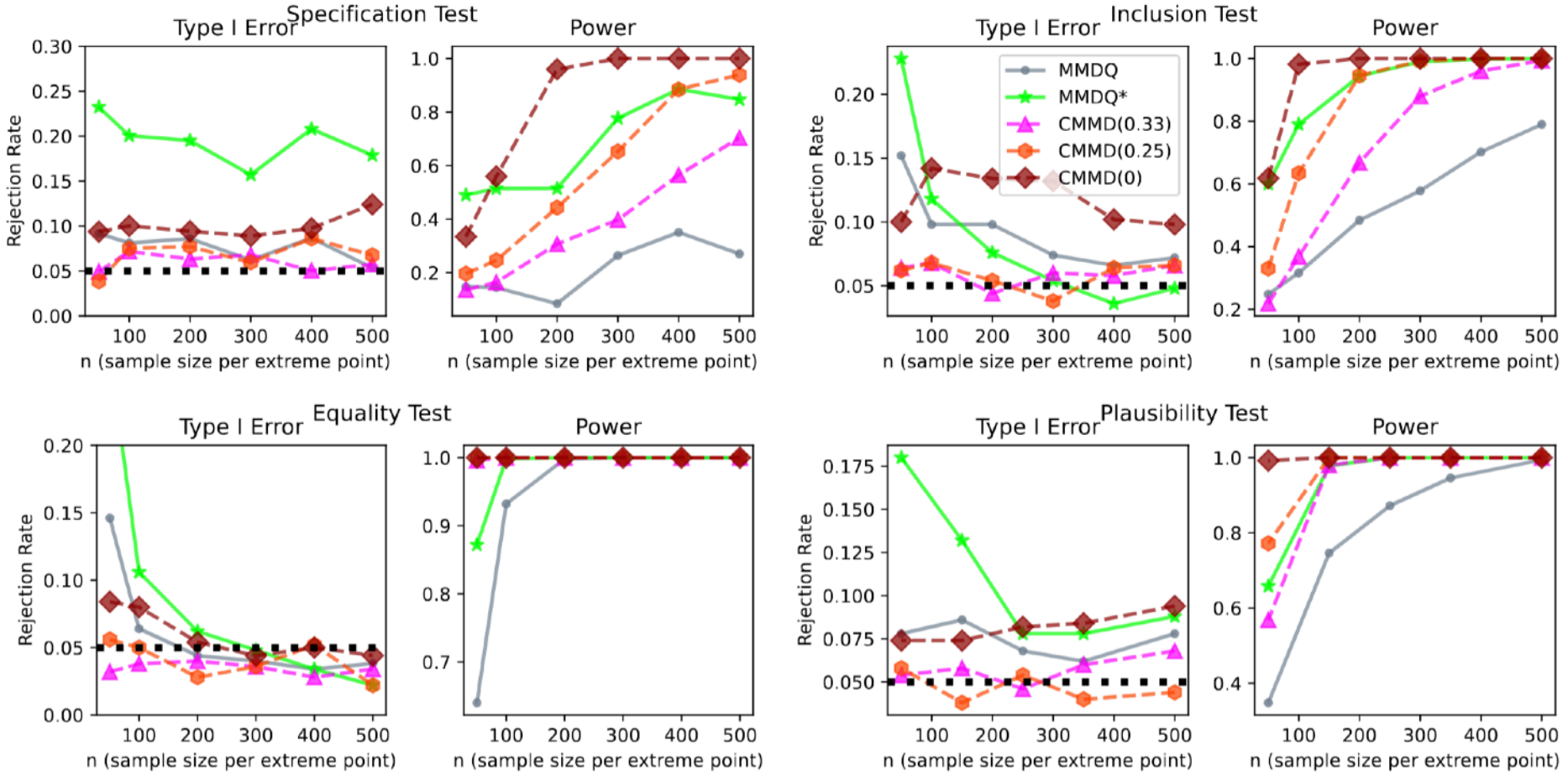


Figure 15: MNIST Credal Testing Experimental Results. In addition to the consistent Type I error inflation for CMMD(0) using a fixed splitting strategy, we observe that MMDQ* also exhibits significantly inflated Type I error rates for small sample sizes. In the specification test, MMDQ* fails to achieve any Type I error control.

Takeaways

- Hypothesis testing, especially two-sample testing, is a fundamental procedure for evidence-based decision-making, but is limited to **precise distributions** only.

- Under **dataset uncertainty**, how should the scientist **adapts**, **represents**, and **propagate** such uncertainty into the testing procedure?

- **Four null credal hypothesis** and their testing procedures are proposed to answer this!
  - Specification $P_X \in C_Y$
  - Inclusion $C_X \subseteq C_Y$
  - Equality $C_X = C_Y$
  - Plausibility $C_X \cap C_Y \neq \varnothing$

Domain Generalisation via Imprecise Learning

Anurag Singh
CISPA

Siu Lun Chau
CISPA

Shahine Bouabid
MIT

Krikamol Muandet
CISPA

ICML International Conference On Machine Learning

Spotlight

Domain Generalisation via Imprecise Learning

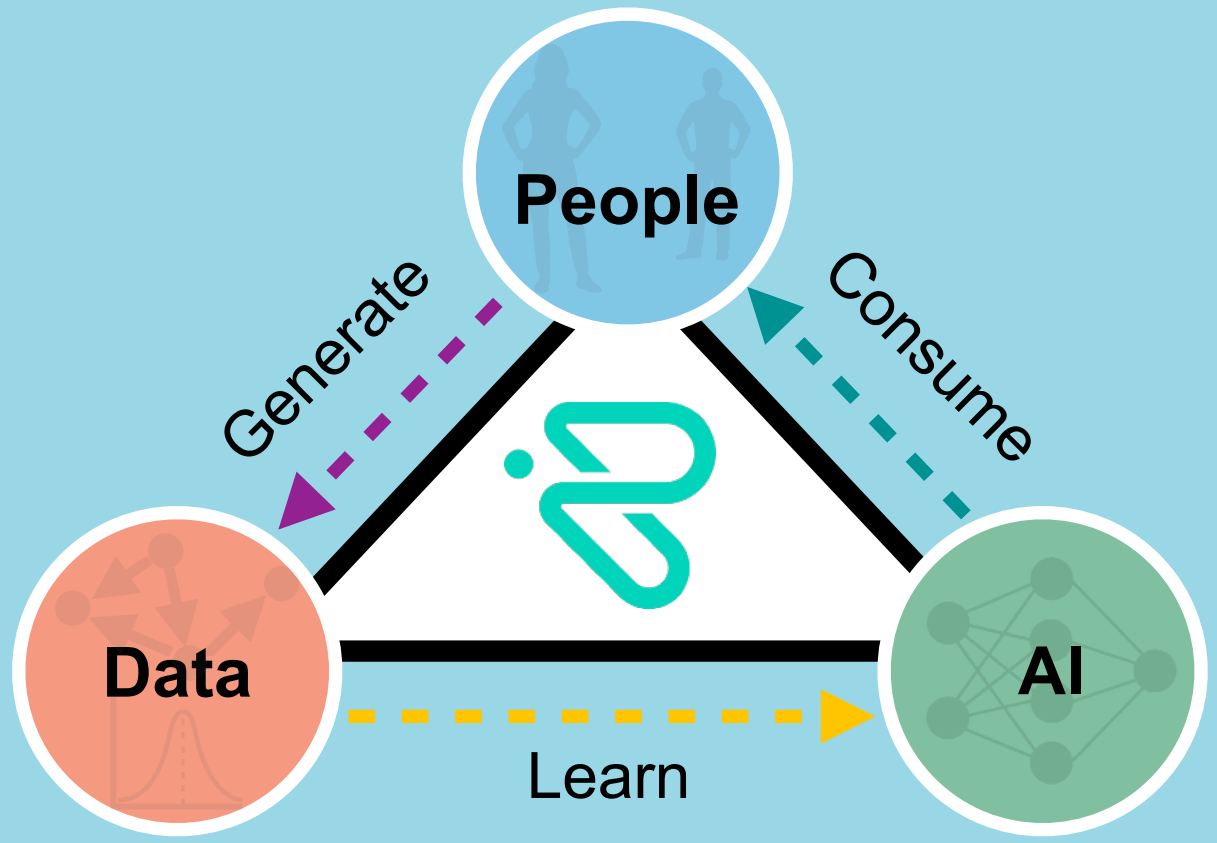
Anurag Singh [1] **Siu Lun Chau** [1] **Shahine Bouabid** [2] **Krikamol Muandet** [1]

Abstract

Out-of-distribution (OOD) generalisation is challenging because it involves not only learning from empirical data, but also deciding among various (LLM) that surpass human-level generalisation capabilities in specific domains.

Despite notable achievements, these systems may catastrophically fail when operated on out-of-domain (OOD)

Rational Intelligence Lab



https://ri-lab.org

muandet@cispa.de