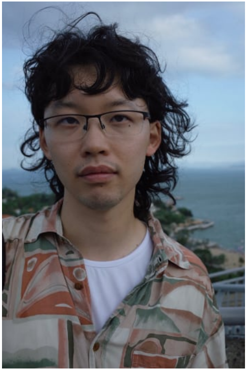
Rank-transformed subsampling: inference for exchangeable p-values

Rajen D. Shah (Statistical Laboratory, University of Cambridge)

Warwick Statistics CRiSM Seminar
26 April 2023

Joint work with:



Richard Guo

Outline

- Randomised tests are useful
- Drawbacks of randomised tests
- Rank-transformed subsampling
- Applications

Randomised tests

Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

1. **'Hunt'**: Use Part A to determine which test to use to target the alternative the data appear to have come from.
2. **Test**: Apply the test to Part B.

Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

1. **'Hunt'**: Use Part A to determine which test to use to target the alternative the data appear to have come from.
2. **Test**: Apply the test to Part B.

In the second step, we may treat the chosen test as fixed in advance and there is no need to account for the 'hunting' in step 1.

Hunt step can be as elaborate as needed in order to find an appropriate test.

Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis H_0 given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

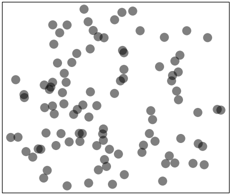
1. **'Hunt'**: Use Part A to determine which test to use to target the alternative the data appear to have come from.
2. **Test**: Apply the test to Part B.

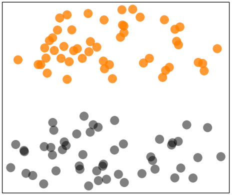
In the second step, we may treat the chosen test as fixed in advance and there is no need to account for the 'hunting' in step 1.

Hunt step can be as elaborate as needed in order to find an appropriate test.

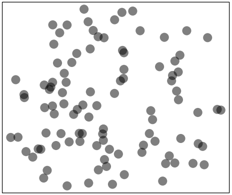
Strategy particularly useful when $H_0 = \cap_{\delta \in \mathcal{D}} H_0(\delta)$, so $H_1 = \cup_{\delta \in \mathcal{D}} H_0^c(\delta)$.

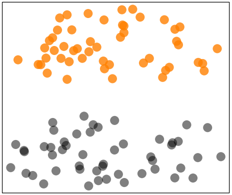
Testing for clustering structure





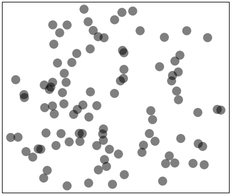
Testing for clustering structure





Clustering algorithms cannot be used directly, as they may return clusters when none are truly present.

Testing for clustering structure

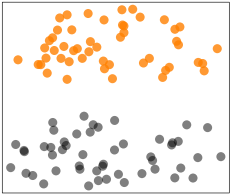


Clustering algorithms cannot be used directly, as they may return clusters when none are truly present.

We can formalise our null hypothesis as testing for unimodality.

Various notions exist in multiple dimensions including *linear unimodality*:

$$X \in \mathbb{R}^p \text{ is unimodal if } a^T X \text{ is unimodal } \forall\, a \neq \mathbf{0}.$$

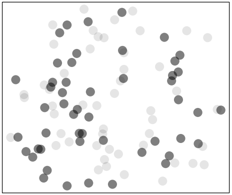


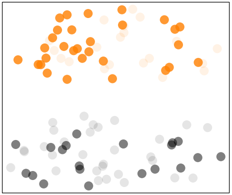
That is,

$$H_0 : \bigcap_{a \neq \mathbf{0}} \{a^T X \text{ is unimodal}\},$$

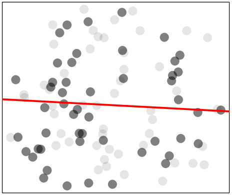
$$H_1 : \exists\, a \neq 0 \text{ such that } a^T X \text{ is not unimodal.}$$

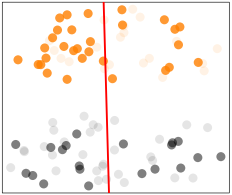
Testing for clustering structure



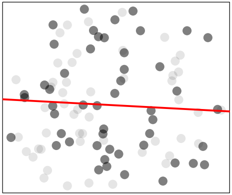


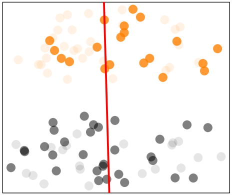
Testing for clustering structure



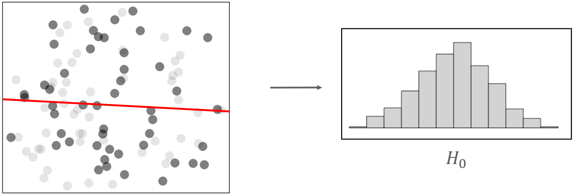


Testing for clustering structure

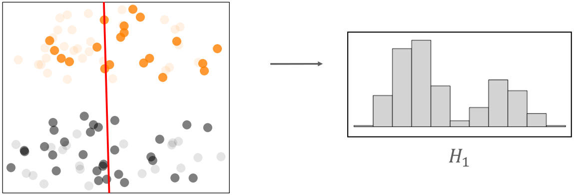




Testing for clustering structure

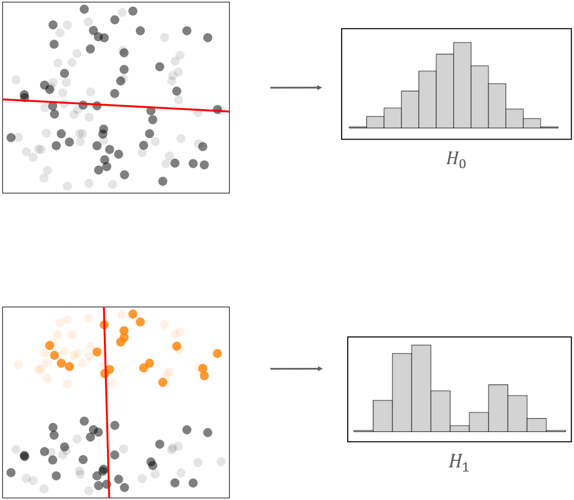


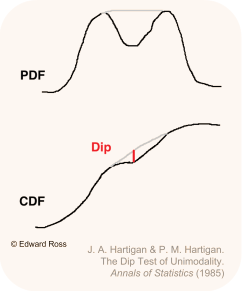
H_0



H_1

Testing for clustering structure





Testing under distributional shifts (Thams et al., 2023)

Consider testing the conditional independence $H_0 : X \perp\!\!\!\perp Y \,|\, Z$ given iid copies $(X_i, Y_i, Z_i)_{i=1}^n$.

Suppose instead we had samples from a distribution with density

$$\begin{aligned} p(x,y,z)\frac{q(z)}{p(z|x)} &= p(y|x,z)p(z|x)p(x)\frac{q(z)}{p(z|x)} \\ &= p(y|x,z)q(z)p(x) \\ &\overset{\text{under } H_0}{=} p(y|z)q(z)p_X(x). \end{aligned}$$

Thus the reweighted marginal distribution of (X, Y) would be $p(y)p(x)$, so $X \perp\!\!\!\perp Y$.

In the reweighted distribution, the null of conditional independence becomes the simpler null of independence.

Testing under distributional shifts (Thams et al., 2023)

Consider testing the conditional independence $H_0 : X \perp\!\!\!\perp Y \mid Z$ given iid copies $(X_i, Y_i, Z_i)_{i=1}^n$.

Suppose instead we had samples from a distribution with density

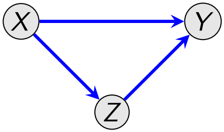
$$\begin{aligned} p(x,y,z)\frac{q(z)}{p(z|x)} &= p(y|x,z)p(z|x)p(x)\frac{q(z)}{p(z|x)} \\ &= p(y|x,z)q(z)p(x) \\ &\overset{\text{under } H_0}{=} p(y|z)q(z)p_X(x). \end{aligned}$$

Thus the reweighted marginal distribution of (X, Y) would be $p(y)p(x)$, so $X \perp\!\!\!\perp Y$.

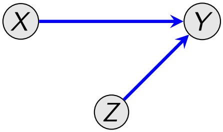
In the reweighted distribution, the null of conditional independence becomes the simpler null of independence.

If we know $p(z|x)$ we can always obtain a sample from the reweighted distribution through e.g. rejection sampling.

Testing generalised conditional independencies (Robins, 1999)

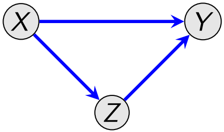


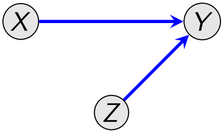
$p(y|x,z)p(z|x)p(x)$



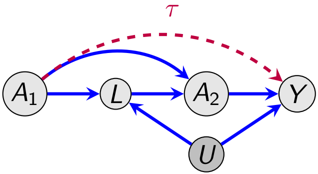
$p(y|x,z)q(z)p(x)$

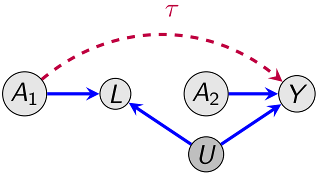
Testing generalised conditional independencies (Robins, 1999)





$p(y|x,z)p(z|x)p(x)$ $\qquad$ $p(y|x,z)q(z)p(x)$





Reweighting by $\frac{q(a_2)}{p(a_2|a_1,l)}$

A_1, A_2: 1st and 2nd treatments. L, Y: 1st and 2nd outcomes, confounded by e.g. health status U.

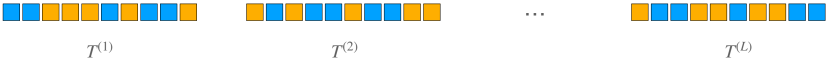
Drawbacks of randomised tests

Replicability and Power

Replicability: Conclusions may depend delicately on the random seed used.

Power loss: Tests may not be making full use of the data.

Consider repeatedly applying the same randomised procedure to the same data.



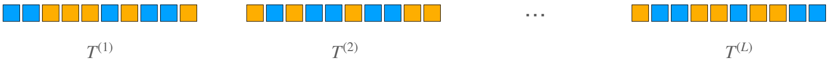
Test statistics $T^{(1)}, \ldots, T^{(L)}$ are exchangeable.

Replicability and Power

Replicability: Conclusions may depend delicately on the random seed used.

Power loss: Tests may not be making full use of the data.

Consider repeatedly applying the same randomised procedure to the same data.



Test statistics $T^{(1)}, \ldots, T^{(L)}$ are exchangeable.

Suppose $T^{(1)} \sim \mathcal{N}(\mu, 1)$ and we want to test

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0.$$

Consider aggregating them by, e.g.,

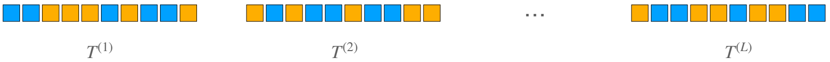
$$S = \left(T^{(1)} + \cdots + T^{(L)}\right)/L.$$

Replicability and Power

Replicability: Conclusions may depend delicately on the random seed used.

Power loss: Tests may not be making full use of the data.

Consider repeatedly applying the same randomised procedure to the same data.



Test statistics $T^{(1)}, \dots, T^{(L)}$ are exchangeable.

Suppose $T^{(1)} \sim \mathcal{N}(\mu, 1)$ and we want to test

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0.$$

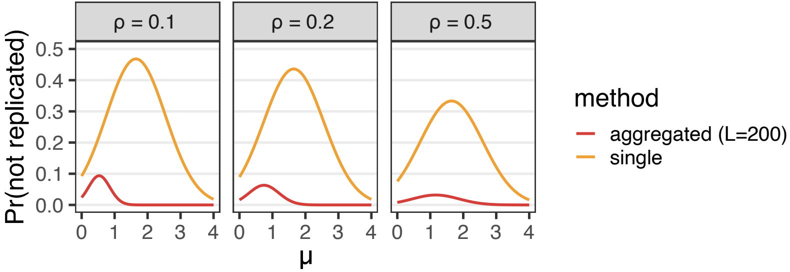
Consider aggregating them by, e.g.,

$$S = \left(T^{(1)} + \cdots + T^{(L)} \right) / L.$$

Let us further model $\left(T^{(1)}, \dots, T^{(L)}\right)$ as jointly normal with correlation $\rho > 0$.

Toy example: replicability

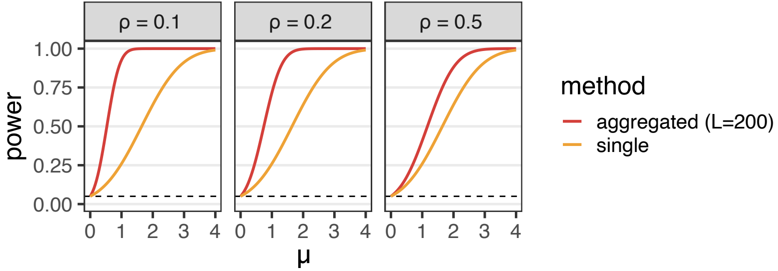
Single-split test	Reject H_0 when $T^{(1)}$ is large
Aggregated test	Reject H_0 when $S = \left(T^{(1)} + \cdots + T^{(L)}\right)/L$ is large.



"Not replicated" = one acceptance and one rejection in two runs.

Toy example: Power

Single-split test Reject H_0 when $T^{(1)}$ is large
Aggregated test Reject H_0 when $S = \left(T^{(1)} + \cdots + T^{(L)}\right)/L$ is large.



Complex dependence in practice

In the toy example, the aggregated test was based on

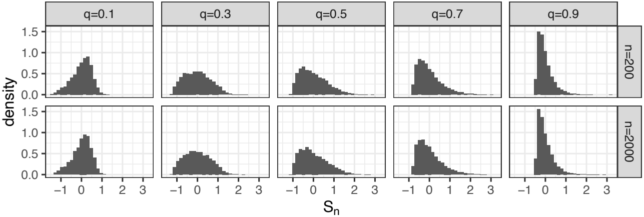
$$S \sim \mathcal{N}(\mu, 1/L + \rho(L-1)/L).$$

Complex dependence in practice

In the toy example, the aggregated test was based on

$$S \sim \mathcal{N}(\mu, 1/L + \rho(L-1)/L).$$

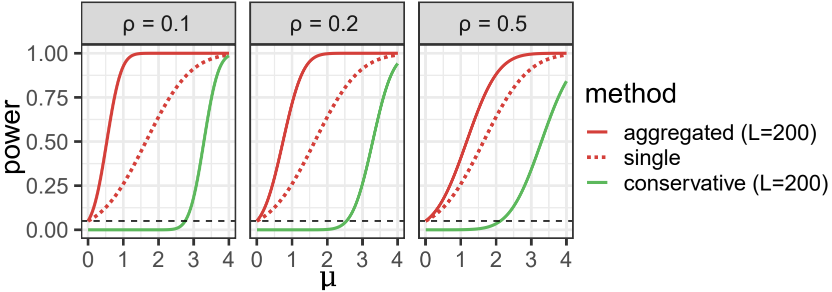
In reality, however, the dependence among $T^{(1)}, \ldots, T^{(L)}$ can be complex and there is no good description or approximation (beyond symmetry).



Distribution of S in a real hunt-and-test example case from Kim & Ramdas (2020)

Toy example: Power

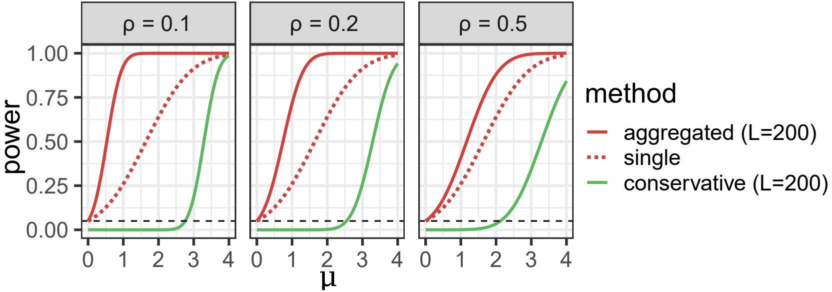
Single-split test	Reject H_0 when $T^{(1)}$ is large
Aggregated test	Reject H_0 when $S = \left(T^{(1)} + \cdots + T^{(L)}\right)/L$ is large.



The conservative rule rejects when $S > 2z_\alpha$.

Toy example: Power

Single-split test	Reject H_0 when $T^{(1)}$ is large
Aggregated test	Reject H_0 when $S = \left(T^{(1)} + \cdots + T^{(L)}\right)/L$ is large.



The conservative rule rejects when $S > 2z_\alpha$.

Other conservative approaches similarly lose power. (Vovk & Wang, 2020; Vovk et al., 2021; DiCiccio et al., 2020; Meinshausen et al., 2009;...

Rank-transformed subsampling

Setup

Have exchangeable test statistics $T^{(1)}, \ldots, T^{(L)}$.

A1 Under $P \in H_0$, T_n is asymptotically $U(0,1)$. (Also works for $T_n \xrightarrow{d} \mathcal{N}(0,1)$).

Setup

Have exchangeable test statistics $T^{(1)}, \dots, T^{(L)}$.

A1 Under $P \in H_0$, T_n is asymptotically $U(0,1)$. (Also works for $T_n \xrightarrow{d} \mathcal{N}(0,1)$).

Choose a deterministic aggregation function $S : \mathbb{R}^L \to \mathbb{R}$ to give $S_n := S(T^{(1)}, \dots, T^{(L)})$.

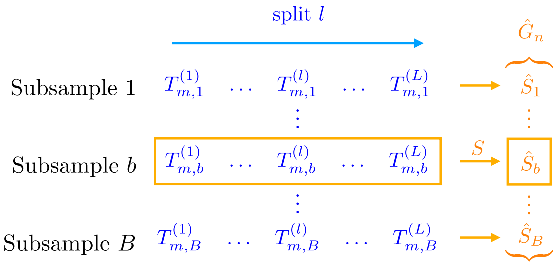
A2 Under $P \in H_0$, S_n converges to (unknown) distribution G_P with bounded density.

We wish to construct a test / form a p-value based on S_n.

Subsampling (e.g. Politis et al. (1999))

We use subsampling to estimate the asymptotic distribution G_P.

Choose $b = 1, \dots, B$ subsamples of size $m := \lfloor n / \log n \rfloor$.



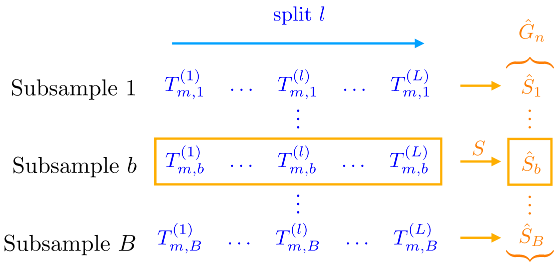
$\hat{G}_n := \text{ECDF}\,\{\hat{S}_1, \dots, \hat{S}_B\}$.

By standard consistency of subsampling (e.g. Politis et al. (1999)) for $P \in H_0$, $\|\hat{G}_n - G_P\|_\infty \xrightarrow{p} 0$.

Subsampling (e.g. Politis et al. (1999))

We use subsampling to estimate the asymptotic distribution G_P.

Choose $b = 1, \ldots, B$ subsamples of size $m := \lfloor n / \log n \rfloor$.



$\hat{G}_n := \text{ECDF}\,\{\hat{S}_1, \ldots, \hat{S}_B\}$.

By standard consistency of subsampling (e.g. Politis et al. (1999)) for $P \in H_0$, $\|\hat{G}_n - G_P\|_\infty \xrightarrow{p} 0$.

But $\hat{G}_n$ will continue to approximate G_P under local alternatives.

Rank transform

We have not yet used that we *know* the asymptotic null distribution of $T_n^{(1)}$ (A1).

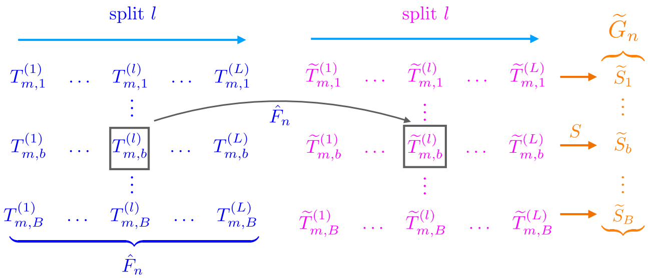
Rank transform

We have not yet used that we *know* the asymptotic null distribution of $T_n^{(1)}$ (A1).

Replace each $T_{m,b}^{(l)}$ by its normalised rank within the matrix:

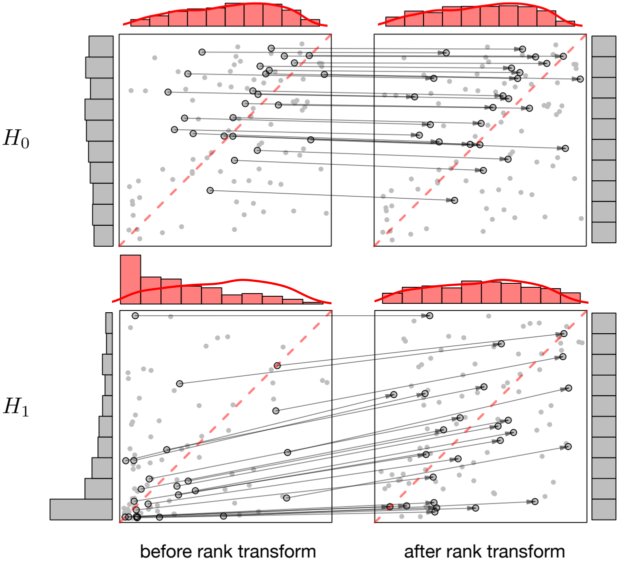
$$\tilde{T}_{m,b}^{(l)} := \#\{T_{m,b'}^{(l')} \leq T_{m,b}^{(l)}\}/BL = \hat{F}_n\left(T_{m,b}^{(l)}\right),$$

where $\hat{F}_n$ is the ECDF of $\{T_{m,b}^{(l)}\}$.



We use $\tilde{G}_n :=$ empirical measure of $\{\tilde{S}_b\}$ as our reference for testing.

Rank transform: illustration



Theory

Theorem (Size control)

Under (A1) and (A2), for $P \in H_0$, $\mathbb{P}_P \left\{ S_n < \tilde{G}_n^{-1}(\alpha) \right\} \to \alpha.$

Theory

Theorem (Size control)

Under (A1) and (A2), for $P \in H_0$, $\mathbb{P}_P \left\{ S_n < \tilde{G}_n^{-1}(\alpha) \right\} \to \alpha$.

Theorem (Local alternatives)

Suppose additionally that for every sequence $P_n \in \mathcal{P}$ *(where* $\mathcal{P}$ *is the set of possible distributions) such that* $P_n \xrightarrow{d} P \in H_0$*, it holds that*

$$\left(F_{n,P_n}(T_n^{(1)}), \ldots, F_{n,P_n}(T_n^{(L)}) \right) \to_d (C^{(1)}, \ldots, C^{(L)}) \tag{3.1}$$

for some $(C^{(1)}, \ldots, C^{(L)})$ *whose distribution does not depend on the sequence* P_n*. Then, for every sequence* $P_n \in \mathcal{P}$ *such that* $P_n \xrightarrow{d} P$*, we have that*

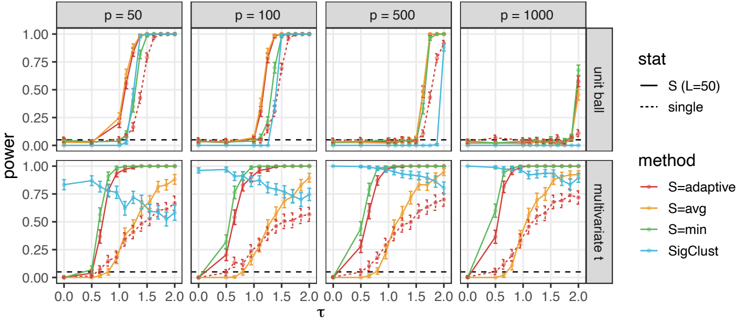
$$\tilde{G}_n^{-1}(1-\alpha) \xrightarrow{p} G_P^{-1}(1-\alpha),$$

i.e. the critical value of our test converges to that of the 'oracle' test.

Applications

Testing unimodality

Hunt: 2-means clustering, Test: T_n = asymptotic dip test p-value. $L = 50$ splits.
Setting: Mixture of two d-dimensional (unit ball, multivariate t) distributions separated τ away.
Aggregation: Consider S = avg, S = min.



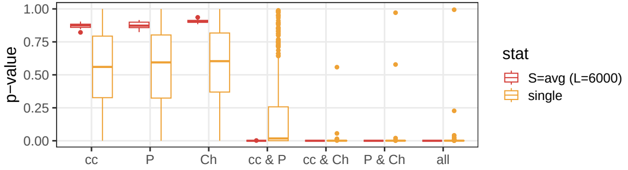
SigClust is a competing method based on multivariate normal mixture.

Adaptive algorithm version available that adapts to the aggregation function with better performance.

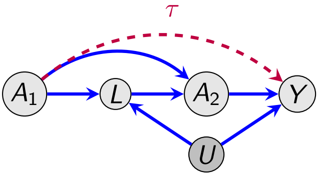
Gene expression of cancer subtypes

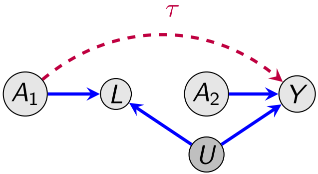
Three types of renal cell carcinoma:
clear cell (ccRCC), papillary (PRCC) and chromophobe (ChRCC).

ICGC/TCGA Pan-Cancer dataset: Expression levels of 1,000 genes. $L = 6000$ splits.



Generalised conditional independence



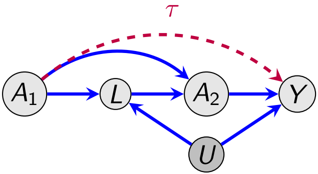


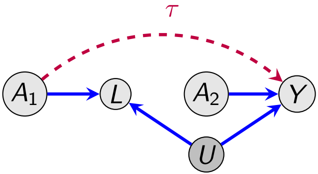
Reweighting by $\frac{q(a_2)}{p(a_2|a_1,l)}$ to give distribution Q

A_1, A_2: 1st and 2nd treatments. L, Y: 1st and 2nd outcomes, confounded by e.g. health status U.

H_0: A_1 has no direct effect on Y.

Generalised conditional independence





Reweighting by $\frac{q(a_2)}{p(a_2 | a_1, l)}$ to give distribution Q

A_1, A_2: 1st and 2nd treatments. L, Y: 1st and 2nd outcomes, confounded by e.g. health status U.

H_0: A_1 has no direct effect on Y.

We can construct tests based on $\mathrm{Cov}_Q(A_1, Y) = 0$.
IPW test (Robins, 1999) proposes

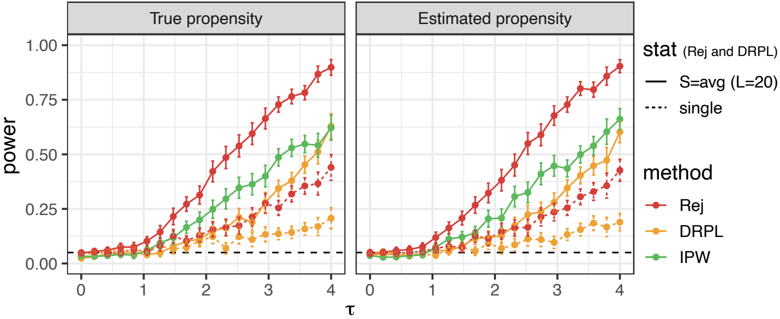
$$Z_i := \frac{Y_i(A_{1,i} - \mathbb{E}A_{1,i})}{p(A_{2,i} \mid A_{1,i}, L_i)}, \quad T := \sum_i Z_i \Big/ \sqrt{\sum_i Z_i^2} \xrightarrow{d} \mathcal{N}(0, 1).$$

Simple setting

Setting: $A_1 \sim \text{Ber}(1/2)$, $A_2 \sim \text{Ber}(\text{expit}(2A_1 - L + 2))$ and

$$U \sim \mathcal{N}_4(0, \Sigma_{ij} = 2^{-|i-j|}), \quad L = A_0 + \beta_{U,L}^T U + \varepsilon_L, \quad Y = \tau A_1 - A_2 + \beta_{U,Y}^T U + \varepsilon_Y.$$

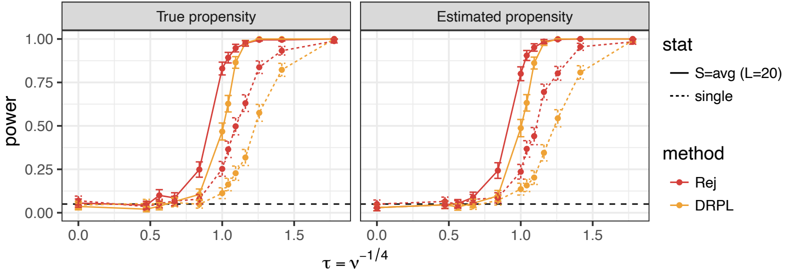
Also consider alternative to rejection sampling: 'distinct replacement sampling' (DRPL) (Thams et al., 2021).



S is the average of permutation p-values computed on $L = 20$ accepted/resampled data.

More complex setting

A more difficult setting where IPW is inapplicable.



T_n: permutation p-value with HSIC (Gretton et al., 2012) as the statistic on the accepted/resampled data.

Summary

- Randomised tests can be useful for a variety of applications:
  - Testing for clustering structure, testing for the presence of signal in high-dimensional data, goodness-of-fit testing, (nonparametric) variable significance testing,...
  - Testing under distributional shifts
  - Testing or confidence interval construction based on double / debiased machine learning
- Replicability and power issues may hamper their adoption in practice.
  - Conservative aggregation rules improve replicability, but power can degrade significantly.
- While a naive subsampling also suffers from power loss, rank-transformed subsampling uses knowledge of the null distribution to avoid these issues.

Thank you for listening.