Credal Two-sample Tests of Epistemic Uncertainty

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

[1]Nanyang Technological University, Singapore [2]University College London, United Kingdom
[3]University of Adelaide, Australia, [4]CISPA Helmholtz Center for Information Security, Germany

Table 1. *Different hypotheses to compare credal sets.*

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

This poster presents our recent work at AISTATS 2025 [1], aimed at tackling a long-standing problem in the IP community: *How to statistically compare two credal sets based on samples?*

Introduction. We introduce *credal two-sample tests*, a new hypothesis testing framework for comparing credal sets using samples drawn i.i.d. from each extreme distribution. Unlike classical two-sample tests, which focus on comparing precise distributions, the framework integrates epistemic uncertainty in testing and accommodates a broader and more versatile range of hypotheses. By generalising two-sample testing to comparing credal sets, our framework supports reasoning about equality, inclusion, intersection, and mutual exclusivity — each offering distinct insights into the modeller's epistemic beliefs. Our approach faithfully incorporates epistemic uncertainty into hypothesis testing, leading to more robust and credible conclusions, with kernel-based implementations supporting real-world applications.

Credal hypotheses. Let $\mathbf{P}_X := \{P_X^{(1)}, \dots, P_X^{(\ell)}\}$ and $\mathbf{P}_Y := \{P_Y^{(1)}, \dots, P_Y^{(r)}\}$ be our sets of distributions where we can obtain samples from, and $\mathcal{C}_X :=$ ConvexHull$(\mathbf{P}_X)$, $\mathcal{C}_Y :=$ ConvexHull$(\mathbf{P}_Y)$ the corresponding finitely generated credal set. In Table 1, we list out the hypotheses considered in our paper. Our hypotheses can find their use in the following applications:

1. **Specification** test can be used for finite-mixture model tests and credal set calibration tests.
2. **Inclusion** test can compare whether one credal set contains more imprecision than another one.
3. **Equality** test can be seen as a generalisation of precise two-sample test under distributional ambiguity.
4. **Plausibility** test is a distributionally robust two-sample test, where rejection of the test implies a significant difference in distributions despite ambiguity.

Non-parametric testing procedures. We have sets of i.i.d. samples $\mathbf{S}_X = \{S_X^{(i)}\}_{i=1}^{\ell}$ and $\mathbf{S}_Y = \{S_Y^{(j)}\}_{j=1}^{r}$, each of size n, from each distribution in $\mathbf{P}_X$ and $\mathbf{P}_Y$ respectively. Starting with splitting the samples for estimation (size n_e) and testing (size n_t), the whole procedure consists of two stages:

1. **Epistemic alignment:** Solve the following biconvex optimisation using Maximum Mean Discrepancy:
$$\lambda^e, \boldsymbol{\eta}^e = \arg\min_{\boldsymbol{\eta}\in\Delta_{\ell-1}, \lambda\in\Delta_{r-1}} \widehat{\mathrm{MMD}^2}\left(\lambda^\top \mathbf{P}_X, \boldsymbol{\eta}^\top \mathbf{P}_Y\right)$$
2. **Hypothesis testing.** Resample observations from $\lambda^{e\top}\mathbf{P}_X, \boldsymbol{\eta}^{e\top}\mathbf{P}_Y$ and conduct a precise kernel two-sample test [2].

Theoretical guarantees. Since we test with estimated (e.g. $P_X = \boldsymbol{\eta}^{e\top}\mathbf{P}_Y$) rather than true convex weights where the population credal sets intersect (e.g. $P_X = \boldsymbol{\eta}_0^\top \mathbf{P}_Y$), estimation error could invalidate the procedure. Adaptive sample splitting ensures validity. To illustrate, we present results for the specification test.

Theorem 1 (Validity and consistency of our test)**.** *Under $H_{0,\in}$ and regularity assumptions, when n is large,*
$$\left| n_t \widehat{\mathrm{MMD}^2}(P_X, \boldsymbol{\eta}^{e\top}\mathbf{P}_Y) - n_t \widehat{\mathrm{MMD}^2}(P_X, \boldsymbol{\eta}_0^\top \mathbf{P}_Y) \right| = O\left(\sqrt{n_t/n_e}\right)$$
Therefore, if the split is chosen such that $n_t/n_e \to 0$, then $n_t \widehat{\mathrm{MMD}^2}(P_X, \boldsymbol{\eta}^{e\top}\mathbf{P}_Y) \xrightarrow{D} \sum_{i=1}^{\infty} \zeta_i Z_i^2$. That is, the same limiting distribution as if no estimation had happened. Furthermore, under $H_{A,\in}$, $n_t \widehat{\mathrm{MMD}^2}(P_X, \boldsymbol{\eta}^{e\top}\mathbf{P}_Y) \to \infty$.

Experiments. We validate our results with numerical experiments on synthetic and MNIST datasets, which align with our theoretical analysis: valid Type I control and diminishing Type II error.

References

[1] Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. "Credal Two-sample Tests of Epistemic Uncertainty". In: *International Conference on Artificial Intelligence and Statistics* (2025).

[2] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.