Rank-transformed subsampling: Inference for multiple data splitting and exchangeable p-values

F. Richard Guo *[1] and Rajen D. Shah †[2]

[1]Department of Statistics, University of Michigan, Ann Arbor, USA
[2]Statistical Laboratory, University of Cambridge, Cambridge, UK

September 5, 2024

Abstract

Many testing problems are readily amenable to randomised tests such as those employing data splitting. However despite their usefulness in principle, randomised tests have obvious drawbacks. Firstly, two analyses of the same dataset may lead to different results. Secondly, the test typically loses power because it does not fully utilise the entire sample. As a remedy to these drawbacks, we study how to combine the test statistics or p-values resulting from multiple random realisations such as through random data splits. We develop rank-transformed subsampling as a general method for delivering large sample inference about the combined statistic or p-value under mild assumptions. We apply our methodology to a wide range of problems, including testing unimodality in high-dimensional data, testing goodness-of-fit of parametric quantile regression models, testing no direct effect in a sequentially randomised trial and calibrating cross-fit double machine learning confidence intervals. In contrast to existing p-value aggregation schemes that can be highly conservative, our method enjoys type-I error control that asymptotically approaches the nominal level. Moreover, compared to using the ordinary subsampling, we show that our rank transform can remove the first-order bias in approximating the null under alternatives and greatly improve power.

Keywords—Data-splitting; Cross-fitting; Goodness-of-fit; Rejection sampling; Subsampling; Unimodality; Verma constraint.

1 Introduction

Many modern statistical procedures are randomised in the sense that the output is a random function of the data. A prominent class of such procedures are hypothesis testing methods that involve splitting the dataset into independent parts (Moran, 1973; Cox, 1975). These procedures randomly divide the data into two non-overlapping subsets, A and B, and then perform two steps which can be described as "hunt and test": first, sample A is used to choose one from among a collection of test statistics; next the chosen statistic is applied to sample B to produce the final test statistic. This approach is attractive because in the first stage an arbitrarily complicated procedure may be employed to *hunt* for an appropriate test. Clearly, were we to ignore the fact that our test statistic was selected from data and simply apply it to sample A, we would fail to control the type-I error, a phenomenon sometimes referred to as "double dipping"

*ricguo@umich.edu
†r.shah@statslab.cam.ac.uk

or "data snooping". However, because the data in A and B are independent, we can, in the second stage, effectively forget that the test statistic was chosen in a data-driven way, which permits straightforward calibration even when using a complicated "hunting" procedure. This strategy is particularly useful in settings with complex alternatives, as the test may be chosen to target the particular alternative from which the data appear to have arisen. This approach has been used for a variety of problems, such as testing the location of multiple samples (Cox, 1975), constructing conformal prediction intervals (Lei et al., 2018; Solari and Djordjilović, 2022), goodness-of-fit testing (Janková et al., 2020), conditional (mean) independence testing (Scheidegger et al., 2022; Lundborg et al., 2022), and conducting inference that is agnostic to the asymptotic regime (Kim and Ramdas, 2024), to list just a few. As we show in this work (see Section 4.1.2), a hunt and test approach can also be used to test for a clustering structure, i.e., for testing the null of unimodality, in high-dimensional data.

Another use of data splitting is related to nonparametric or semiparametric methods where the estimator for the parameter of interest depends on nuisance parameters that must also be estimated. To ensure proper asymptotic behaviour of the final estimator, the bias from nuisance parameter estimation needs to be controlled, and this may be achieved by employing a form of sample splitting known as cross-fitting. Here, the data are first split into folds (i.e., parts) of roughly equal size, and then estimators are computed on each fold using nuisance parameters estimated from out-of-fold data. The per-fold estimates are then combined to produce the final estimate. The independence afforded by data splitting permits the use of flexible machine learning methods to estimate these parameters, as adopted by targeted estimation or double/debiased machine learning methods. This second use of sample splitting has recently become popular in practice, although the idea has a long history; see also Newey and Robins (2018); Chernozhukov et al. (2018) and references therein.

Randomised procedures also arise naturally in settings where a null hypothesis one wishes to test may more easily be stated in terms of a reweighted distribution. Thams et al. (2023) show how a wide range of problems may be cast in this framework, including testing properties of a new policy in a contextual bandits setting, model selection after covariate shift and testing so-called generalised conditional independencies (also known as dormant independencies, or Verma constraints; see Section 4.2 for further discussion). In this work we focus on the latter, which after appropriate reweighting may be reduced to simpler independence testing problems. Thams et al. (2023) further propose to use resampling or rejection sampling to select from the original set of observations, a random subset that behaves like a sample from the reweighted distribution, to which an off-the-shelf testing procedure may then be applied.

Despite its simplicity and broad applicability, as pointed out by Cox (1975), randomised procedures have obvious drawbacks. Firstly, the extra randomness hinders replication of the analysis, an issue that is particularly concerning in view of today's "replication crisis" in many scientific disciplines (Ioannidis, 2005; Open Science Collaboration, 2015; Baker, 2016). Although one may insist that the random seed used in a randomised algorithm should be part of the replication, when the substantive result of an analysis hinges on a particular seed, considering this as replicable is questionable. Moreover, although it is sometimes argued that replicability is less of an issue when the sample size is large, as we show in Example 1 below, it remains a problem when the effect size is moderate.

A second major issue in the context of hypothesis testing is that when sampling or data-splitting is used in the construction of a test, we may expect a loss of power due to the sample not being fully utilised; see, e.g., Kim and Ramdas (2024, Theorem 2.6) for a concrete case. In the context of debiased machine learning, cross-fitting can sometimes be applied to alleviate the two drawbacks, to a certain extent. Ideally, the per-fold estimates are asymptotically independent and jointly Gaussian, thus giving an approximately Gaussian final averaged statistic. However,

in finite samples or less ideal settings, these estimates are correlated and can result in under-coverage of standard confidence intervals; we discuss this further in Section 4.3.

To illustrate the two main drawbacks mentioned above, consider the following toy example.

Example 1 (Gaussian location experiment)**.** Let $T^{(1)}, T^{(2)}, \dots$ be test statistics resulting from repeatedly applying a randomised (e.g., data splitting) procedure to a given dataset. Because $T^{(1)}, T^{(2)}, \dots$ are iid conditional on the data, *unconditionally* they are exchangeable in the sense that the joint distribution is invariant under any finite permutation (Kallenberg, 2005, §1.1). Suppose the statistic is constructed such that the location of $T^{(1)}$ captures the signal. In particular, suppose $T^{(1)} \sim \mathcal{N}(\mu, 1)$ marginally and we are interested in testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$. The corresponding α-level test is to reject H_0 when $T^{(1)} > z_\alpha := \Phi^{-1}(1-\alpha)$, which we refer to as the single-split test. We compare this to a test that aggregates the values from L realizations (e.g., random splits of the data). In this case, it is natural to aggregate by taking the average $\bar{T}_{1:L} := (T^{(1)} + \cdots + T^{(L)})/L$. The aggregated test rejects H_0 when $\bar{T}_{1:L} > \bar{t}_\alpha$, where the critical value $\bar{t}_\alpha$ is determined by the null distribution of $\bar{T}_{1:L}$.

In order to permit closed form expressions for the distribution of $\bar{T}_{1:L}$ and hence $\bar{t}_\alpha$, let us assume $T^{(1)}, T^{(2)}, \dots$ are jointly normal, i.e., follow a Gaussian process. One can show that the likelihood ratio test against any fixed $\mu > 0$ is monotone in $\bar{T}_{1:L}$. Due to the exchangeability, the distribution is parametrised by (μ, ρ), where $\rho \in [0, 1)$ is the pairwise correlation between $T^{(i)}$ and $T^{(j)}$ for every $i \neq j$. In practice, we expect $\rho > 0$.

Let $\phi_{T^{(l)}} := \mathbb{I}\{T^{(l)} > z_\alpha\}$ be the test corresponding to the l-th randomised test statistic. For each μ and ρ, we may compute the probability that the test cannot be replicated by another application of the same procedure on the same dataset: $\mathbb{P}(\phi_{T^{(1)}} \neq \phi_{T^{(2)}})$. Similarly, the probability that the aggregated test does not replicate is $\mathbb{P}(\phi_{\bar{T}_{1:L}} \neq \phi_{\bar{T}_{(L+1):2L}})$, where $\phi_{\bar{T}_{1:L}} := \mathbb{I}\{\bar{T}_{1:L} > \bar{t}_\alpha\}$. The two probabilities (see Appendix A for their expressions) are compared in Fig. 1. Note that the probability of non-replication for $\phi_{T^{(1)}}$ can be quite high when the effect size (relative to the sample size for constructing $\phi_{T^{(1)}}$) is neither too weak nor too strong. The probability is significantly reduced by aggregating $L = 200$ realizations and approaches zero as L tends to infinity.

Moreover, the aggregation also boosts the power as evidenced by Fig. 1. One can show that the power of the aggregated test is given by

$$\mathbb{E}\,\phi_{\bar{T}_{1:L}} = \Phi\left(\frac{\mu}{\sqrt{1/L + \rho(L-1)/L}} - z_\alpha\right),$$

where $L = 1$ corresponds to the power of the single-split test. For a small μ and a large L, we have

$$\mathbb{E}\,\phi_{\bar{T}^{(1)}} \approx \alpha + \phi(z_\alpha)\mu, \quad \mathbb{E}\,\phi_{\bar{T}_{1:L}} \approx \alpha + \phi(z_\alpha)\mu/\sqrt{\rho},$$

so the slope of local power is improved by a factor of $1/\sqrt{\rho}$. This improvement can be particularly significant when ρ takes a small positive value, which is not uncommon for the settings considered in the paper such as statistics resulting from two random splits of data.

From the example above, one may be tempted to conclude that to perform the aggregated test, we only need to estimate the correlation ρ (e.g., through the empirical variance computed from $T^{(1)}, T^{(2)}, \dots$). However, the joint normality assumption on the statistics $T^{(1)}, T^{(2)}, \dots$, which further implies the normality of $\bar{T}_{1:L}$, need not hold in practice. In other words, even if every $T^{(l)}$ is marginally normal or asymptotically normal, the *dependence* among the statistics need not be a normal copula even in large samples. Consider the following example due to Kim and Ramdas (2024).

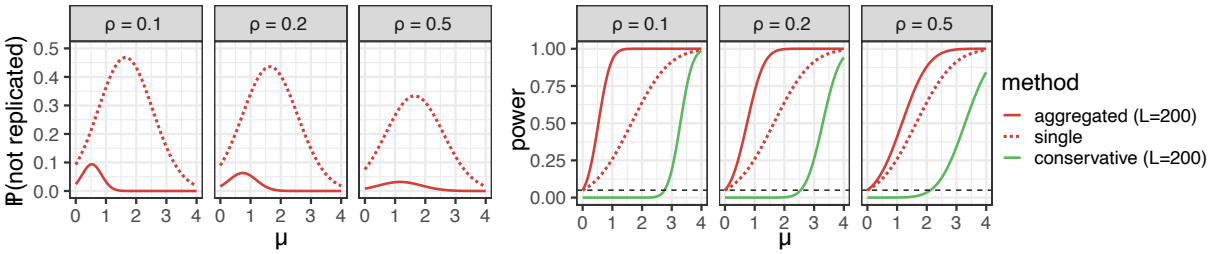

Figure 1: Probability of non-replication (left) and power (right) for the Gaussian location experiment in Example 1 ($\alpha = 0.05$, $L = 200$). The conservative aggregation rule rejects H_0 when $\bar{T}_{1:L} > 2z_\alpha$ (see Appendix B).

Example 2. Let $X_1, \ldots, X_n$ be iid random vectors in $\mathbb{R}^p$ with mean μ and covariance Σ. We are interested in testing $H_0 : \mu = \mathbf{0}$ against $H_1 : \mu \neq \mathbf{0}$. Let I_n be a random subset of $\{1, \ldots, n\}$ of size n_1 and let $n_2 := n - n_1$. Let $\widehat{\mu}_1 := n_1^{-1} \sum_{i \in I_n} X_i$ and $\widehat{\mu}_2 := n_2^{-1} \sum_{i \in I_n^c} X_i$. We expect $\widehat{\mu}_1^\top \widehat{\mu}_2 > 0$ if μ deviates from zero. Treating $\widehat{\mu}_1$ as a fixed vector we see that $\widehat{\mu}_1^\top \widehat{\mu}_2 = n_2^{-1} \sum_{i \in I_n^c} X_i^\top \widehat{\mu}_1$ is simply a sample mean computed from I_n^c. By studentising this quantity, we may obtain via a central limit theorem (under appropriate conditions), a test statistic

$$T_n(X_1, \ldots, X_n; I_n) := \frac{\sqrt{n_2} \widehat{\mu}_1^\top \widehat{\mu}_2}{\sqrt{\widehat{\mu}_1^\top \widehat{\Sigma}_2 \widehat{\mu}_1}} \to_d \mathcal{N}(0, 1) \quad \text{under } H_0,$$

as $n \to \infty$ and $n_2/n \to q \in (0, 1)$, where $\widehat{\Sigma}_2$ is the empirical covariance computed from I_n^c. We reject H_0 when T_n is large compared to $\mathcal{N}(0, 1)$. One benefit of using such a test statistic is that unlike for example a norm of the empirical mean from the whole sample, the limit distribution does not depend delicately on the asymptotic limit of p/n.

As indicated earlier, however, a disadvantage is that the approach does not fully utilise the information in the sample. One might hope that this can be alleviated by considering the cross-fitted statistic $[T_n(X_1, \ldots, X_n; I_n) + T_n(X_1, \ldots, X_n; I_n^c)]/2$. However, Kim and Ramdas (2024, Proposition A.1) showed that this does not have a normal limit, rendering calibration challenging. Further, let $I_n^{(l)}$ for $l = 1, \ldots, L$ be independent random subsets of size $\lfloor qn \rfloor$ and define $T_n^{(l)} := T_n(X_1, \ldots, X_n; I_n^{(l)})$ and the aggregated statistic $S_n := \sum_{l=1}^L T_n^{(l)}/L$. We see in Fig. 2 that its sampling distribution under $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$ is clearly non-normal even in large sample.

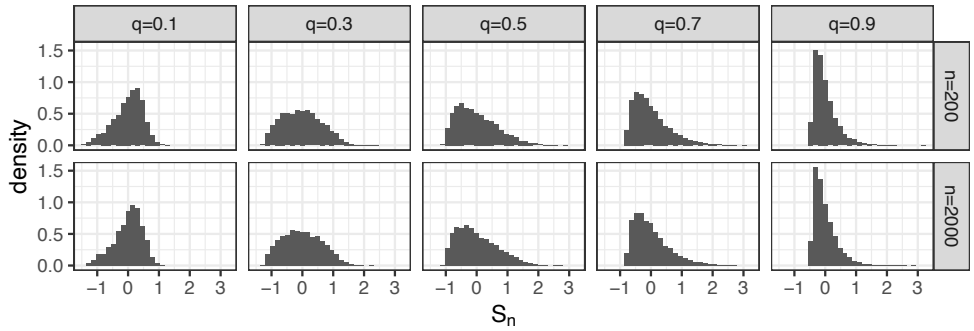

Figure 2: Non-normal null distribution of the aggregated statistic S_n in Example 2 under $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$ with $L = 200$, $p = 3$ and $\Sigma_{ij} = 2^{-|i-j|}$.

Using an aggregate S_n of the exchangeable test statistics $T_n^{(1)}, \ldots, T_n^{(L)}$ is more sensitive to departures from the null as it makes better use of the full data. In addition, because the conditional variability in S_n given the data decreases in L (e.g., like $1/L$ when S_n is the average), by taking a relatively large L, S_n is effectively derandomised; we will revisit Example 2 in Section 4.1.1 and Appendix H.1 to demonstrate the improvement from using S_n. However it should be clear from this very simple example that the main challenge in general lies in calibrating S_n. In particular, we need to handle the unknown, potentially complicated dependence among $T_n^{(1)}, \ldots, T_n^{(L)}$, which typically causes S_n to not follow any textbook distribution. To solve this problem, we develop a data-driven calibration scheme based on subsampling to obtain tests with size that under mild assumptions will approach the nominal level. Further, we will demonstrate that through the rank transform we introduce, our method is able to accurately approximate the null distribution even when data is drawn from a local alternative, and this leads to a power almost as good as an oracle procedure. Our approach is applicable to all of the randomised tests mentioned above, and as we will see, by inverting particular aggregated hypothesis tests, we can also obtain confidence intervals for cross-fitted double machine learning that can reliably deliver coverage when the standard confidence intervals cannot. Before introducing our method, we briefly summarise some existing proposals in the literature.

1.1 Existing proposals and related literature

There is a long line of work that has considered how to combine multiple test statistics. Most approaches either aim to control the type-I error under an arbitrary dependence of the test statistics, or make some specific assumptions about their dependence, such as a Gaussian copula.

Most approaches of the first type consider aggregating p-values, i.e., when $T^{(1)}$ is marginally (super-)uniformly distributed under the null. Classical examples include Bonferroni correction, i.e., taking the minimum p-value and multiplying by the total number of p-values, and the average of the p-values multiplied by 2 (Rüschendorf, 1982; Meng, 1994). Related results have been shown for quantiles (Meinshausen et al., 2009; DiCiccio et al., 2020) and generalised means (Vovk and Wang, 2020). Other aggregation rules include taking a weighted sum of Cauchy transformations (Liu and Xie, 2020), those developed through concentration inequalities (DiCiccio et al., 2020) and those involving converting p-values to the so-called e-values (Vovk and Wang, 2021), to name a few.

While these methods have the attractive guarantee of giving finite-sample valid p-values, as they necessarily must cater for the worst-case dependence, this benefit comes with the downside of conservativeness. Indeed, as pointed out by DiCiccio et al. (2020), when used with p-values produced through sample-splitting, these methods can sometimes be outperformed by using a single-split test; see Appendix H.4 for several such numerical examples from applications considered in this paper. For combining Z-statistics, the right panel of Fig. 1 illustrates a similar phenomenon in the context of Example 1 with the 'conservative' rule that rejects when $\bar{T}_{1:L} > 2z_\alpha$, where the extra factor of 2 guarantees its validity; see Theorem B.1 in the supplementary material. Indeed, as we show in Proposition B.1, directly comparing $\bar{T}_{1:L}$ to a standard normal can lead to a size of 2α, even when enforcing exchangeability of the underlying Z-statistics. This may come as a surprise given that by Jensen's inequality, $\bar{T}_{1:L}$ has variance at most 1 for example. An analogous negative result for averaging p-values, which shows that the doubling rule cannot be improved under exchangeability, is proved by Choi and Kim (2022).

In view of this, other work has considered approaches relying on asymptotic joint normality of the underlying test statistics $T_n^{(1)}, T_n^{(2)}, \ldots$; see, e.g., Romano and DiCiccio (2019, Theorem 4.1), DiCiccio (2018, Theorem 3.2), Tian et al. (2023) and Liu et al. (2022). However, as evidenced by Example 2, such a dependence structure among the test statistics is unlikely to

hold in practice, especially for data splitting.

Our work also connects more broadly to a body of literature on subsampling; see Politis et al. (1999) for a monograph on the topic. Berg et al. (2010); McMurry et al. (2012) consider subsampling for hypothesis testing and use the data to centre the estimated null distribution to improve power; this is in similar spirit to our rank transform introduced in Section 2.2, which is a more aggressive form of centring that enforces the mean and all other moments of the null. Subsampling has also been used to reduce the variance of an estimator through what is known as 'subagging' (Bühlmann and Yu, 2002). Stability selection procedures exploit this for variable selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) and can provide a form of finite-sample error control for a user-chosen variable selection method. More broadly, in the literature of resampling-based inference, ranks from bootstrap can be used to "prepivot" (Beran, 1987, 1988) a statistic to improve error control under the null, though this is somewhat different from our use of ranks to improve power; this is discussed further in Appendix G.

A large number of randomised algorithms have been developed for various testing problems to which our method is applicable. Some of these have been mentioned in the introduction under the umbrellas of "hunt and test" and reweighting. Other methods that do not explicitly fall within these categories include approximate co-sufficient sampling (Barber and Janson, 2022) for goodness-of-fit testing and several approaches for assessing variable importance nonparametrically (Cai et al., 2022; Williamson et al., 2023; Dai et al., 2022; Tansey et al., 2022).

1.2 Main contributions and organisation of the paper

We develop a general framework for hypothesis testing with an aggregated statistic $S_n := S(T_n^{(1)}, \ldots, T_n^{(L)})$, where $S(\cdot)$ is a user-specified, symmetric aggregation function such as the arithmetic mean. To fully exploit the signal in S_n, it is essential that the aggregated test is not conservative. Therefore, we construct a test from S_n whose type-I error asymptotically approaches the nominal level α, and which outperforms existing conservative aggregation rules.

To achieve this, we employ subsampling (Politis and Romano, 1994; Politis et al., 1999), a generic tool for approximating sampling distributions under minimal assumptions. However, using subsampling alone is not enough, because it not only approximates the sampling distribution under the null, but under alternatives also picks up a visible, finite-sample bias from the sampling distribution (we formalise this in a first-order asymptotic analysis of subsampling in Appendix D.3). Hence, naively comparing S_n to its subsampling critical value leads to a test with suboptimal power. To fix this crucial issue, we exploit the fact that the asymptotic null distribution of $T_n^{(1)}$ is typically known, e.g., $\mathrm{unif}(0,1)$ for a p-value or $\mathcal{N}(0,1)$ for a Z-statistic. We introduce a rank transform that we apply to the subsampled statistics to enforce the known null marginal distribution of $(T_n^{(1)}, \ldots, T_n^{(L)})$. In other words, to approximate the null distribution of $(T_n^{(1)}, \ldots, T_n^{(L)})$ and hence the aggregated S_n, we effectively combine the known *marginal* null distribution with the unknown *copula* estimated from subsampling. We demonstrate favourable performance of our method with three types of applications.

(i) "Hunt-and-test" procedures: we use our framework to develop new tests for the goodness of fit of parametric quantile regression models, and for testing unimodality in high-dimensional data. We illustrate the latter on cancer gene expression data to detect the presence of cancer subtypes.

(ii) Testing hypotheses under reweighting or "distributional shift" with resampling or rejection sampling: specifically we study testing the sharp null of no direct treatment effect in the context of a sequentially randomised trial.

(iii) Calibrating confidence intervals for cross-fitted, double/debiased machine learning estimators: in our simulations we look in particular at confidence intervals for partially linear models, though the methodology we develop is applicable more broadly.

The rest of this paper is organised as follows. In Section 2, we introduce our rank-transformed subsampling procedure. We present an aggregated, multiple-split test (Algorithm 2) and a variant (Algorithm 3) that allows for several user-specified aggregation functions and adapts to the best one. In Section 3, we study the theoretical properties of our method. We show that these algorithms give tests that asymptotically have size equal to a given significance level α. Further, we show that if the test statistic and the aggregated statistic converge uniformly under the null, our procedures inherit such uniformity in terms of type-I error control. In terms of power, we show under mild conditions that our test is as powerful as an oracle test that has access to the null distribution of the aggregated statistic. Moreover, we show that the power gap between the oracle test and our test is smaller than the gap between the oracle test and a test based on ordinary subsampling (i.e., without the rank transform) by an asymptotic order, and this leads to a significant power improvement in practice. We establish this result under general conditions that go beyond the settings where the Edgeworth expansion, the standard tool for higher-order asymptotic analysis in the literature, can be applied. In Section 4, we demonstrate our method with a variety of applications as mentioned above. Finally, we conclude with a discussion in Section 5 outlining possible directions for future research. The supplementary material contains all proofs, additional theoretical and numerical results; all the appendices can be found there. An R package `MultiSplit` implementing our method and scripts for reproducing numerical results are available from https://github.com/richardkwo/MultiSplit.

2 Method

2.1 Setup

Let $X_1, \ldots, X_n \in \mathcal{X}$ be data points drawn iid from an underlying distribution P. We are interested in testing

$$H_0 : P \in \mathcal{P}_0 \quad \text{vs} \quad H_1 : P \in \mathcal{P} \setminus \mathcal{P}_0,$$

where $\mathcal{P}$ is the set of relevant laws of X that includes both the null and the alternative. Let $T_n^{(1)}, \ldots, T_n^{(L)}$ be test statistics that can be computed from sample $X := (X_1, \ldots, X_n)$ and a piece of external randomness Ω generated by the analyst. Throughout, we assume that the random vector

$$(T_n^{(1)}, \ldots, T_n^{(L)}) \quad \text{is exchangeable} \tag{1}$$

under $(X, \Omega) \sim P^n \times P_\Omega$ for every $P \in \mathcal{P}$, where P_Ω is the distribution of Ω.

Often, such $T_n^{(1)}, \ldots, T_n^{(L)}$ are obtained by applying the same randomised procedure L times on X. That is, we have

$$T_n^{(l)} = T_n(X_1, \ldots, X_n; \Omega^{(l)}), \quad l = 1, \ldots, L. \tag{2}$$

where, without loss of generality, we assume $\Omega^{(l)} \overset{\text{iid}}{\sim} \text{unif}(0,1)$ independently from X. Formally, for $n = 1, 2, \ldots$, the randomised test is a measurable map $T_n : \mathcal{X}^n \times [0,1] \to \mathbb{R}$. For example, $\Omega^{(l)}$ can be used by T_n to realise a random data split or a sequence of $\text{unif}(0,1)$ random variables for acceptance-rejection sampling (e.g., by splitting the bits in a binary expansion). In this case, $\Omega = (\Omega^{(1)}, \ldots, \Omega^{(L)})$.

Alternatively, it can be that every $T_n^{(l)}$ is a deterministic function of $X_1, \ldots, X_n$. This can happen, for example, when there are L pre-specified ways of splitting the full sample and every

such way looks no different from any other way. We will study this case in the context of cross-fitting in Section 4.3.

Throughout, we require that under the null, $T_n^{(1)}$ converges to a known, continuous distribution F_0, such as unif$(0,1)$ or $\mathcal{N}(0,1)$. Without loss of generality, we assume H_0 is rejected for large values of $T_n^{(1)}$; other cases can be handled by redefining $T_n^{(1)}$, e.g., replacing $T_n^{(1)}$ with $|T_n^{(1)}|$ for a two-sided test, or with $1 - T_n^{(1)}$ for a p-value. To abuse the term slightly, we call the test that rejects when $T_n^{(1)} > F_0^{-1}(1-\alpha)$ the "single-split" test, even though $T_n^{(1)}$ itself may not be constructed with data splitting. Consider the aggregated, "multiple-split" statistic

$$S_n := S(T_n^{(1)}, \ldots, T_n^{(L)})$$

constructed with a symmetric, continuous aggregation function $S : \mathbb{R}^L \to \mathbb{R}$, such as the arithmetic mean or the maximum. By taking L reasonably large, we can expect that the conditional variance of S_n given $X_1, \ldots, X_n$ is small enough such that the aggregated test statistic is effectively derandomised. Note the restriction that S is symmetric is rather reasonable: it follows from the Neyman–Pearson lemma that a most powerful test (for a simple null against a simple alternative) necessarily combines the exchangeable test statistics in some symmetric fashion; see Proposition C.1 in the supplementary material.

We will make the mild assumption (see Assumption 1 and the following discussion) that S_n converges to *some* distribution G_P under the null; in practice the limit G_P is typically an unknown and often non-Gaussian continuous distribution that depends on $P \in \mathcal{P}_0$. Our aggregated test rejects H_0 for large values of S_n, and under the null aims to mimic an oracle procedure that rejects whenever S_n exceeds the unknown upper α quantile of G_P. To do this, we use subsampling to compute $\widetilde{G}_n$, an approximation to G_P, and use its quantile to determine the critical values for S_n. As mentioned earlier in Section 1.2, to have good power, however, $\widetilde{G}_n$ must continue to closely mimic the null sampling distribution of S_n even when data is generated under alternatives. For example, when $T_n^{(1)} \sim \mathcal{N}(0,1)$ under H_0 and $S_n := (T_n^{(1)} + \cdots + T_n^{(L)})/L$, under alternatives, $\widetilde{G}_n$, as expected from an oracle procedure, should maintain zero mean even when S_n takes a positive mean. This is achieved by the rank transform introduced below.

2.2 Rank-transformed subsampling

In this section, we describe our procedure when using a single aggregation function S. We first introduce some notation relating to distribution functions and then set out our subsampling scheme.

Notation Given a set of points $\{x_i\}$ on the real line, we use $\mathbb{F}_{\{x_i\}}$ to denote their empirical distribution function. For a real-valued function F, let $\|F\|_\infty := \sup_x |F(x)|$. For a distribution function F, its upper α quantile is defined as $F^{-1}(1-\alpha) := \inf\{x : F(x) \geq 1-\alpha\}$.

Subsampling Our method is based on subsampling, which ensures type-I error control under minimal assumptions. In general, subsampling cannot be replaced by the bootstrap without sacrificing the wide applicability of our method; we explain this in Appendix F. Let $m < n$ be a user-chosen subsample size. Throughout the paper, we require $m \to \infty$ and $m/n \to 0$; for the description of the algorithms and all the numerical experiments in this paper, we use $m = \lfloor n/\log n \rfloor$ (see Section 5 for a discussion). We randomly select a total of B sets of indices, each of size m, such that there is a sufficiently low degree of overlap among the sets. To do this, we first choose a positive integer (e.g., $J = 100$) and let $B := J\lfloor n/m \rfloor$. Then our collection of sets of indices $\mathcal{B} := \{(i_{1,b}, \ldots, i_{m,b}) : b = 1, \ldots, B\}$ is formed using Algorithm 1.

Algorithm 1 Generate ordered tuples

Input: Sample size n, subsample size m, positive integer J.

$\mathcal{B} \leftarrow \emptyset$.
for $j = 1, \ldots, J$ **do**
  $\pi \leftarrow$ a random permutation of $\{1, \ldots, n\}$.
  $\mathcal{B} \leftarrow \mathcal{B} \cup \big\{(\pi_1, \ldots, \pi_m), (\pi_{m+1}, \ldots, \pi_{2m}), \ldots, (\pi_{(\lfloor n/m \rfloor - 1)m+1}, \ldots, \pi_{\lfloor n/m \rfloor m})\big\}$.
end for
return $\mathcal{B}$

Note that the construction guarantees that $\mathcal{B}$ contains J collections of $\lfloor n/m \rfloor$ sets that are non-overlapping, and so statistics computed on these subsamples are independent. This will allow us to obtain guarantees for subsampling that do not rely on approximating a scheme (e.g., Politis et al., 1999, §2.4) where statistics on every possible subsample of size m are evaluated.

Let $\widehat{\mathbf{H}} = (\widehat{H}_{b,l})$ be a $B \times L$ matrix consisting of rows

$$\widehat{\mathbf{H}}_{b,\cdot} := \Big(T_m^{(1)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}}), \quad \ldots \quad , T_m^{(L)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}})\Big), \quad b = 1, \ldots, B, \tag{3}$$

i.e., by computing $(T_m^{(1)}, \ldots, T_m^{(L)})$ on each subsample listed in $\mathcal{B}$. When the statistic is a randomised test in the form of Eq. (2), the external random number is regenerated for every entry of $\widehat{\mathbf{H}}$; that is,

$$\widehat{H}_{b,l} = T_m(X_{i_{1,b}}, \ldots, X_{i_{m,b}}; \Omega^{(b,l)}), \quad \Omega^{(b,l)} \overset{\text{iid}}{\sim} \text{unif}(0,1), \quad b = 1, \ldots, B,\ l = 1, \ldots, L.$$

Note that although we have arranged the subsampled test statistics into a matrix, entries in the same column but different rows do not correspond directly to one another, that is, $\widehat{H}_{b,l}$ is no more related to $\widehat{H}_{b',l}$ than $\widehat{H}_{b',l'}$ for $b' \neq b$ and $l' \neq l$.

Now if we were to apply the aggregation function S to each row of $\mathbf{H}$, we would obtain

$$\widehat{S}_b := S(\widehat{H}_{b,1}, \ldots, \widehat{H}_{b,L}), \quad b = 1, \ldots, B,$$

whose empirical distribution function $\widehat{G}_n(x) := \mathbb{F}_{\{\widehat{S}_b\}}(x)$ is the natural subsampling estimate for $G_P(x)$. By the standard consistency result of subsampling (Politis et al., 1999; see also Appendix C.1), we have $\|\widehat{G}_n - G_P\|_\infty \to_p 0$ under $P \in \mathcal{P}_0$. However, directly using $\widehat{G}_n$ to construct the test is suboptimal because under a sequence of local alternatives, $\widehat{G}_n$ contains an upward bias from sampling under alternative. Although such a bias may vanish asymptotically, the rate at which it vanishes can be rather slow and this can severely reduce power; this is illustrated in the bottom left panel of Fig. 3. In the supplementary material, we formalise this point in Theorem D.2 and numerically demonstrate the bias in Appendix D.5. Therefore, instead of using $\widehat{G}_n$ to calibrate our test statistic S_n, we perform the rank transform introduced below.

Rank transform Using exchangeability, we can pool the entries of $\widehat{\mathbf{H}}$ and let $\mathbb{F}_{\widehat{\mathbf{H}}}$ be the resulting empirical distribution function. With this, we form a rank-transformed version of $\widehat{\mathbf{H}}$, denoted by $\widetilde{\mathbf{H}} = (\widetilde{H}_{b,l})$, filled with entries

$$\begin{aligned} \widetilde{H}_{b,l} &:= F_0^{-1}\Big(\mathbb{F}_{\widehat{\mathbf{H}}}(\widehat{H}_{b,l}) - 1/(2B\,L)\Big) \\ &= F_0^{-1}\left(\frac{(\text{rank of } \widehat{H}_{b,l} \text{ among entries in } \widehat{\mathbf{H}}) - 1/2}{B\,L}\right), \end{aligned} \tag{4}$$

where the subtraction of $1/2$ from the ranks is simply a finite sample correction to prevent infinity being produced when applying F_0^{-1}. We then compute the aggregated statistics

$$\widetilde{S}_b := S(\widetilde{H}_{b,1}, \ldots, \widetilde{H}_{b,L}), \quad b = 1, \ldots, B$$

and their resulting empirical distribution function $\widetilde{G}_n := \mathbb{F}_{\{\widetilde{S}_b\}}(x)$, which we then use to determine the critical value $\widetilde{G}_n^{-1}(1-\alpha)$ for S_n. The full procedure is given in Algorithm 2.

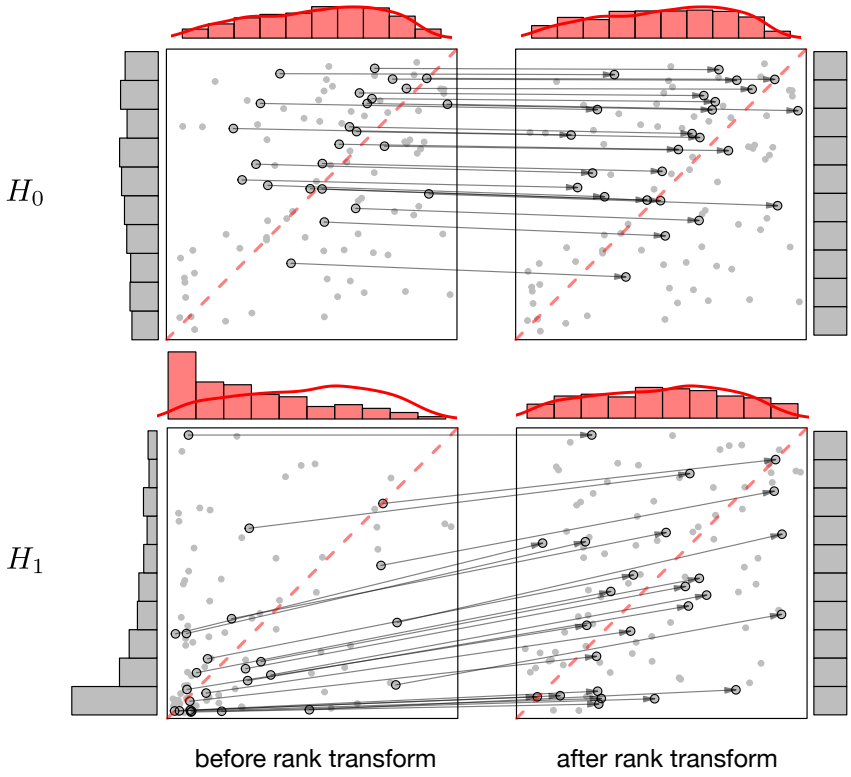

Figure 3: Illustration of the rank transform. Here $L = 2$ and the rows of $\widehat{\mathbf{H}}$ and $\widetilde{\mathbf{H}}$ are plotted as points in two dimensions in the left and right panels respectively. Arrows indicate for certain points their image after the rank transform. We consider aggregation function S as the arithmetic mean which may be visualised as projection onto the red dashed line. Top (red) histograms: distributions $\widehat{G}_n$ and $\widetilde{G}_n$ (curve: null density of S_n); side (grey) histograms: marginal distributions $\mathbb{F}_{\widehat{\mathbf{H}}}$ and $\mathbb{F}_{\widetilde{\mathbf{H}}}$ ($F_0 = \text{unif}(0,1)$).

We provide some intuition on why the rank transform works. Consider first the null case where $P \in \mathcal{P}_0$. As $n \to \infty$ (and hence $n/m \to \infty$, $B \to \infty$), by consistency of subsampling and exchangeability of $T_m^{(1)}, \ldots, T_m^{(L)}$ (so they share the same marginal distribution), we expect $\mathbb{F}_{\widehat{\mathbf{H}}}(\cdot) - 1/(2BL) \approx F_0(\cdot)$ in Eq. (4). Therefore, $\widetilde{\mathbf{H}} \approx \widehat{\mathbf{H}}$ under the null. Because $\widetilde{G}_n$ is computed from $\widetilde{\mathbf{H}}$ in the same way as $\widehat{G}_n$ is computed from $\widehat{\mathbf{H}}$, we can expect that $\widetilde{G}_n \approx \widehat{G}_n \approx G_P$ with high probability under the null. This is formalised in Theorem 1 and illustrated in the top panel of Fig. 3, from which we see that under the null the rank-transform leaves the points almost unchanged, and both $\widehat{G}_n$ and $\widetilde{G}_n$ well-approximate the sampling distribution of S_n.

Under an alternative $P \in \mathcal{P} \setminus \mathcal{P}_0$, we expect that $\mathbb{F}_{\widehat{\mathbf{H}}}(\cdot) - 1/(2BL) \approx F_P(\cdot)$, with F_P the distribution function of the test statistic $T_m^{(1)}$ corresponding to the subsample size m. Thus, from Eq. (4) we have

$$\widetilde{H}_{b,l} = F_0^{-1}\left(\mathbb{F}_{\widehat{\mathbf{H}}}(\widehat{H}_{b,l}) - 1/(2BL)\right) \approx F_0^{-1}\left(F_P(\widehat{H}_{b,l})\right) \approx F_0^{-1}\left(U_{b,l}\right)$$

for some $U_{b,l} \sim \text{unif}(0,1)$. In this way, the rank transform enforces the marginal distribution of $\widetilde{H}_{b,l}$ to be F_0, the asymptotic null distribution of $\widehat{H}_{b,l}$, as we can observe from the side histograms in Fig. 3. The dependency among the $\widehat{H}_{b,1}, \ldots, \widehat{H}_{b,L}$ in contrast is left unchanged. However, particularly under local alternatives where $T_n^{(1)}$ contains just enough information to detect deviation from the null, we would certainly expect the *dependency* among test statistics constructed from smaller subsampled data to be indistinguishable from that under the null; we will formalise this notion in Definition 1. In sum then, $\widetilde{H}_{b,1}, \ldots, \widetilde{H}_{b,L}$ should continue approximating the null distribution of $T_n^{(1)}, \ldots, T_n^{(L)}$, as we desire. Further, such an approximation should be more accurate than directly using $\widehat{H}_{b,1}, \ldots, \widehat{H}_{b,L}$, because any bias stemming from the difference in the marginal distribution has already been removed. Indeed, this underlies the effectiveness of the rank transform in restoring null-like behaviour under the alternative, as demonstrated in the bottom panel of Fig. 3 and formalised by our theory in Section 3.2.

Algorithm 2 Aggregated multiple-split test

Input: Data $(X_1, \ldots, X_n)$, exchangeable single-split test statistics $(T_n^{(1)}, \ldots, T_n^{(L)})$, asymptotic null distribution function F_0, aggregation function S, significance level $\alpha \in (0,1)$, positive integer J.

$m \leftarrow \lfloor n/\log n \rfloor$, $B \leftarrow J\lfloor n/m \rfloor$.
Run Algorithm 1 to obtain $\mathcal{B} = \{(i_{1,b}, \ldots, i_{m,b}) : b = 1, \ldots B\}$.
Initialise $B \times L$ matrices $\widehat{\mathbf{H}}$, $\widetilde{\mathbf{H}}$ and B-dimensional vector $\widetilde{S}$.
for $b = 1, \ldots, B$ **do**
    $\widehat{\mathbf{H}}_{b,\cdot} \leftarrow (T_m^{(1)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}}), \ldots, T_m^{(L)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}}))$
end for
for $b = 1, \ldots, B$ **do**
    for $l = 1, \ldots, L$ **do**
        $\widetilde{H}_{b,l} \leftarrow F_0^{-1}\left((\{\text{rank of } \widehat{H}_{b,l} \text{ in } \widehat{\mathbf{H}}\} - 1/2)/B\,L\right)$
    end for
    $\widetilde{S}_b \leftarrow S(\widetilde{H}_{b,1}, \ldots, \widetilde{H}_{b,L})$
end for
$\widetilde{G}_n \leftarrow \mathbb{F}_{\{\widetilde{S}_b\}}$
Compute $S_n \leftarrow S(T_n^{(1)}, \ldots, T_n^{(L)})$ from $X_1, \ldots, X_n$.
Reject H_0 if $S_n > \widetilde{G}_n^{-1}(1-\alpha)$ and report p-value $1 - \widetilde{G}_n(S_n)$.

2.3 Adapting to the best aggregation function

We can further improve power by choosing a good aggregation function S. The performance of an aggregation function, however, depends on the joint behaviour of the single-split tests under the alternative of interest, which is usually unknown. For example, we expect $S = (T_n^{(1)} + \cdots + T_n^{(L)})/L$ to work particularly well if most of the single-split statistics are large under the alternative; in contrast, $S = \max(T_n^{(1)}, \ldots, T_n^{(L)})$ should perform better if only a few of them are large under the alternative. This motivates us to allow the user to specify multiple candidate aggregation functions $S^1, \ldots, S^W$, which are expected to accommodate different cases. In Algorithm 3, we present a variant of our procedure that aims to adapt to the best aggregation function among $S^1, \ldots, S^W$.

The algorithm rejects for large values of

$$R_n := \max\left(\widetilde{G}_n^1(S_n^1), \ldots, \widetilde{G}_n^W(S_n^W)\right), \tag{5}$$

where $\widetilde{G}_n^w$ and S_n^w respectively are the counterparts of $\widetilde{G}_n$ and S_n in Algorithm 2 but relate to the w-th aggregation function. The quantity R_n in Eq. (5) is therefore one minus the minimum p-value corresponding to each of the aggregation functions. Thus if any one of the aggregation functions yields good power, we should expect R_n to be large: in this way, the test statistic aims to achieve power close to that of the best S^w under consideration.

We could calibrate R_n using a Bonferroni correction, but this would give a conservative test potentially sacrificing any power we might have gained in using multiple aggregation functions. Instead, we can *reuse* our subsampling aggregate statistics $\widetilde{S}_b^w$ to approximate the sampling distribution of R_n under the null; the subsampled versions of R_n used for this are computed in lines 9–11 of Algorithm 3. The advantage of this approach is that it properly takes account of the dependence among $\widetilde{G}_n^1(S_n^1), \ldots, \widetilde{G}_n^W(S_n^W)$ involved in the construction of R_n. As a consequence, the resulting test has asymptotic size equal to its prescribed level (Theorem 2), and so in this sense no power is lost.

Algorithm 3 Aggregated multiple-split test that adapts to the best aggregation function

Input: As in Algorithm 2 but with the single aggregation function S replaced by a collection $S^1, \ldots, S^W$.

1: Run up to line 10 of Algorithm 2 to obtain $\widetilde{H}_{b,l}$ for $b = 1, \ldots, B$ and $l = 1, \ldots, L$.
2: Initialise B-dimensional vectors $\widetilde{R}$, $\widetilde{S}^1, \ldots, \widetilde{S}^W$.
3: **for** $w = 1, \ldots, W$ **do**
4: **for** $b = 1, \ldots, B$ **do**
5: $\widetilde{S}_b^w \leftarrow S^w(\widetilde{H}_{b,1}, \ldots, \widetilde{H}_{b,L})$
6: **end for**
7: $\widetilde{G}_n^w \leftarrow \mathbb{F}_{\{\widetilde{S}_b^w : b=1,\ldots,B\}}$
8: **end for**
9: **for** $b = 1, \ldots, B$ **do**
10: $\widetilde{R}_b \leftarrow \max\left(\widetilde{G}_n^1(\widetilde{S}_b^1), \ldots, \widetilde{G}_n^W(\widetilde{S}_b^W)\right)$
11: **end for**
12: $\widetilde{Q}_n \leftarrow \mathbb{F}_{\{\widetilde{R}_b\}}$
13: **for** $w = 1, \ldots, W$ **do**
14: Compute $S_n^w \leftarrow S^w(T_n^{(1)}, \ldots, T_n^{(L)})$ from $X_1, \ldots, X_n$.
15: **end for**
16: $R_n \leftarrow \max\left(\widetilde{G}_n^1(S_n^1), \ldots, \widetilde{G}_n^W(S_n^W)\right)$
17: Reject H_0 if $R_n > \widetilde{Q}_n^{-1}(1-\alpha)$ and report p-value $1 - \widetilde{Q}_n(R_n)$.

3 Theory

3.1 Behaviour under the null

In this section, we establish that our algorithms lead to asymptotically level α tests under a set of mild assumptions, which ensure the consistency of subsampling and the validity of rank transform.

Condition 1 (Asymptotic pivotal null)**.** *For every $P \in \mathcal{P}_0$, under $(X, \Omega) \sim P^n \times P_\Omega$, as $n \to \infty$ it holds that*

$$T_n^{(1)} \to_d F_0,$$

where F_0 is a known, continuous distribution function.

Note that when $T_n^{(l)}$ is a randomised test of the form in Eq. (2), a sufficient condition for the above is that for every fixed $\omega \in (0,1)$ and every $P \in \mathcal{P}_0$, it holds that $T_n(X_1, \ldots, X_n; \omega) \to_d F_0$.

We study the tests constructed from rank-transformed subsampling under the null for two leading cases, when (i) F_0 is unif$(0,1)$ (one minus p-value), (ii) $F_0 = \Phi$ (Z-statistic). For other null distributions, probability integral transform and its inverse can be applied to convert the statistic to one of these cases.

First, we show that Algorithm 2 is an asymptotic level α test under mild assumptions. Further, when $T_n^{(1)}$ and S_n converge uniformly to their limiting distributions over the null, under a mild finite density condition, we show that the test also controls size below α *uniformly* over the null. It can be argued that uniform asymptotic size control, as opposed to pointwise asymptotic size control, is more relevant to practice because in contrast to the latter, it ensures that the sample size required to control the actual type-I error below, say 0.051, does not depend on the underlying $P \in \mathcal{P}_0$; see Lehmann and Romano (2005, §11.1). As uniform size control involves consideration of the behaviour of random variables under different P (rather than a single P in pointwise asymptotics), in the below, we will use a subscript in $\mathbb{P}_P(\cdot)$ to denote that $(X, \Omega) \sim P^n \times P_\Omega$.

Recall that Algorithm 2 rejects H_0 whenever S_n exceeds the upper α quantile of the rank-transformed subsampling distribution $\widetilde{G}_n$. Control of the size is therefore intimately linked to the behaviour of $\widetilde{G}_n$, for which we will require the following. Note that in the below, all densities are with respect to the Lebesgue measure.

Condition 2 (Lipschitz aggregation)**.** *The aggregation function S is Lipschitz continuous in $\|\cdot\|_\infty$ with Lipschitz constant 1.*

Examples include $S_n = (T_n^{(1)} + \cdots + T_n^{(L)})/L$ and $S_n = \max(T_n^{(1)}, \ldots, T_n^{(L)})$. Note that given any Lipschitz continuous S, by scaling, the Lipschitz constant 1 above is not a restriction; however such a scaling affects the value of $g_{P,\max}$ in Assumption 1 below, which we require to be finite.

Assumption 1 (Stability of S_n)**.** *For every $P \in \mathcal{P}_0$, under $(X, \Omega) \sim P^n \times P_\Omega$, it holds that*

$$S_n \to_d G_P,$$

where G_P is a continuous distribution that can depend on P and can be unknown. Further, G_P has a density function g_P such that $g_{P,\max} := \sup_x g_P(x) < \infty$.

If $(T_n^{(1)}, \ldots, T_n^{(L)})$ has a limiting joint distribution under every $P \in \mathcal{P}_0$, then the first part of Assumption 1 holds by definition of S_n. Given that $(T_n^{(1)}, \ldots, T_n^{(L)})$ is exchangeable with a limit marginal law (Condition 1), we typically expect the joint distribution to be stable as well (see Fig. 2 for an example). In fact, for a given P, the sequence $(S_n)_{n=1}^\infty$ is uniformly tight, and so by Prohorov's theorem (van der Vaart, 2000, Thm. 2.4), there always exists a subsequence that converges in distribution; see Proposition C.3 in the supplementary material. Thus the only way the stability assumption can fail is when the copula $(F_{n,P}(T_n^{(1)}), \ldots, F_{n,P}(T_n^{(L)}))$ in some sense "oscillates" as $n \to \infty$, where $F_{n,P}$ is the distribution function of $T_n^{(1)}$.

Below we present results on both pointwise and uniform asymptotic size control for the test in Algorithm 2.

Theorem 1 (Validity of Algorithm 2)**.** *Let $(X, \Omega) \sim P^n \times P_\Omega$ for $P \in \mathcal{P}_0$. Suppose $T_n^{(1)}, \ldots, T_n^{(L)}$ are exchangeable and Condition 1 holds with $F_0 = \text{unif}(0,1)$ or $F_0 = \mathcal{N}(0,1)$. Suppose S is chosen such that Condition 2 holds. Then for all $\alpha \in (0,1)$, the following hold:*

(i) Under Assumption 1, the test in Algorithm 2 is pointwise asymptotically level α.

(ii) Suppose $T_n^{(1)}$ *and* S_n *converge to their respective limit distributions uniformly over the null, i.e., for every* $x \in \mathbb{R}$

$$\sup_{P\in\mathcal{P}_0} |F_{n,P}(x) - F_0(x)| \to 0, \quad \sup_{P\in\mathcal{P}_0} |G_{n,P}(x) - G_P(x)| \to 0,$$

where $F_{n,P}$ *and* $G_{n,P}$ *respectively denote the distribution function of* $T_n^{(1)}$ *and* S_n*. Also, suppose* $\sup_{P\in\mathcal{P}_0} g_{\max,P} < \infty$*. Then, the test in Algorithm 2 is uniformly asymptotically level* α*, i.e.,*

$$\sup_{P\in\mathcal{P}_0} \left| \mathbb{P}_P \left\{ S_n > \widetilde{G}_n^{-1}(1-\alpha) \right\} - \alpha \right| \to 0.$$

Now we establish similar results for the adaptive test in Algorithm 3, under the following joint stability assumption on the chosen aggregation functions $(S^1, \ldots, S^W)$.

Assumption 2 (Joint stability of multiple aggregation functions)**.** *For every* $P \in \mathcal{P}_0$*, under* $(X, \Omega) \sim P^n \times P_\Omega$*, it holds that*

$$(S_n^1, \ldots, S_n^W) \to_d (S^1, \ldots, S^W),$$

where every S^w *has a continuous distribution function* G_P^w *that can depend on* P*. Further, suppose* G_P^w *permits a density* g_P^w *such that* $g_{P,\max}^w := \sup_x g_P^w(x) < \infty$ *for* $w = 1, \ldots, W$*.*

Theorem 2 (Validity of Algorithm 3)**.** *Let* $(X, \Omega) \sim P^n \times P_\Omega$ *for* $P \in \mathcal{P}_0$*. Suppose* $T_n^{(1)}, \ldots, T_n^{(L)}$ *are exchangeable and Condition 1 holds with* $F_0 = \text{unif}(0,1)$ *or* $F_0 = \mathcal{N}(0,1)$*. Also, suppose Condition 2 holds for every* S^w $(w = 1, \ldots, W)$*. Then for every* $\alpha \in (0,1)$ *the following hold:*

(i) Under Assumption 2, the test in Algorithm 3 has pointwise asymptotic level α*.*

(ii) Suppose $T_n^{(1)}, S_n^1, \ldots, S_n^W$ *and* $\max\{G_P^1(S_n^1), \ldots, G_P^W(S_n^W)\}$ *converge to their respective limit distributions uniformly over the null, i.e., for every* $x \in \mathbb{R}$ *and each* $w = 1, \ldots, W$*,*

$$\sup_{P\in\mathcal{P}_0} |F_{n,P}(x) - F_0(x)| \to 0, \quad \sup_{P\in\mathcal{P}_0} |G_{n,P}^w(x) - G_P^w(x)| \to 0, \quad \sup_{P\in\mathcal{P}_0} |Q_{n,P}(x) - Q_P(x)| \to 0,$$

where $Q_{n,P}$ *is the distribution function of* $\max\{G_P^1(S_n^1), \ldots, G_P^W(S_n^W)\}$*, and* $G_{n,P}^w$ *is the distribution function of* S_n^w*. Also, suppose* $\max_w \sup_{P\in\mathcal{P}_0} g_{\max,P}^w < \infty$*. Then, the test in Algorithm 3 is uniformly asymptotically level* α*, i.e.,*

$$\sup_{P\in\mathcal{P}_0} \left| \mathbb{P}_P \left\{ R_n > \widetilde{Q}_n^{-1}(1-\alpha) \right\} - \alpha \right| \to 0.$$

3.2 Power

In this section, we study the power of rank-transformed subsampling and establish its advantage over ordinary subsampling, i.e., subsampling without the rank transform. We will analyse power under a sequence of local alternatives that converge "in copula" to a null case — such a null is typically also the limit that the sequence of local alternatives weakly converges to. For any sequence $P_n \in \mathcal{P}$, let F_{m,P_n} be the distribution function of $T_m^{(1)}(X_1, \ldots, X_m)$ under $(X, \Omega) \sim P_n^m \times P_\Omega$ and let U_m denote the copula:

$$U_m := (U_m^{(1)}, \ldots, U_m^{(L)}) := \left(F_{m,P_n}(T_m^{(1)}), \ldots, F_{m,P_n}(T_m^{(L)}) \right). \tag{6}$$

Definition 1 (Convergence in copula)**.** Let U_m be the copula of $(T_m^{(1)}, \ldots, T_m^{(L)})$ under $P_n^m \times P_\Omega$ given by Eq. (6). We say P_n converges in copula to P_0, if there exists some $P_0 \in \mathcal{P}_0$ such that under $P_0^n \times P_\Omega$, $(T_n^{(1)}, \ldots, T_n^{(L)})$ converges to a limit distribution with copula $C = (C_1, \ldots, C_L)$ satisfying $U_m \to_d C$.

Convergence in copula is a rather weak notion of convergence for two reasons. Firstly, it involves the lower sample size m (recall $m = o(n)$) rather than n. Consider a sequence of alternatives P_n that are only just distinguishable from the null at sample size n. At sample size m, the null and P_n should be indistinguishable; that is, the behaviour of the entire vector of test statistics $(T_m^{(1)}, \ldots, T_m^{(L)})$ under $P_n^m \times P_\Omega$ and $P_0^m \times P_\Omega$ should be asymptotically identical, and in particular convergence in copula would hold. Secondly, Definition 1 is completely insensitive to the marginal distribution of the test statistics, and so in fact we can even expect a stronger version of the convergence above to hold with m replaced by n. In particular, when P_n is a sequence of local alternatives that converges to $P_0 \in \mathcal{P}_0$ in a way such that under $P_0^n \times P_\Omega$, $(T_n^{(1)}, \ldots, T_n^{(L)})$ and the log-likelihood ratio $\log(\mathrm{d}P_n^n / \mathrm{d}P_0^n)$ jointly converge to a normal limit, then by Le Cam's third lemma (van der Vaart, 2000, Example 6.7), $(T_n^{(1)}, \ldots, T_n^{(L)})$ under $P_n^n \times P_\Omega$ must also converge to a normal limit with the same covariance, and so the same copula, as its null limit.

Our next result shows that under a sequence of local alternatives that converge in copula to a null $P_0 \in \mathcal{P}_0$, the test in Algorithm 2 asymptotically has the same critical value and hence achieves the same power as an oracle test that has access to the asymptotic null distribution of S_n under P_0. In stating Theorems 3 and 4 below, we use G_{P_0} to denote the limit null distribution function of $S_n := S(T_n^{(1)}, \ldots, T_n^{(L)})$ under $P_0^n \times P_\Omega$.

Theorem 3 (Zeroth-order behaviour of rank-transformed subsampling)**.** *Let $(T_n^{(1)}, \ldots, T_n^{(L)})$ be exchangeable. Suppose Condition 1 holds with $F_0 = \mathrm{unif}(0,1)$ or $F_0 = \mathcal{N}(0,1)$ and S is chosen to satisfy Condition 2. Consider a sequence $P_n \in \mathcal{P}$ that converges in copula to some $P_0 \in \mathcal{P}_0$. Under Assumption 1, for any $\alpha \in (0,1)$, we have*

$$\widetilde{G}_n^{-1}(1-\alpha) \to_p G_{P_0}^{-1}(1-\alpha),$$

where $\widetilde{G}_n$ is computed using Algorithm 2.

In fact, when P_n is a sequence of local alternatives that converges to $P_0 \in \mathcal{P}_0$ such that P_n^n is contiguous to P_0^n (i.e., absolutely continuous asymptotically; see van der Vaart, 2000, Ch. 6), we also expect $\widehat{G}_n^{-1}(1-\alpha) \to_p G_{P_0}^{-1}(1-\alpha)$ for $\widehat{G}_n$ obtained from ordinary subsampling (Politis et al., 1999, Theorem 2.6.1). Hence, to capture the power improvement from the rank transform, we need a finer analysis. To this end, we characterise the first-order asymptotic behaviour of rank-transformed subsampling in the next theorem, of which the full statement can be found in Appendix D.2. In the below, $\mathrm{d_{TV}}(X, Y)$ denotes the total variation distance between X's distribution and Y's distribution. The regularity condition would require S to be non-decreasing in each coordinate, a condition that holds for $S = \mathrm{avg}$, $S = \max$ and other reasonable choices.

Theorem 4 (First-order behaviour of rank-transformed subsampling)**.** *Suppose Condition 1 holds and $(T_n^{(1)}, \ldots, T_n^{(L)})$ is exchangeable. Consider a sequence $P_n \in \mathcal{P}$ that converges in copula to some $P_0 \in \mathcal{P}_0$ in the sense of Definition 1 such that*

$$\mathrm{d}_{TV}(U_m, C) = o(\sqrt{m/n}).$$

Suppose the distribution of C is absolutely continuous with respect to the Lebesgue measure. Let $\widetilde{G}_n$ denote the rank-transformed subsampling distribution function obtained with a variant of

Algorithm 2 that uses two independent copies of the data under $(X, X', \Omega) \sim P_n^n \times P_n^n \times P_\Omega$ *(see Appendix D.2).*

Suppose Assumption 1 holds and fix $\alpha \in (0,1)$ *such that the density* G'_{P_0} *is strictly positive and continuous in a neighbourhood of* $G_{P_0}^{-1}(1-\alpha)$*. Then, under regularity conditions posed on the copula and* S *(see Appendix D.2), for any* $M > 0$*, we have*

$$\mathbb{E}\left[-M \vee \sqrt{n/m}\,\left(\widetilde{G}_n^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha)\right) \wedge M\right] \to 0.$$

Further, let G_{n,P_n} *be the distribution function of* $S(T_n^{(1)}, \ldots, T_n^{(L)})$ *under* $P_n^n \times P_\Omega$*. Suppose additionally that* $\|G_{n,P_n} - G_{alt}\|_\infty = o(\sqrt{m/n})$ *holds for some distribution function* G_{alt} *that is differentiable at* $G_{P_0}^{-1}(1-\alpha)$*. Then, for any* $M > 0$*, we also have*

$$\mathbb{E}\left[-M \vee \sqrt{n/m}\,\left(G_{n,P_n}(\widetilde{G}_n^{-1}(1-\alpha)) - G_{n,P_n}(G_{P_0}^{-1}(1-\alpha))\right) \wedge M\right] \to 0.$$

To interpret this result, let us take M to be a large constant and choose $m = \lfloor n/\log n \rfloor$. Then, the first statement above says that up to the first order (with scaling factor $\sqrt{\log n}$), the rank-transformed subsampling delivers an approximation to the oracle critical value that is asymptotically unbiased. In contrast, in Theorem D.2 (see the supplementary material) we show that the ordinary subsampling approximation to the oracle critical value is biased upwards, and typically the bias grows with the effect size of the alternative. This formalises our observation from Fig. 3: under H_1, before applying the rank transform, subsampling is biased towards the alternative sampling distribution and we can see a clear discrepancy between the subsampling distribution (red histogram) and the desired null distribution (red curve). Along a sequence of contiguous local alternatives, although this discrepancy vanishes asymptotically, this occurs rather slowly ($1/\sqrt{\log n} \approx 1/4$ when $n = 10^7$) and can result in a significant loss of power in practice.

Recall that G_{n,P_n} denotes the distribution function of S_n under $P_n^n \times P_\Omega$. The power of the oracle test is $\mathbb{P}(S_n > G_{P_0}^{-1}(1-\alpha)) = 1 - G_{n,P_n}(G_{P_0}^{-1}(1-\alpha))$. If we ignore the dependence between the estimated critical value and the test statistic, the power of our test can be written similarly as

$$\mathbb{P}(S_n > \widetilde{G}_n^{-1}(1-\alpha)) \approx 1 - \mathbb{E}\, G_{n,P_n}(\widetilde{G}_n^{-1}(1-\alpha)).$$

Consequently, the second statement of Theorem 4 implies

$$\text{pow(rank-transformed subsampling)} \approx \text{pow(oracle)} - o(1/\sqrt{\log n}). \tag{7}$$

Meanwhile, with G_{m,P_n} denoting the distribution function of S_m under $P_n^m \times P_\Omega$, suppose $(n/m)^\beta(G_{m,P_n}^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha))$ converges to $\tau > 0$ that measures the effect size. Then, in contrast to the above, Theorem D.2 implies

$$\text{pow(ordinary subsampling)} \approx \text{pow(oracle)} - \kappa_{\alpha,\tau}\,\tau\,/(\log n)^\beta, \quad 0 < \beta \le 1/2 \tag{8}$$

for some $\kappa_{\alpha,\tau} > 0$. Typically, we expect $\kappa_{\alpha,\tau}\,\tau$ to grow as τ increases from zero up to a certain value; see Appendix D.5 for a concrete example.

For example, for testing a hypothesis of a regular parameter, under a $\sqrt{n}$ local alternative, we may expect $G_{n,P_n}^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha) \to \tau$ and hence $\sqrt{m/n}(G_{m,P_n}^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha)) \to \tau$. We illustrate such a case using numerical results for Example 2, which considers testing the mean of a random vector. The details of the simulation study will be described in Section 4.1.1. For now, let us focus on Fig. 4, which shows the power of several aggregated, multiple-split tests against the effect size of local alternatives. Indeed, we can see that the rank-transform

(Algorithm 2) has a clear power advantage over ordinary ('`no rank`') subsampling, and this advantage enlarges with the effect size, exactly as we expect from comparing Eqs. (7) and (8) under $\beta = 1/2$. Meanwhile, in every setting, the power of the rank-transformed subsampling closely tracks that of the oracle test regardless of the effect size, confirming Eq. (7).

We prove Theorem 4 in Appendix D using the functional delta method, where a major technical challenge is a certain Hadamard differentiability we establish for handling the errors introduced by the rank transform. For technical reasons, Theorem 4 is proved for a variant of Algorithm 2 that has access to two independent copies of the data, but we expect a similar result to hold for the original algorithm as well; see Appendix D.5 for a concrete example with supporting numerical results. Further, in Appendix D.4, we show that when C follows a Gaussian copula, the regularity conditions in Theorem 4 are satisfied by choices of F_0 and S considered in this paper.

4 Applications

We illustrate our method with three types of applications. First, we study data-splitting, hunt-and-test procedures: specifically, we revisit Example 2 for testing the zero mean of a high-dimensional random vector; we develop a new test for unimodality in high dimensions; and introduce a simple, flexible approach for goodness-of-fit testing of parametric regression models such as parametric quantile regression. Next, we consider using the data from a distribution P to test a property of a different distribution Q, where Q is related to P through reweighting. We study this in the context of causal inference, where P is the observational distribution and Q is an intervened distribution. Finally, we study the inference of cross-fitted, double/debiased machine learning estimators. We show that the cross-fold dependence in these estimators, though often argued to be asymptotically negligible in standard well-specified, low-dimensional settings, can, in finite sample or under misspecification, lead to under-coverage of confidence intervals (see Jiang et al., 2022 for a high-dimensional setting not considered in this paper where this issue also arises). We present an alternative construction of confidence intervals using our method that captures such dependence and restores the desired coverage. These confidence intervals can even maintain coverage when the model for a nuisance parameter in doubly robust estimation is misspecified, which we illustrate in Appendix E.

For each application, we present numerical results to illustrate and benchmark the new methods we develop. Additional numerical results, including the performance of conservative aggregation rules mentioned in Section 1.1, can be found in Appendix H.

4.1 Hunt and test

In this section, we consider testing a hypothesis that can be expressed as a conjunction of simpler hypotheses

$$H_0 = \bigcap_{\delta \in \mathcal{D}} H_0(\delta), \tag{9}$$

where we already have an off-the-shelf test for each $H_0(\delta)$. As explained in the introduction, such null hypotheses are amenable to a hunt-and-test approach that employs data splitting, where one part of the data is used to find an appropriate $\widehat{\delta}$ and the remaining data is used to test $H_0(\widehat{\delta})$. Here is another perspective due to Moran (1973). Consider testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1, \tag{10}$$

where Θ_1 does not contain θ_0. The alternative parameter space Θ_1 might be so large or heterogeneous that a reasonable test for H_0 only has power against certain alternatives in Θ_1. Again,

we can split our data and perform hunt and test: use the first part to estimate $\widehat{\theta}_1 \in \Theta_1$ and then use the second part to test $\theta \in \Theta_0$ versus $\theta = \widehat{\theta}_1$.

4.1.1 Revisiting Example 2

In Example 2, we considered a hunt-and-test approach for testing $H_0 : \mu = \mathbf{0}$ versus $H_1 : \mu \neq \mathbf{0}$ with iid random vectors. Clearly, the problem is an instance of Eq. (10); it can also be viewed as an instance of Eq. (9), namely $H_0 = \cap_{\delta \in \mathbb{R}^p} \{\mu : \mu^\top \delta = 0\}$. Fig. 4 compares the performance of various tests based on aggregated $S_n = (T_n^{(1)} + \cdots + T_n^{(L)})/L$ with $L = 200$ and the single-split test based on $T_n^{(1)}$ alone. In our simulation, we draw $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma)$ where $\Sigma \in \mathbb{R}^{3\times 3}$ has entries given by $\Sigma_{ij} = 2^{-|i-j|}$ and $\mu = \tau n^{-1/2} v_1$, where v_1 is the normalised principal eigenvector of Σ. When $\tau = 0$, we see that our method (Algorithm 2) controls the type-I error at the nominal level. Further, as τ grows, its power clearly dominates both the single-split test and the ordinary subsampling test, while closely tracking the power of the oracle test in all regimes. These observations align with our theory presented in Section 3.2. Also, note that the power of our test is insensitive to the split ratio q.

Meanwhile, our aggregated test significantly reduces the chance of non-replication. For example, when $q = 0.5$, for a random dataset, there is less than 5% chance that two applications of Algorithm 2 will give contradicting results, while the probability can be as large as 30% for the single-split test; see Appendix H.1.

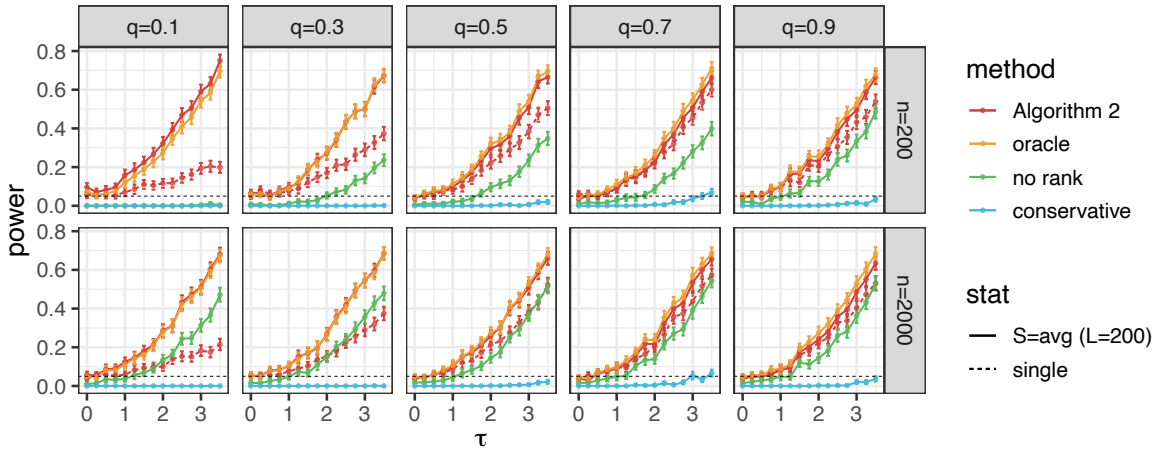

Figure 4: Testing $\mu = \mathbf{0}$ in Example 2: power (95% CI) at level $\alpha = 0.05$ (dashed horizontal). Location is $\mu = \tau n^{-1/2} v_1$ with v_1 the normalised principal eigenvector of Σ; q is the proportion of the test sample. We compare the single-split test (`single`) with tests based on S_n being the arithmetic mean (`S=avg`): Algorithm 2 is our method; '`oracle`' is an oracle test that compares S_n to its null distribution; '`no rank`' uses the ordinary subsampling distribution $\widehat{G}_n$ to determine the critical value; '`conservative`' compares $S_n/2$ to a standard normal, where division by two ensures validity when the statistics are not jointly normal as evidenced by Fig. 2 (see also Appendix B). Observe that Algorithm 2 has a power advantage over '`no rank`' and this advantage grows with τ; meanwhile, Algorithm 2 closely tracks '`oracle`' regardless of τ. These observations match our first-order theory presented in Section 3.2.

4.1.2 Testing multivariate unimodality

Testing non-trivial clustering structure of a high-dimensional dataset is a long-standing problem. The problem cannot be directly answered by clustering algorithms, because typically these (e.g.,

k-means or hierarchical clustering) always return clusters even when the data comes from a homogenous population (Huang et al., 2015a). This problem is closely connected to selecting the number of clusters as a trivial clustering structure corresponds to the true number being one. Here, we work with Euclidean data and we take the perspective that there is only one cluster if the population distribution is unimodal.

A univariate distribution is called unimodal if there is a point a such that the distribution function is convex on $(-\infty, a)$ and concave on $(a, +\infty)$ (Khintchine, 1938). While there are different notions of multivariate unimodality, we take *linear unimodality* as our definition: we say a random vector $(X_1, \ldots, X_p)$ is unimodal if $\sum_i a_i X_i$ is unimodal for every non-zero coefficient vector $a = (a_1, \ldots, a_p)$. That is,

$$H_0 : \bigcap_{a \neq \mathbf{0}} \left\{ \sum_{i=1}^{p} a_i X_i \text{ is unimodal} \right\}. \tag{11}$$

Linear unimodality is implied by several other notions related to multivariate unimodality such as log-concavity (Dharmadhikari and Joag-Dev, 1988, Lemma 2.1); see also Dharmadhikari and Joag-Dev (1988, Theorem 2.15). The formulation in Eq. (11) naturally leads to the following hunt-and-test procedure after randomly splitting the data into two parts A and B:

1. Identify a direction $\widehat{a}$ using any suitable clustering algorithm on part A of the data;
2. Test univariate unimodality of $\sum_i \widehat{a}_i X_i$ on part B.

We note that the idea of reducing to a univariate test is not new, and is for example used by Ahmed and Walther (2012) in projecting data onto its principal curve, and in a likelihood ratio test for log-concavity using random projection and data-splitting (Dunn et al., 2024).

Dip hunting test To identify a good direction $\widehat{a}$, on part A we run a 2-means algorithm (initialised with k-means++ by Vassilvitskii and Arthur, 2006) and choose $\widehat{a}$ as the normalised vector connecting the two cluster centres. Then to test for unimodality, we use a test based on the dip statistic due to Hartigan and Hartigan (1985), which we describe below. Fig. 5 shows a schematic of our procedure, which may be described as "dip hunting" by analogy with the bump hunting procedure of Good and Gaskins (1980).

Let $\mathbb{F}_n$ be the empirical distribution function of Y and let $\mathcal{U}$ be the set of unimodal univariate distributions. The dip statistic is defined as

$$\rho_n := \inf_{Q \in \mathcal{U}} \|\mathbb{F}_n - Q\|_\infty,$$

and may be computed efficiently using R package `diptest` (Maechler, 2021). Hartigan and Hartigan (1985) recommend comparing ρ_n to the dip statistic of a sample drawn from unif$(0, 1)$, which serves as the least favourable null distribution. However, this approach typically results in very conservative p-values. To avoid this problem, Cheng and Hall (1998) show that when the density f of Y is unimodal, under mild regularity conditions, we have $2n^{3/5}\rho_n \to_d c\mathcal{Z}$, where $\mathcal{Z}$ is a particular function of a standard Wiener process. The constant c depends on the density f and is given by

$$c = \left\{ f(x_0)^3 / |f''(x_0)| \right\}^{1/5},$$

where x_0 is the unique mode of f. The only unknown quantity c can be estimated with $\widehat{c}$, which is a plugin from kernel density estimates $\widehat{f}$ and $\widehat{f''}$ evaluated at $\widehat{x}_0 = \arg\max \widehat{f}$. We use R package `kedd` (Guidoum, 2015) to estimate $\widehat{f}$ and $\widehat{f''}$, for which the respective bandwidths are selected with maximum likelihood cross validation (function `h.mlcv`, Habbema et al., 1974;

Duin, 1976). From our experience, for $\widehat{c}$ to behave properly, it is essential to centre and rescale Y so that its value lies between 0 and 1, which does not affect c. Because the asymptotic distribution of the dip statistic only depends on c, Cheng and Hall (1998) suggest the following approach to obtain the corresponding p-value, which we adopt here. Given our observed dip statistic, we compare this to the distribution of ρ_n based on samples drawn from a known distribution whose c equals $\widehat{c}$; three families of such distributions covering the range of c are provided by Cheng and Hall (1998).

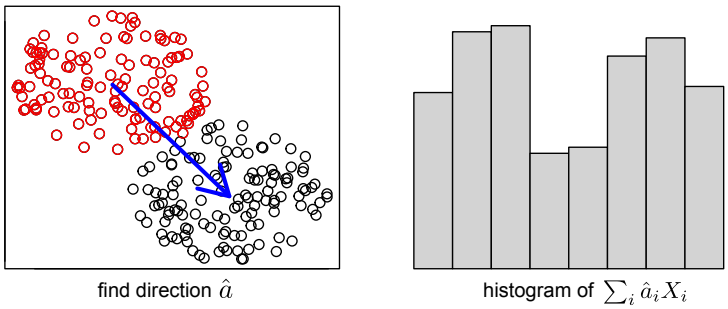

Figure 5: Dip hunting test for multivariate unimodality

Simulations We consider the following settings.

1. **Mixture of unit balls**. Let $\mathfrak{B}(p)$ be the unit p-dimensional ball centred at the origin. Consider the following density with bounded support:

$$f(x) = 1/2\,\mathrm{unif}_{\mathfrak{B}(p)}(x) + 1/2\,\mathrm{unif}_{\mathfrak{B}(p)}(x - x_0).$$

   Here, $\mathrm{unif}_{\mathfrak{B}(p)}$ denotes the uniform density on the unit ball $\mathfrak{B}(p)$ and $\|x_0\|$ is the Euclidean distance between the centres of the two balls. We set $\|x_0\| = 2\tau/\sqrt{2+p}$ so that the density in the direction connecting the two ball centres[1] becomes bimodal roughly when $\tau \geq 1$.

2. **Mixture of multivariate t's**. We consider a heavy-tailed setting

$$f(x) = 1/2\,t_4(x;\Sigma) + 1/2\,t_4(x - x_0;\Sigma),$$

   where t_4 is the density of the p-dimensional multivariate t-distribution with 4 degrees of freedom, mean zero and scale matrix $\Sigma \in \mathbb{R}^{p\times p}$ with entries $\Sigma_{ij} = 2^{-|i-j|}$. We set $x_0 = \tau v_2\sqrt{p}$, where v_2 is the second normalised eigenvector of Σ.

Fig. 6 shows the results based on sample size $n = 1000$. We compare the single-split dip hunting test (`single`) with the multiple-split versions that aggregate $L = 50$ dip hunting p-values, including $S = \mathrm{avg}$ and $S = \min$ (Algorithm 2), as well as the adaptive test (Algorithm 3) with $(S^1 = \mathrm{avg}, S^2 = \min)$. We compare dip hunting to `SigClust` (Liu et al., 2008; Huang et al., 2015b, 2022), a widely-used clustering significance testing method based on a Gaussian mixture model. Other methods include a nonparametric bootstrap approach suitable for ellipsoidal clusters (Maitra et al., 2012), and an approach based on simulating from an estimated Gaussian copula model (Helgeson et al., 2021); see also the review paper Adolfsson et al. (2019) and references therein. We see that at the null ($\tau = 0$), `SigClust` incurs a large type-I error for the

[1]The one-dimensional projection of $\mathrm{unif}_{\mathfrak{B}(p)}(x)$ has density $f(x) = \Gamma(p/2+1)/[\sqrt{\pi}\Gamma(p/2+1/2)](1-x^2)^{(p-1)/2}$ for $x \in [-1,1]$, which leads to variance $1/(2+p)$.

multivariate t settings, while all the dip hunting tests maintain the correct level in both settings; note that this is to be expected as `SigClust` assumes a Gaussian distribution under the null. For the unit ball setting, `SigClust` loses power as p increases. We can see that $S = \text{avg}$ is more powerful than $S = \min$ for the unit ball setting and conversely for the multivariate t setting. Nevertheless, our adaptive test is able to achieve the better performance between the two and also shows significant power improvement over the single-split test.

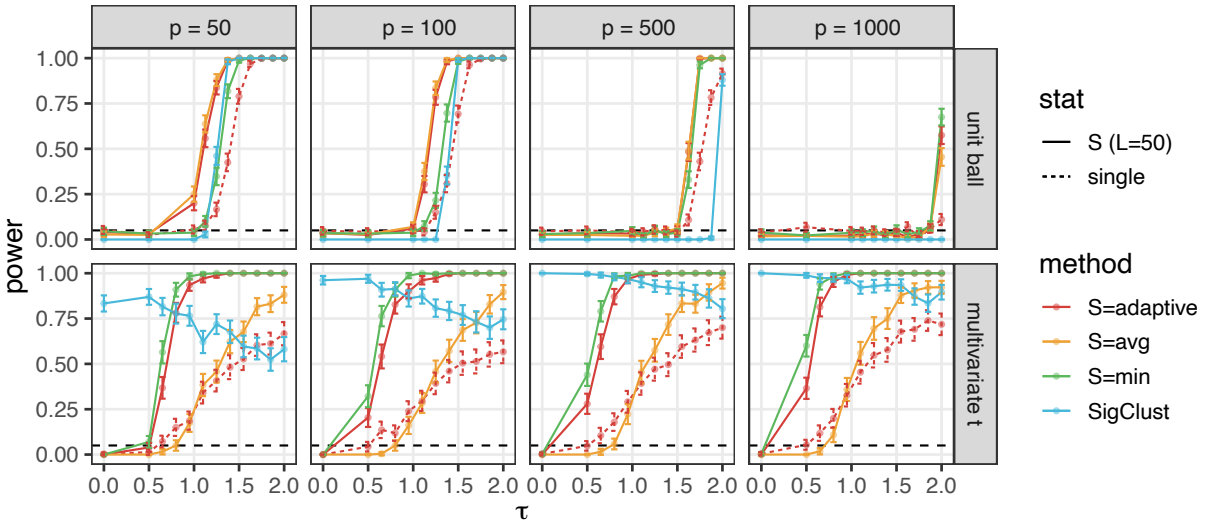

Figure 6: Detecting a mixture of two p-dimensional unimodal components whose centres are separated $\propto \tau$ away: power (95% CI) at level $\alpha = 0.05$ (dashed horizontal). The multivariate density is linearly unimodal when $\tau = 0$. We compare dip hunting with clustering significance testing method `SigClust`. Test `single` is the single-split dip hunting test; `S=avg`, `S=min` and `S=adaptive` are multiple-split tests that aggregate $L = 50$ dip hunting p-values. The `S=adaptive` test is Algorithm 3 with ($S^1 = \text{avg}, S^2 = \min$), which is able to adapt to the better performance between the two. See also Fig. H.4 for the performance of conservative p-value aggregation rules.

Gene expression of cancer subtypes We apply our test to gene expression data on renal cell carcinoma (RCC), which mainly consists of three subtypes: clear cell (ccRCC), papillary (PRCC) and chromophobe (ChRCC). We use the ICGC/TCGA Pan-Cancer dataset (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020, available from `https://bit.ly/3r65AUw`), which contains mRNA expression levels (FPKM-UQ normalised) from 111 kidney samples, including 37 cc, 31 P and 43 Ch RCC cases. We use the expression levels of 1,000 genes that are most relevant to RCC by selecting those with the highest $|\mu_1 - \mu_0|/\sigma_0$, where μ_1 and μ_0 are case and control means, σ_0 is the control standard deviation. We apply both the single-split and the aggregated (S is the arithmetic mean, $L = 6000$) dip hunting tests to every subtype, every mixture of two subtypes and the whole sample. Fig. 7 shows the distribution of p-values. The aggregated test produces stable p-values, which indicate clear separation between subtypes and relative homogeneity of each subtype. In contrast, it is more difficult to disentangle subtypes ccRCC and PRCC from the single-split dip hunting test.

4.1.3 Goodness-of-fit testing for parametric regression models

As another application, we use hunt and test to construct flexible goodness-of-fit tests for parametric regression models. The approach we take here is closely related to the generalised residual

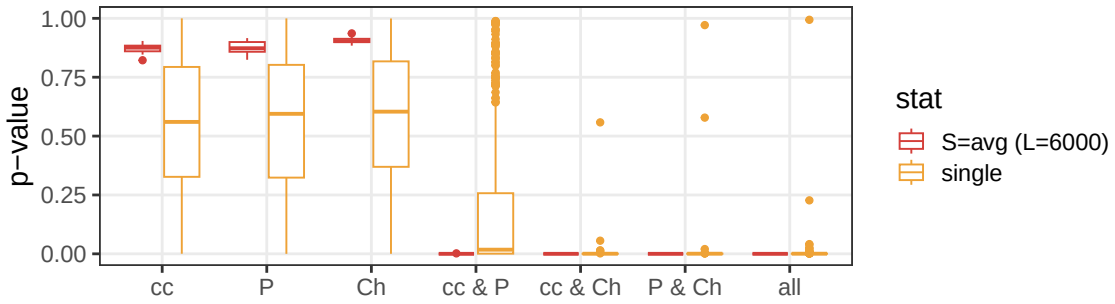

Figure 7: Testing homogeneity of gene expression levels of three subtypes of renal cell carcinoma: clear cell (cc), papillary (P) and chromophobe (Ch). The plot shows p-values resulting from repeatedly applying single-split and aggregated ($S = \text{avg}$, $L = 6000$) dip hunting tests to each subtype, every mixture of two subtypes and the whole sample.

prediction (RP) test (Janková et al., 2020; see also Shah and Bühlmann, 2018) for assessing the goodness of fit of (potentially high-dimensional) generalised linear models. Generalised RP tests also employ sample splitting and could equally well benefit from our rank transform p-value aggregation scheme. However, our construction here is applicable more broadly to testing model specification of the form

$$h(X) = \beta^\top X, \tag{12}$$

where $h(X)$ is the conditional mean or a conditional quantile of outcome $Y \in \mathbb{R}$ given $X \in \mathbb{R}^p$.

The starting point of our approach is the simple observation that $h(X) = \beta^\top X$ is equivalent to having $\gamma = 0$ in $h(X) = \beta^\top X + \gamma g(X)$ where $g : \mathbb{R}^p \to \mathbb{R}$ is any nonlinear (measurable) function. For a given g, we may test for whether $\gamma = 0$ by regressing Y on $(X, g(X))$ and utilising existing inference tools for the model at hand to assess the significance of $g(X)$. To obtain good power under an alternative, we would like to pick an appropriate g to expose the lack of fit present in the data. This suggests a hunt-and-test procedure where we randomly divide our data into parts A and B, and use part A to hunt for a suitable g, and part B to assess the significance of our artificially constructed additional covariate $g(X)$.

To find an appropriate g, we take inspiration from gradient boosting (Friedman, 2001) and proceed as follows. Let $(X_i, Y_i)_{i=1}^{n'}$ be iid covariate–response pairs in part A. Suppose we have an M-estimator $\widehat{\beta}$ for estimating β in Eq. (12) that minimises $\sum_i \ell(Y_i - \beta^\top X_i)$ for some loss function $\ell : \mathbb{R} \to \mathbb{R}$. Defining residual $r_i := Y_i - \widehat{\beta}^\top X_i$, upon introducing a potential new covariate $g(X)$, the loss can be locally approximated as

$$\sum_i \ell(Y_i - \widehat{\beta}^\top X_i - \gamma g(X)) \approx \sum_i \ell(r_i) - \gamma \sum_i g(X_i)\, \ell'(r_i),$$

where ℓ' is the derivative of the loss function. Hence, when $\gamma > 0$ and fixing $\big(\sum_i g(X_i)^2\big)^{1/2}$, to locally decrease the loss by the greatest amount, we should attempt to choose g such that approximately $(g(X_i))_{i=1}^{n'} \propto (\ell'(r_i))_{i=1}^{n'}$. To achieve this, we regress $(\ell'(r_i))_{i=1}^{n'}$ onto the covariates using any flexible regression or machine learning method, and take the fitted regression function to be g. Since we expect that under an alternative the resulting $g(X)$ should have a positive coefficient, we take as our single-split statistic T_n a Z-statistic for the significance of $g(X)$ computed on part B. We expect T_n to be large and positive under an alternative.

We demonstrate the effectiveness of this approach for quantile regression. Consider a quantile regression model specified as $q_\tau(X) = \beta_0 + \beta^\top X$ for a fixed $\tau \in (0, 1)$, where $q_\tau(X)$ is the

τ-th conditional quantile of Y given p-dimensional covariates X. The construction of goodness-of-fit tests, or more commonly called lack-of-fit tests in the related literature, have largely relied on asymptotic properties of certain statistics or processes concerning the residual; see, e.g., Horowitz and Spokoiny (2002); He and Zhu (2003); Escanciano and Velasco (2010); Escanciano and Goh (2014). These tests tend to have difficulty scaling up to more than a handful of covariates (Conde-Amboage et al., 2015). Recently, Dong et al. (2019) recast the goodness-of-fit problem as a two-sample test problem and developed a different, highly competitive method that can handle moderate or large p. It is worth mentioning that, unlike our approach that directly repurposes existing parameter inference for quantile regression (Koenker, 2005, Chap. 3), these aforementioned methods rely on asymptotic results that can require substantial development.

For quantile regression, we have $\ell_\tau(r) = r(\tau - \mathbb{I}_{r<0})$ and $\ell'_\tau(r) = \tau - \mathbb{I}_{r<0}$ for $r \neq 0$. Therefore, we use part A to train a classifier (e.g., random forest, Breiman, 2001) that predicts the sign of the residual from X; we take $g : \mathbb{R}^p \to \{-1, 1\}$ to be the resulting prediction function. To improve numerical stability, we also partial out X from $g(X)$ on part B before adding $g(X)$ to the covariates. Define the single-split test statistic as $T_n := \sqrt{n/2}\widehat{\beta}'/\widehat{\sigma}$ (the prime indicates fitted from part A), where $\widehat{\sigma}$ is estimated from bootstrap. The statistic can be readily computed using R package `quantreg` (Koenker, 2022).

Fig. 8 shows results from a simulation study under sample size $n = 1000$. Covariate vector X is drawn from p-dimensional Gaussian with covariance $\Sigma_{ij} = 2^{-|i-j|}$. We fix $\tau = 0.5$ and consider two specifications

$$\begin{aligned} &\text{(i)} \quad Y = 1 + \beta_0^\top X + (v/\sqrt{n})\eta(X) + (1 + X_2 + X_3)\varepsilon, \\ &\text{(ii)} \quad Y = 1 + \beta_0^\top X + \left[1 + X_2 + X_3 + (v/\sqrt{n})\eta(X)\right]\varepsilon, \end{aligned} \tag{13}$$

where non-linear function $\eta(X) = 4\sqrt{X_1^2 + X_2^2}$ introduces misspecification. We set $\beta_0^\top = (-1, 2, 0, -1, 2, 0, \dots)$. Error ε is drawn from $\{\mathsf{Exp}(1), t_3\}$ for specification (i) and from $\mathsf{Exp}(1)$ for specification (ii). Because $\mathsf{Exp}(1)$ has a non-zero median, observe that the quantile regression model $q_\tau(X) = \beta_0 + \beta^\top X$ is well-specified if and only if $v = 0$. We choose g to be a random forest trained with R package `ranger` (Wright and Ziegler, 2017). We run Algorithm 2 with S being the arithmetic mean. Our approach is already competitive with the state-of-the-art method of Dong et al. (2019); see Appendix H.2 for results in another setting.

4.2 Testing generalised conditional independence

In this section, we consider the use of randomised procedures in causal inference. To draw causal conclusions from data, we are often faced with the challenge that the quantity of interest is defined with respect to an "intervened" distribution, which is related to but different from the data-generating distribution. This is also known as a "distributional shift" in machine learning (Shimodaira, 2000).

For example, consider a two-stage sequentially randomised trial represented by graph $\mathcal{G}$ in Fig. 9: A_1 is the first treatment, L is the first outcome, A_2 is the second treatment and Y is the final outcome. The first treatment A_1 is completely randomised; the second treatment A_2 is randomised according to A_1 and L. Variable U represents an unobserved confounder, for example, the underlying health status that affects both outcomes and is unobserved. Graph $\mathcal{G}$, typically with extra base covariates that we omit here, is often employed to represent observational or follow-up studies where the treatment is time-varying and is affected by previous outcomes. For example, in HIV studies, we have $A_t = 1$ $(t = 1, 2)$ if the individual receives antiretroviral therapy at time t and $A_t = 0$ otherwise. Outcomes L and Y denote CD4 cell counts that measure the effectiveness of the therapy. See Hernán and Robins (2020, Chap. 19) for more background.

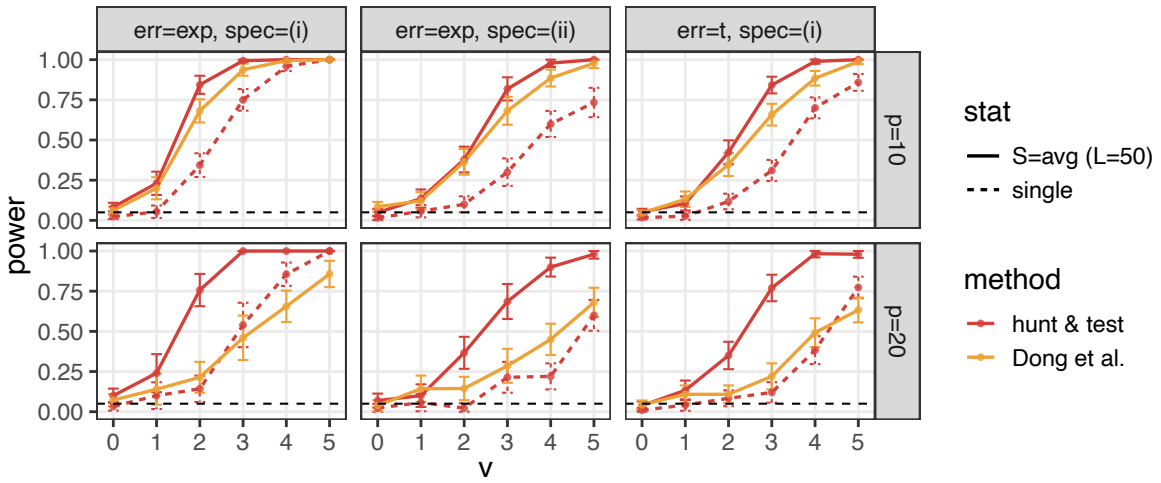

Figure 8: Testing goodness-of-fit of a quantile regression model $q_{0.5}(X) = \beta_0 + \beta^\top X$: power (95% CI) at level $\alpha = 0.05$ (dashed horizontal) under $n = 1000$. The model is well-specified if and only if $v = 0$. The non-linear function in Eq. (13) is $\eta(X) = 4\sqrt{X_1^2 + X_2^2}$. See also Figs. H.2 and H.5 in the supplementary material.

Suppose we are interested in the first treatment's *direct effect* τ on the final outcome (i.e., not through the second treatment), represented by the dashed edge in $\mathcal{G}$. Because A_1 also affects Y through A_2, we cannot learn τ by regressing Y on A_1 alone. Moreover, because of the latent U that affects both L and Y (such variables are called "phantoms" by Bates et al., 2022), nor can we learn τ by additionally adjusting for L or A_2 (or both) in the regression.

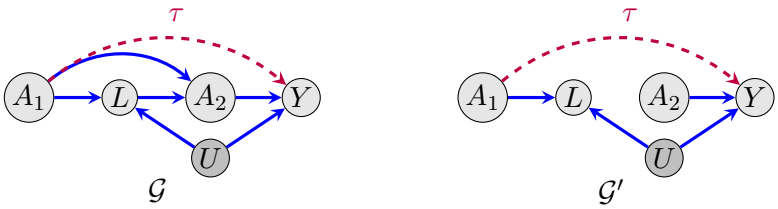

Figure 9: Graph $\mathcal{G}$ depicts a sequentially randomised trial: A_1 is the first treatment, L is the first outcome, A_2 is the second treatment, and Y is the final outcome. Latent variable U represents the underlying health status that affects both outcomes. The dashed edge represents the direct effect from A_1 on Y. Graph $\mathcal{G}'$ represents the population where both A_1 and A_2 are completely randomised. In this case, if A_1 has no direct effect on Y (i.e., the dashed edge is absent), then we can observe $A_1 \perp\!\!\!\perp Y$.

To learn τ, it is useful to imagine another trial, drawn as $\mathcal{G}'$ in Fig. 9, where both treatments A_1 and A_2 are completely randomised so that A_1 only affects Y directly. There, we can easily learn τ by regressing Y on A_1 in the $\mathcal{G}'$ distribution. Although we did not carry out the $\mathcal{G}'$ trial, its data distribution can be approximated by reweighting our data obtained from the $\mathcal{G}$ trial according to the inverse propensity of A_2 given A_1 and L. Further, we can even artificially *simulate* data from the $\mathcal{G}'$ trial by resampling (e.g., importance resampling or rejection sampling) our data from the $\mathcal{G}$ trial. In other words, reweighting or resampling provides access to our distribution of interest. This idea underlies a general approach known as the g-methods (Robins, 1986; Naimi et al., 2017).

Specifically, let us consider testing the sharp null hypothesis

$$H_0 : A_1 \text{ has no individual direct effect on } Y, \tag{14}$$

which is represented by the dashed edge from A_1 to Y being absent from $\mathcal{G}$. As explained earlier, this amounts to $A_1 \perp\!\!\!\perp Y$ in the population Q represented by $\mathcal{G}'$, given by

$$dQ/dP = q(A_2)/p(A_2 \mid A_1, L), \tag{15}$$

where q is an arbitrary positive distribution over A_2. Constraints as such, which prescribe independence or conditional independence in a reweighted distribution, are called generalised conditional independence or Verma constraints in the literature; see Richardson et al. (2023) and the references therein.

Because the independence holds under Q instead of P, the usual permutation test is not applicable; nor is it applicable through simple reweighting as employed by Berrett et al. (2020) for testing ordinary conditional independence. Instead, the standard approach in causal inference is through inverse probability weighting (IPW). Here, we consider an alternative approach: we resample our data to represent Q according to Eq. (15) and then use the resampled data to test $A_1 \perp\!\!\!\perp Y$, e.g., by a permutation test or any off-the-shelf test for independence.

More generally, as demonstrated by Thams et al. (2023), such a test-after-subsampling procedure is applicable to testing any property under "distributional shift". Suppose we observe iid sample $X_1, \dots X_n \sim P$ but we are interested in testing a property of a different target distribution Q. Distribution Q is related to P through a density ratio $r = \mathrm{d}Q/\,\mathrm{d}P$, which is either known or can be estimated from P. Suppose we already have a suitable test for the property based on a test statistic $T_n := T_n(\widetilde{X}_1, \dots, \widetilde{X}_n)$ for iid sample $\widetilde{X}_1, \dots, \widetilde{X}_n \sim Q$. Assume r is bounded from above by a constant C. Then, we can test the property in two steps:

1. Obtain a sample from Q through rejection sampling;
2. Test the property with the test statistic computed from the accepted sample.

The next result shows that the test statistic computed as such inherits the desired asymptotic distribution while permitting the use of an estimated density ratio $\widehat{r}_n$.

Proposition 1 (Test after rejection sampling)**.** *Let $T_n(X_1, \dots, X_n)$ be a test statistic. Define $T_0 := 0$. With $X := (X_1, \dots, X_n)$, suppose $T_n(X_1, \dots, X_n) \to_d T$ under $X \sim Q^n$. Let P be a distribution and C be a positive constant, such that Q is absolute continuous with respect to P and $r := \mathrm{d}Q/\,\mathrm{d}P$ satisfies*

$$r(x) \le C, \quad P\text{-almost every } x.$$

Let $\widehat{r}_n$ be a random sequence of approximate density ratios such that almost surely,

$$\int \widehat{r}_n \,\mathrm{d}P = 1, \quad 0 \le \widehat{r}_n(x) \le C \text{ for } P\text{-almost every } x.$$

Also, suppose

$$\mathbb{E} \int |\widehat{r}_n - r| \,\mathrm{d}P = o(n^{-1}). \tag{16}$$

With $U := (U_1, \dots, U_n)$, consider $(X, U) \sim P^n \times \mathrm{unif}^n(0,1)$. For $i = 1, \dots, n$, let the i-th sample be accepted if $U_i < \widehat{r}_n(X_i)/C$. Let the set of accepted sample be denoted as $\widetilde{X}_1^n, \dots, \widetilde{X}_{\Gamma_n}^n$, where $0 \le \Gamma_n \le n$ is the number of acceptances. Then it holds that

$$T_{\Gamma_n}(\widetilde{X}_1^n, \dots, \widetilde{X}_{\Gamma_n}^n) \to_d T.$$

While the rate in Eq. (16) seems to demand a large separate sample for estimating r, we expect that such rate can be relaxed under additional conditions on the density and the test statistic. For example, when the density ratio is parametrically specified and the test statistic is simple, see Proposition C.2 and the following remark in the supplementary material.

When the density ratio r is unbounded, the rejection sampling can be replaced by the distinct replacement sampling (DRPL) introduced by Thams et al. (2023), which selects a resampled dataset $(\widetilde{X}_1 = X_{i_1}, \ldots, \widetilde{X}_k = X_{i_k})$ of given size k by sampling distinct indices $(i_1, \ldots, i_k)$ with probability proportional to $\prod_{l=1}^{k} r(X_{i_l})$. It is shown that, under a moment condition on r, by choosing $k = o(\sqrt{n})$, the statistic T_n evaluated on the DRPL resampled data converges to its asymptotic distribution under Q (Thams et al., 2023, Theorem 1). However, DRPL produces an $o(\sqrt{n})$ sample while rejection sampling produces an $O(n)$ sample with high probability.

Both rejection sampling and DRPL lead to tests that are randomised and potentially of low power due to a reduced sample size. However, as we demonstrate here in the case of testing Eq. (14) in Fig. 9, our aggregation scheme is able to restore highly competitive levels of power. For simplicity, we choose the post-rejection/DRPL-sampling test statistic T_n as the permutation p-value for $|\operatorname{cov}_Q(A_1, Y)|$, which is uniformly distributed when $A_1 \perp\!\!\!\perp Y$ holds under Q. As a benchmark, we also consider testing Eq. (14) with the standard IPW version of the statistic Robins (1999, p. 10), which directly uses the sample under P. Let

$$Z_i := \frac{Y_i(A_{1,i} - \mathbb{E}\, A_1)}{p(A_{2,i} \mid L_i, A_{1,i})},$$

where $\mathbb{E}\, A_1$ and $p(A_2 \mid L, A_1)$ can be replaced by their consistent, asymptotically linear estimators. Under H_0, we have

$$\chi_n := \frac{\sum_i Z_i}{\sqrt{\sum_i Z_i^2}} \to_d \mathcal{N}(0,1), \tag{17}$$

which leads to a two-sided test; see also Proposition C.4 in the supplementary material.

We consider the following data generating mechanism. We have binary treatments A_1, A_2 and real-valued outcomes L, Y. The latent U is a 4-dimensional random vector. Let

$$A_1 \sim \mathsf{Ber}(1/2), \quad A_2 \sim \mathsf{Ber}(\operatorname{expit}(2A_1 - L + 2)),$$

and

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

where $\beta_{U,L} = (1, 1, -2, 2)^\top$, $\beta_{U,Y} = (2, -1, 3, -10)^\top$ and $\varepsilon_L, \varepsilon_Y \sim \mathcal{N}(0,1)$ independently. Parameter τ controls the effect size: H_0 in Eq. (14) holds if and only if $\tau = 0$.

Fig. 10 shows the result for sample size $n = 1000$ under P. For rejection sampling, $q(A_2)$ in Eq. (15) is chosen to maximise the acceptance rate; for DRPL, we choose $q(A_2) = p(A_2)$ and $k = \lfloor \sqrt{n} \rfloor$. We use Algorithm 2 to aggregate $L = 20$ post-rejection/DRPL-sampling permutation p-values (S=avg). Presumably because rejection sampling generates a larger sample than DRPL, rejection sampling tests are more powerful than DRPL tests here. For both rejection sampling and DRPL, 'avg' considerably boosts the power while maintaining the correct level at $\tau = 0$. Perhaps surprisingly, the 'avg' rejection sampling test outperforms the IPW test based on Eq. (17), even though they target the same population quantity $\operatorname{cov}_Q(A_1, Y)$.

Our strategy is particularly useful when it is difficult to detect the dependence between A_1 and Y under Q with simple, "mean-like" statistics such as covariance and a more complex statistic must be employed. In such cases, it is not always clear how to reweight and calibrate the statistic under P using IPW. In contrast, post-resampling test works out of the box whenever a test under Q can be constructed. We present such an example in Appendix H.3.

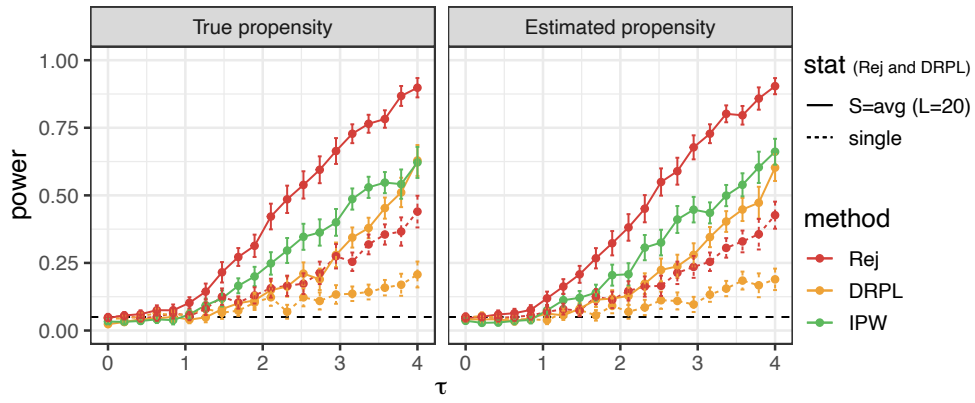

Figure 10: Power for testing no individual direct effect of A_1 on Y in graph $\mathcal{G}$ of Fig. 9, based on $n = 1000$ sample at level 0.05 (dashed horizontal). The null hypothesis corresponds to $\tau = 0$. `Rej`: post-rejection-sampling permutation test for $|\operatorname{cov}_Q(A_1, Y)|$, `DRPL`: post-DRPL-sampling (Thams et al., 2023) permutation test for $|\operatorname{cov}_Q(A_1, Y)|$, `IPW`: inverse probability weighted test based on $\operatorname{cov}_Q(A_1, Y) = 0$ given by Eq. (17). `Rej` and `DRPL` are randomised tests ('single'). Their 'S=avg' version is the aggregated test by taking the average of 20 p-values, calibrated by Algorithm 2. Left panel: the true propensity $p(A_2 \mid A_1, L)$ is used for sampling; right panel: the propensity used for sampling is estimated using logistic regression. See also Fig. H.6 in the supplementary material for comparison with conservatively merged p-values.

4.3 Calibrating cross-fitted double machine learning

Double/debiased machine learning (DML) methods, also known as doubly robust targeted estimation, employ flexible, machine learning methods to estimate low-dimensional target parameters. These methods are widely used for a variety of problems; see Díaz (2020); Kennedy (2022) for recent reviews. In this section, we follow the setup of Chernozhukov et al. (2018) and focus on estimation with Neyman orthogonal scores, but our method can be extended to other settings such as those that are based on influence functions with a mixed bias property (Rotnitzky et al., 2021).

Suppose we have an estimating equation $\psi(W; \theta, \eta)$ for an unknown, real-valued target parameter θ_0 such that $\mathbb{E}\,\psi(W; \theta_0, \eta_0) = 0$. Here W denotes the data point and η_0 denotes two or more unknown nuisance parameters. Suppose $\psi(W; \theta, \eta)$ satisfies the Neyman orthogonality condition (Chernozhukov et al., 2018, Definition 2.1) and takes the form

$$\psi(W; \theta, \eta) = \psi^a(W; \eta)\,\theta + \psi^b(W; \eta).$$

The DML approach proposes to estimate η_0 using flexible machine learning methods, and employs sample-splitting to control bias that may otherwise be introduced by overfitting. Suppose iid data $W_1, \ldots, W_n$ are split into L equal-sized folds. For $l = 1, \ldots, L$, let $\widehat{\eta}^{(-l)}$ be the nuisance parameters learned from all the data except the l-th fold. Then, let $\widehat{\theta}^{(l)}$ be the solution to

$$\mathbb{P}_n^{(l)} \psi(W; \theta, \eta^{(-l)}) = 0, \quad l = 1, \ldots, L,$$

where $\mathbb{P}_n^{(l)}$ denotes the empirical measure of data in the l-th fold. Under regularity conditions and a sufficiently fast rate of nuisance estimation (Chernozhukov et al., 2018, Assumptions 3.1 & 3.2), it can be shown that the single-split statistic

$$T_n^{(l)} := \frac{\sqrt{n/L}(\widehat{\theta}_n^{(l)} - \theta_0)}{\sigma} \to_d \mathcal{N}(0, 1), \quad l = 1, \ldots, L, \tag{18}$$

where the asymptotic variance is

$$\sigma^2 = \mathbb{E}\,\psi^2(W;\theta_0,\eta_0)/\{\mathbb{E}\,\psi^a(W;\eta_0)\}^2. \tag{19}$$

The so-called cross-fitted DML-1 estimator is simply the average $\widehat{\theta}^{\text{dml}} := (\widehat{\theta}^{(1)} + \cdots + \widehat{\theta}^{(L)})/L$. Under the same conditions as above, it holds that

$$\sqrt{L}S_n = \frac{1}{\sqrt{L}}\sum_{l=1}^{T} T_n^{(l)} = \frac{\sqrt{n}(\widehat{\theta}_n^{\text{dml}} - \theta_0)}{\sigma} \to_d \mathcal{N}(0,1), \tag{20}$$

where S_n is the average of exchangeable statistics $T_n^{(1)},\ldots,T_n^{(L)}$. Provided with a consistent estimator of σ, standard large-sample DML confidence intervals are constructed from Eq. (20).

In view of Eq. (18), Eq. (20) effectively ignores the between-fold correlations among statistics $T_n^{(1)},\ldots,T_n^{(L)}$ because their contribution is asymptotically negligible under the conditions. However, in practice, such correlations can be visible in finite sample and hence undermine the coverage of normal confidence intervals constructed from Eq. (20). When the between-fold correlation is ρ, the variance of $\sqrt{n}(\widehat{\theta}_n^{\text{dml}} - \theta)$ is roughly $\sigma^2(1+\rho(L-1))$ instead of σ^2. We propose to use Algorithm 2 to account for such correlation and construct confidence intervals with better finite-sample coverage. In this context, $\widehat{H}_{b,l}$ in Algorithm 2 is the subsample counterpart of Eq. (18), which involves unknown σ and θ_0. Nevertheless, by monotonicity, the rank-transformed $\widetilde{\mathbf{H}}$ can be computed from the ranks of the subsampling counterparts of $\widehat{\theta}_n^{(l)}$ directly as

$$\widetilde{H}_{b,l} = \Phi^{-1}\left(\mathbb{F}_{\{\widehat{\theta}_{m,b}^{(l)}\}}(\widehat{\theta}_{m,b}^{(l)}) - 1/(2BL)\right), \quad b = 1,\ldots,B, \quad l = 1,\ldots,L.$$

Here $\widehat{\theta}_{m,b}^{(l)}$ is the l-th single-split estimator computed from the b-th subsample. Further, because $\widetilde{H}_{b,l} \approx \widehat{H}_{b,l} = \sqrt{m/L}(\widehat{\theta}_{m,b}^{(l)} - \theta_0)/\sigma$ with high probability, we get the following estimator for σ as a by-product:

$$\widehat{\sigma}_{\text{ls}} := \sqrt{m/L}/\widehat{\beta}_1.$$

Here $\widehat{\beta}_1$ the slope from least squares $\widetilde{H}_{b,l} \sim \beta_1\widehat{\theta}_{m,b}^{(l)} + \beta_0$. We perform this least squares regression only using those points for which $\mathbb{F}_{\{\widehat{\theta}_{m,b}^{(l)}\}}(\widehat{\theta}_{m,b}^{(l)}) \in (\varepsilon/2, 1-\varepsilon/2)$ for some small $\varepsilon > 0$ to avoid high-leverage points influencing the fit too much; in our numerical experiments we use $\varepsilon = 0.1$.

Proposition 2. *Suppose $\widehat{\theta}_n^{(1)},\ldots\widehat{\theta}_n^{(L)}$ are exchangeable and Eq. (18) holds. Then, $\widehat{\sigma}_{ls} \to_p \sigma$.*

Consequently, our rank-transformed subsampling confidence interval for θ_0 is

$$\left[\widehat{\theta}_n^{\text{dml}} - \sqrt{L/n}\,\widehat{\sigma}_{\text{ls}}\,\widetilde{G}_n^{-1}(1-\alpha/2),\ \widehat{\theta}_n^{\text{dml}} - \sqrt{L/n}\,\widehat{\sigma}_{\text{ls}}\,\widetilde{G}_n^{-1}(\alpha/2)\right]. \tag{21}$$

As a simple example, consider inferring θ_0 in a partially linear model

$$\begin{aligned} D &= m_0(X) + V, \quad \mathbb{E}[V \mid X] = 0, \\ Y &= D\theta_0 + g_0(X) + \xi, \quad \mathbb{E}[\xi \mid D, X] = 0 \end{aligned}$$

using ψ given by Robinson's score function (Chernozhukov et al., 2018, §4.1)

$$\psi(W;\theta,\eta) = \{Y - l(X) - \theta(D - m(X))\}(D - m(X)), \quad l(X) := \mathbb{E}[Y \mid X],\ m(X) = \mathbb{E}[D \mid X].$$

In our numerical example, we choose heteroscedastic errors V and ξ (see Appendix E for details) and fit nuisance parameters $\eta := (l, m)$ with random forests. Table 1 compares the coverage of

our confidence interval Eq. (21) with the coverage of standard DML normal confidence interval based on Eq. (20). For the latter, we use a plugin estimate of σ that replaces expectations in Eq. (19) by their empirical counterparts. While DML confidence intervals tend to undercover for smaller samples, where $\rho(L-1)$ is larger, our confidence intervals are well calibrated.

When one of the two nuisance functions above is misspecified, under assumptions, $\widehat{\theta}_n^{\mathrm{dml}}$ can still be consistent and asymptotically normal due to the double robustness of the estimating equation. However, the estimators $\widehat{\theta}_n^{(1)}, \ldots \widehat{\theta}_n^{(L)}$ are correlated, even asymptotically. In such a case, the standard DML confidence interval no longer has the desired asymptotic coverage but our confidence interval Eq. (21) should still work. See Appendix E for such an example; see also Benkeser et al. (2017) for related methods.

Table 1: Coverage of nominal 95% confidence intervals (brackets: median width of intervals; ρ is the between-fold correlation of $T_n^{(l)}$)

	$n = 500$		$n = 1000$		$n = 2000$	
method	$L = 2$	$L = 5$	$L = 2$	$L = 5$	$L = 2$	$L = 5$
$\rho(L-1)$	0.46	0.31	0.36	0.18	0.25	0.14
Eq. (21)	0.94 [0.21]	0.93 [0.19]	0.95 [0.15]	0.95 [0.13]	0.96 [0.10]	0.95 [0.09]
DML	0.86 [0.15]	0.88 [0.15]	0.88 [0.11]	0.92 [0.11]	0.91 [0.08]	0.92 [0.08]

5 Discussion

Rank-transformed subsampling provides a framework that properly aggregates results from multiple applications of a randomised statistical test. Backed by this framework, we are free to design explicitly randomised tests that employ data splitting, resampling or any other random processing to solve difficult problems. In particular, data splitting and resampling can often reduce a complex hypothesis to a simpler one, for which an off-the-shelf test can be repurposed in a "plug-and-play" fashion. Even though the resulting "single-split" randomised procedure may have high variability or low power, as we have demonstrated, these aspects can be significantly improved by our aggregation algorithms.

There are several aspects worth further investigation. Firstly, it is of interest to go beyond Algorithm 3 and study how to learn an optimal aggregation function. Secondly, with a large L and B, constructing $\widehat{\mathbf{H}}$ can be computationally intensive, although it is embarrassingly parallelisable across both rows and columns. It is desirable to develop faster approximations to $\widehat{\mathbf{H}}$. Thirdly, we find $m = \lfloor n/\log n \rfloor$ works surprisingly well in our numerical studies, in spite of only guaranteeing around 7 independent subsamples for $n = 1000$. Essentially, rank-transformed subsampling is able to approximate the null sampling distribution far better than the theory might suggest, and it would be interesting to investigate why this might be the case, and whether there is a better choice of the subsample size (e.g., using the rule of Bickel and Sakov, 2008).

Acknowledgements The authors thank Richard Moulange and Sach Mukherjee for help with the TCGA data, and Richard Samworth, Qingyuan Zhao and four anonymous referees for their helpful comments. Both authors were funded by EPSRC grants EP/N031938/1 and EP/R013381/1.

References

Andreas Adolfsson, Margareta Ackerman, and Naomi C Brownstein. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26, 2019.

Murat O Ahmed and Guenther Walther. Investigating the multimodality of multivariate data with principal curves. *Computational Statistics & Data Analysis*, 56(12):4462–4469, 2012.

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 2016.

Rina Foygel Barber and Lucas Janson. Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *The Annals of Statistics*, 50(5):2514–2544, 2022.

Stephen Bates, Edward Kennedy, Robert Tibshirani, Valerie Ventura, and Larry Wasserman. Causal inference with orthogonalized regression adjustment: Taming the phantom. *arXiv preprint arXiv:2201.13451*, 2022.

David Benkeser, Marco Carone, MJ Van Der Laan, and Peter B Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

Rudolf Beran. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468, 1987.

Rudolf Beran. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697, 1988.

Rudolf Beran. Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics*, 49:1–24, 1997.

Arthur Berg, Timothy L McMurry, and Dimitris N Politis. Subsampling p-values. *Statistics & Probability Letters*, 80(17-18):1358–1364, 2010.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

Peter J Bickel and Anat Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, pages 967–985, 2008.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Zhanrui Cai, Jing Lei, and Kathryn Roeder. Model-free prediction test with application to genomics data. *Proceedings of the National Academy of Sciences*, 119(34):e2205518119, 2022.

M-Y Cheng and Peter Hall. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):579–589, 1998.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.

Woohyun Choi and Ilmun Kim. Averaging p-values under exchangeability. *Statistics & Probability Letters*, page 109748, 2022. doi: https://doi.org/10.1016/j.spl.2022.109748.

Mercedes Conde-Amboage, César Sánchez-Sellero, and Wenceslao González-Manteiga. A lack-of-fit test for quantile regression models with high-dimensional covariates. *Computational Statistics & Data Analysis*, 88:128–138, 2015.

David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62 (2):441–444, 1975.

Ben Dai, Xiaotong Shen, and Wei Pan. Significance tests of feature relevance for a black-box learner. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Sudhakar Dharmadhikari and Kumar Joag-Dev. *Unimodality, Convexity, and Applications*. Academic Press, San Diego, 1988.

Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2020.

Cyrus J DiCiccio. *Hypothesis Testing Using Multiple Data Splitting*. Stanford University, 2018.

Cyrus J DiCiccio, Thomas J DiCiccio, and Joseph P Romano. Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166:108865, 2020.

Chen Dong, Guodong Li, and Xingdong Feng. Lack-of-fit tests for quantile regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):629–648, 2019.

R. P. W Duin. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, 25(11):1175–1179, 1976.

Robin Dunn, Aditya Gangrade, Larry Wasserman, and Aaditya Ramdas. Universal inference meets random projections: A scalable test for log-concavity. *Journal of Computational and Graphical Statistics*, pages 1–13, 2024. doi: 10.1080/10618600.2024.2347338.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

Juan Carlos Escanciano and Sze-Chuan Goh. Specification analysis of linear quantile models. *Journal of Econometrics*, 178:495–507, 2014.

Juan Carlos Escanciano and Carlos Velasco. Specification tests of parametric dynamic conditional quantiles. *Journal of Econometrics*, 159(1):209–221, 2010.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

IJ Good and RA Gaskins. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75(369):42–56, 1980.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Arsalane Chouaib Guidoum. *kedd: Kernel estimator and bandwidth selection for density and its derivatives.*, 2015. URL http://CRAN.R-project.org/package=kedd. R package version 1.0.3.

JDF Habbema, J. Hermans, and K van den Broek. A stepwise discriminant analysis program using density estimation. In *Compstat 1974: Proceedings in Computational Statistics*, Vienna, 1974. Physica Verlag.

John A Hartigan and Pamela M Hartigan. The dip test of unimodality. *The Annals of Statistics*, pages 70–84, 1985.

Xuming He and Li-Xing Zhu. A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, 98(464):1013–1022, 2003.

Erika S Helgeson, David M Vock, and Eric Bair. Nonparametric cluster significance testing with reference to a unimodal null distribution. *Biometrics*, 77(4):1215–1226, 2021.

M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

Joel L Horowitz and Vladimir G Spokoiny. An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association*, 97(459):822–835, 2002.

Hanwen Huang, Yufeng Liu, David Neil Hayes, Andrew Nobel, J. S Marron, and Christian Hennig. *Significance testing in clustering*, pages 315–335. Chapman and Hall/CRC, 2015a.

Hanwen Huang, Yufeng Liu, Ming Yuan, and J. S Marron. Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, 24(4):975–993, 2015b.

Hanwen Huang, Yufeng Liu, and J. S Marron. *SigClust: Statistical Significance of Clustering*, 2022. URL https://CRAN.R-project.org/package=sigclust. R package version 1.1.0.1.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020. doi: 10.1038/s41586-020-1969-6. URL https://doi.org/10.1038/s41586-020-1969-6.

John PA Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Jana Janková, Rajen D Shah, Peter Bühlmann, and Richard J Samworth. Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):773–795, 2020.

Kuanhao Jiang, Rajarshi Mukherjee, Subhabrata Sen, and Pragya Sur. A new central limit theorem for the augmented ipw estimator: Variance inflation, cross-fit covariance and beyond. *arXiv preprint arXiv:2205.10198*, 2022.

Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, New York, 2005.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

A Ya Khintchine. On unimodal distributions. *Izvestiya Nauchno-Issledovatel'skogo Instituta Matematiki i Mekhaniki*, 2(2):1–7, 1938.

Ilmun Kim and Aaditya Ramdas. Dimension-agnostic inference using cross U-statistics. *Bernoulli*, 30(1):683 – 711, 2024. doi: 10.3150/23-BEJ1613. URL https://doi.org/10.3150/23-BEJ1613.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

Roger Koenker. *quantreg: Quantile Regression*, 2022. URL https://CRAN.R-project.org/package=quantreg. R package version 5.94.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Wanjun Liu, Xiufan Yu, and Runze Li. Multiple-splitting projection test for high-dimensional mean vectors. *Journal of Machine Learning Research*, 23(71):1–27, 2022.

Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

Yufeng Liu, David Neil Hayes, Andrew Nobel, and J. S Marron. Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008. URL https://doi.org/10.1198/016214508000000454.

Anton Rask Lundborg, Ilmun Kim, Rajen D Shah, and Richard J Samworth. The projected covariance measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2211.02039*, 2022.

Martin Maechler. *diptest: Hartigan's Dip Test Statistic for Unimodality - Corrected*, 2021. URL https://CRAN.R-project.org/package=diptest. R package version 0.76-0.

Ranjan Maitra, Volodymyr Melnykov, and Soumendra N Lahiri. Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, 107(497):378–392, 2012.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

Timothy L McMurry, Dimitris N Politis, and Joseph P Romano. Subsampling inference with k populations and a non-standard behrens–fisher problem. *International Statistical Review*, 80 (1):149–175, 2012.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Xiao-Li Meng. Posterior Predictive p-Values. *The Annals of Statistics*, 22(3):1142 – 1160, 1994.

Patrick AP Moran. Dividing a sample into two parts a statistical dilemma. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 329–333, 1973.

Alfred Müller and Dietrich Stoyan. *Comparison Methods for Stochastic Models and Risks*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2002.

Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An introduction to g methods. *International Journal of Epidemiology*, 46(2):756–762, 2017.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.

Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

James M Robins. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. *Computation, Causation, and Discovery*, pages 349–405, 1999.

Joseph P Romano and Cyrus DiCiccio. Multiple data splitting for testing. Technical report, Department of Statistics, Stanford University, 2019. URL https://purl.stanford.edu/fb041jg0790.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022.

Rajen D Shah and Peter Bühlmann. Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):113–135, 2018.

Rajen D Shah and Richard J Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Statistics & Probability Letters*, 184:109395, 2022.

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022.

Nikolaj Thams, Sorawit Saengkyongam, Niklas Pfister, and Jonas Peters. Statistical testing under distributional shifts. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):597–663, 2023.

Jinjin Tian, Xu Chen, Eugene Katsevich, Jelle Goeman, and Aaditya Ramdas. Large-scale simultaneous inference under dependence. *Scandinavian Journal of Statistics*, 50(2):750–796, 2023.

Aad van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Sergei Vassilvitskii and David Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006.

Roman Vershynin. *High-Dimensional Probability: an Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

John von Neumann. Various techniques used in connection with random digits. *Applied Math Series*, 12(36-38):3, 1951.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Yuhao Wang and Rajen D Shah. Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders. *arXiv preprint arXiv:2011.08661*, 2020.

Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

A Gaussian location experiment

Here we derive expressions for the power and probability of replication relating to Example 1. The sequence $T^{(1)}, T^{(2)}, \ldots$ follows a Gaussian process with mean and covariance

$$\mu_i = \mu, \quad \Sigma_\rho = \begin{cases} 1, & i = j \\ \rho, & i \neq j \end{cases}.$$

It follows that

$$\begin{pmatrix} \bar{T}_{1:L} \\ \bar{T}_{(1+L):2L} \end{pmatrix} := \begin{pmatrix} L^{-1} \sum_{l=1}^{L} T^{(l)} \\ L^{-1} \sum_{l=L+1}^{2L} T^{(l)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} 1/L + \rho(L-1)/L & \rho \\ \rho & 1/L + \rho(L-1)/L \end{pmatrix} \right).$$

The critical value is

$$\bar{t}_\alpha = \sqrt{1/L + \rho(L-1)/L}\, z_\alpha,$$

which reduces to z_α when $L = 1$. The power function of the aggregated test given by

$$\begin{aligned} \mathbb{E}\, \phi_{\bar{T}_{1:L}} = \mathbb{P}(\bar{T}_{1:L} > \bar{t}_\alpha) &= \Phi\left(\frac{\mu}{\sqrt{1/L + \rho(L-1)/L}} - z_\alpha \right) \\ &= \alpha + \phi(z_\alpha) \frac{\mu}{\sqrt{\frac{1}{L} + \frac{L-1}{L}\rho}} + o(\mu) \end{aligned}$$

as $\mu \searrow 0$. To derive the expression for $\mathbb{P}(\phi_{\bar{T}_{1:L}} \neq \phi_{\bar{T}_{(L+1):2L}})$, consider the representation

$$\begin{pmatrix} \bar{T}_{1:L} \\ \bar{T}_{(1+L):2L} \end{pmatrix} =_d \begin{pmatrix} \mu + \sqrt{\rho} Z_0 + \sqrt{(1-\rho)/L} Z_1 \\ \mu + \sqrt{\rho} Z_0 + \sqrt{(1-\rho)/L} Z_2 \end{pmatrix},$$

where Z_0, Z_1, Z_2 are independent standard normals. Then we can write

$$\begin{aligned} &\mathbb{P}(\bar{T}_{1:L} > \bar{t}_\alpha, \bar{T}_{(L+1):2L} > \bar{t}_\alpha) \\ &= \mathbb{E}\, \mathbb{P}(\sqrt{(1-\rho)/L} Z_1 > \bar{t}_\alpha - \mu - \sqrt{\rho} Z_0, \sqrt{(1-\rho)/L} Z_2 > \bar{t}_\alpha - \mu - \sqrt{\rho} Z_0 \mid Z_0) \\ &= \mathbb{E}\left\{ 1 - \Phi\left(\frac{\bar{t}_\alpha - \mu - \sqrt{\rho} Z_0}{\sqrt{(1-\rho)/L}} \right) \right\}^2 = \mathbb{E}\, \Phi^2\left(\frac{\mu + \sqrt{\rho} Z_0 - \bar{t}_\alpha}{\sqrt{(1-\rho)/L}} \right), \end{aligned}$$

and similarly

$$\mathbb{P}(\bar{T}_{1:L} \leq \bar{t}_\alpha, \bar{T}_{(L+1):2L} \leq \bar{t}_\alpha) = \mathbb{E}\, \Phi^2\left(-\frac{\mu + \sqrt{\rho} Z_0 - \bar{t}_\alpha}{\sqrt{(1-\rho)/L}} \right).$$

Thus

$$\begin{aligned} \mathbb{P}\left(\phi_{\bar{T}_{1:L}} \neq \phi_{\bar{T}_{(L+1):2L}} \right) &= 1 - \int \left\{ \Phi^2(y) + \Phi^2(-y) \right\} \phi(x)\, \mathrm{d}x, \\ y &= \frac{\mu + \sqrt{\rho} z - \sqrt{1/L + \rho(L-1)/L}\, z_\alpha}{\sqrt{(1-\rho)/L}}, \end{aligned}$$

where $\phi(x)$ is the standard Gaussian density evaluated and x. $\mathbb{P}(\phi_{T^{(1)}} \neq \phi_{T^{(2)}})$ is given by the same expression with $L = 1$.

B Tail bound for the average statistic

Let $T^{(1)}, \ldots, T^{(L)}$ be exchangeable statistics with $T^{(1)} \sim F_0$ under the null. In this Appendix, we consider aggregation rules of the form

$$S := c\,\bar{T} = c\,(T^{(1)} + \cdots + T^{(L)})/L, \tag{22}$$

that guarantee finite sample type-I error control when a test is constructed by comparing S to the α quantile of F_0. No assumption is made on the copula of $(T^{(1)}, \ldots, T^{(L)})$. We specifically consider two cases: $F_0 = \mathrm{unif}(0,1)$ and $F_0 = \mathcal{N}(0,1)$. We summarise results from the literature on the former case, and also develop new results on the latter case.

B.1 The $F_0 = \mathrm{unif}(0,1)$ case

It is well-known that when each $T^{(l)}$ is a valid p-value, Eq. (22) with $c = 2$ is also a valid p-value; see Rüschendorf (1982); Meng (1994); Vovk and Wang (2021). It has been recently shown that the constant $c = 2$ cannot be improved under an exchangeability condition on $(T^{(1)}, \ldots, T^{(L)})$ (Choi and Kim, 2022).

B.2 The $F_0 = \mathcal{N}(0,1)$ case

Without loss of generality, suppose the null hypothesis is rejected whenever $S > \Phi(1-\alpha)$.

Theorem B.1 (Averaging for Z-statistics)**.** *Let $T^{(1)}, \ldots, T^{(L)}$ be exchangeable statistics with $T^{(1)} \sim \mathcal{N}(0,1)$ under the null. Then, for $0 < \alpha \le 1 - \Phi(1) \approx 0.159$, rejecting the null whenever $S = \bar{T}/2 > \Phi(1-\alpha)$ controls the type-I error below α.*

We will show Theorem B.1 using properties of convex ordering on random variables.

Definition 2 (Convex order)**.** For random variables X and Y, we say $X \le_{\mathrm{cx}} Y$ if $\mathbb{E}\, f(X) \le \mathbb{E}\, f(Y)$ for every convex f.

For a random variable X, let its integrated survival function be defined as

$$\pi_X(t) = \int_t^\infty \bar{F}_X(s)\,\mathrm{d}s = \int_t^\infty (1 - F_X(s))\,\mathrm{d}s = \mathbb{E}(X-t)_+.$$

It can be shown that $\pi_X(t)$ is decreasing and convex.

Lemma B.1 (Müller and Stoyan (2002, §1.5))**.** *$X \le_{\mathrm{cx}} Y$ if and only if $\pi_X(t) \le \pi_Y(t)$ for every t and $\mathbb{E}\, X = \mathbb{E}\, Y$.*

Lemma B.2 (Mills' ratio)**.** *For $\lambda > 0$, $\frac{\lambda}{\lambda^2+1}\phi(\lambda) < 1 - \Phi(\lambda) < \frac{1}{\lambda}\phi(\lambda)$.*

The next lemma derives a tail bound for a random variable dominated by a standard Gaussian in the convex order; see Meng (1994, Lemma 1) for a similar result on convex order with respect to a $\mathrm{unif}(0,1)$ random variable.

Lemma B.3. *Suppose $\bar{Z} \le_{\mathrm{cx}} Z$ for $Z \sim \mathcal{N}(0,1)$. Then we have*

$$\mathbb{P}(\bar{Z} > s) = \bar{F}_{\bar{Z}}(s) < \frac{1 - \Phi(\beta)}{\beta(s-\beta)}, \quad 0 < \beta < s.$$

Proof. Because $\pi_{\bar{Z}}$ is convex, we have

$$\pi_{\bar{Z}}(\beta) \geq \pi(s) + \pi'_{\bar{Z}}(s)(\beta - s) = \pi_{\bar{Z}}(s) + \bar{F}_{\bar{Z}}(s)(s - \beta),$$

which gives

$$\bar{F}_{\bar{Z}}(s) \leq \frac{\pi_{\bar{Z}}(\beta) - \pi_{\bar{Z}}(s)}{s - \beta}, \quad \beta < s.$$

By $\bar{Z} \leq_{\text{cx}} Z$ and Lemma B.1, we have $\pi_{\bar{Z}}(\beta) \leq \pi_Z(\beta)$ and it follows that

$$\bar{F}_{\bar{Z}}(s) \leq \frac{\pi_Z(\beta) - \pi_{\bar{Z}}(s)}{s - \beta}, \quad \beta < s.$$

Now we bound $\pi_Z(\beta)$ for $\beta > 0$. Using Lemma B.2, we have

$$\begin{aligned}
\pi_Z(\beta) < \int_\beta^{+\infty} t\phi(t)\,\mathrm{d}t = \int_\beta^{+\infty} t\,\mathrm{d}\Phi(t) &= [t\Phi(t)]|_\beta^{+\infty} - \int_\beta^{+\infty} \Phi(t)\,\mathrm{d}(1/t) \\
&= -\Phi(\beta)/\beta + \int_\beta^{+\infty} \Phi(t)t^{-2}\,\mathrm{d}t \\
&\leq -\Phi(\beta)/\beta + \int_\beta^{+\infty} t^{-2}\,\mathrm{d}t = \frac{1 - \Phi(\beta)}{\beta}.
\end{aligned}$$

Hence, for $0 < \beta < s$, we have

$$\bar{F}_{\bar{Z}}(s) < \frac{(1 - \Phi(\beta))/\beta - \pi_{\bar{Z}}(s)}{s - \beta} \leq \frac{1 - \Phi(\beta)}{\beta(s - \beta)}.$$

□

Proof of Theorem B.1. We first claim that $\bar{T} \leq_{\text{cx}} T^{(1)} \sim \mathcal{N}(0, 1)$. Indeed, given any convex function f, we have

$$\mathbb{E}\, f(\bar{T}) \leq \mathbb{E}\, L^{-1} \sum_{l=1}^{L} f(T^{(l)}) = \mathbb{E}\, f(T^{(1)})$$

if $\mathbb{E}\, f(T^{(1)}) < \infty$, where we used Jensen's inequality and exchangeability. Now, take $\beta = s/2$ for $s \geq 2$ in Lemma B.3. We have

$$\mathbb{P}(\bar{T}/2 > s/2) = \mathbb{P}(\bar{T} > s) < \frac{1 - \Phi(s/2)}{(s/2)^2} \leq 1 - \Phi(s/2).$$

□

From our experience, the aggregation rule Eq. (22) with $c = 1/2$ is often quite conservative; see, for example, Fig. 4. It is sometimes argued that using $c = 1$ seems to work in practice (Wang and Shah, 2020; Lundborg et al., 2022). Proposition B.1 below shows that in certain cases using $c = 1$ can lead to type-I error inflated by a factor of two. Further, Proposition B.2 shows that, when α is small, the actual type-I error is at most $e\alpha$. However, determining the optimal c, which lies between 1/2 and 1, remains an open question.

Proposition B.1. *Under the assumption of Theorem B.1, using $S = \bar{T}$ can lead to size 2α when L is even.*

Proof. Let $0 \le \alpha \le 1/2$ and consider Fig. B.1. Let $U_1 \sim \text{unif}(0,1)$ and let

$$U_2 = \begin{cases} U_1, & U_1 \le 1-2\alpha \\ 2-2\alpha-U_1, & U_1 > 1-2\alpha \end{cases}.$$

Pair U_1, U_2 are exchangeable with uniform margin. Then, let $T^{(1)} = \Phi^{-1}(U_1)$ and $T^{(2)} = \Phi^{-1}(U_2)$. It follows that $T^{(1)}, T^{(2)}$ are exchangeable with standard normal margin. Further, with probability 2α, the graph of $(T^{(1)}, T^{(2)})$ is the red curve, on which the minimum value of $\bar{T}$ is $\Phi^{-1}(1-\alpha)$, when $U_1 = U_2 = 1-\alpha$. Hence,

$$\mathbb{P}(\bar{T} > \Phi^{-1}(1-\alpha)) = 2\alpha,$$

or equivalently,

$$\mathbb{P}(\bar{T} > t) = 2(1-\Phi(t)).$$

The construction is specific to t for $t > 0$.

It is easy to extend this construction for $L = 2k$, $k \ge 1$. To construct $T^{(1)}, \dots, T^{(2k)}$, randomly choose half of them to be equal to $\Phi^{-1}(U_1)$ and the rest to be equal to $\Phi^{-1}(U_2)$. Observe that $T^{(1)}, \dots, T^{(2k)}$ are exchangeable with standard normal margin. Further, $\bar{T} = (\Phi^{-1}(U_1) + \Phi^{-1}(U_2))/2$ so the bound remains valid. □

Proposition B.2 (Tight tail bound up to a constant)**.** *Given $\bar{Z} \le_{\text{cx}} \mathcal{N}(0,1)$, it holds that*

$$\mathbb{P}(\bar{Z} > s) < \frac{1-\Phi(s-1/s)}{1-1/s^2}, \quad s > 1,$$

where RHS $\sim e(1-\Phi(s))$ *as* $s \to \infty$.

Proof. For $s > 1$, choose $\beta = s - 1/s > 0$ in Lemma B.3 and we get

$$\mathbb{P}(\bar{Z} > s) < \frac{1-\Phi(s-1/s)}{1-1/s^2}.$$

By L'Hôpital's rule, it is easy to see that

$$\lim_{s\to+\infty} \frac{1-\Phi(s-1/s)}{1-1/s^2} / (1-\Phi(s)) = \lim_{s\to+\infty} \frac{\phi(s-1/s)}{\phi(s)} = e.$$

□

Remark B.1. This means for small α, comparing $\bar{Z}$ to a standard normal inflates size by e.

C Proofs of theoretical results

In this Appendix, we prove results for a generic $m = m(n)$ sequence such that $1 < m < n$, $m \nearrow \infty$ and $n/m \to \infty$, although we choose $m = \lfloor n/\log n \rfloor$ throughout the paper. We write B_n in lieu of B to highlight its dependence on n.

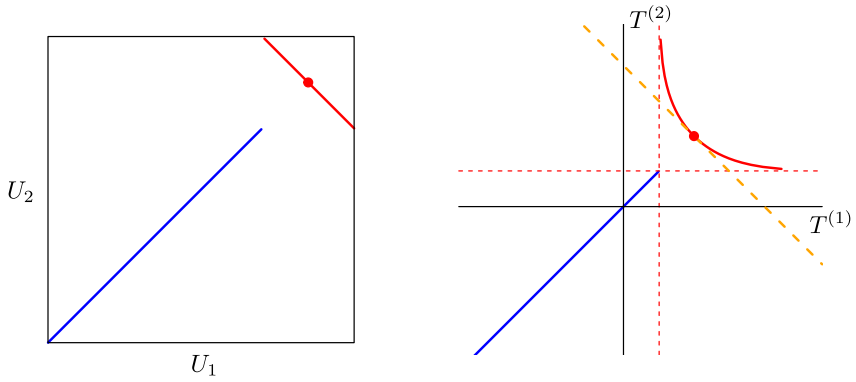

Figure B.1: Construction for exchangeable $(T^{(1)}, T^{(2)})$ with standard normal margin such that $\mathbb{P}(\bar{T} > t) = 2(1 - \Phi(t))$.

C.1 Finite-sample bounds of subsampling

The standard consistency result is given by Politis et al. (1999, Theorem 2.2.1). The following is an extension that allows multiple exchangeable statistics. Throughout, let $\widehat{F}_n := \mathbb{F}_{\widehat{\mathbf{H}}}$, where $\widehat{\mathbf{H}}$ is given by Eq. (3). Let $F_{n,P}$ be the distribution function of $T_n^{(1)}$.

Lemma C.1 (Uniform Hoeffding bound)**.** *Let $T_n^{(1)}, \dots, T_n^{(L)}$ be exchangeable statistics under $(X, \Omega) \sim P^n \times P_\Omega$. We have $\mathbb{E}\,\widehat{F}_n(x) = F_{m,P}(x)$ for every x and*

$$\mathbb{P}(\|\widehat{F}_n - F_{m,P}\|_\infty > t) \le 2\exp(-2\lfloor n/m \rfloor t^2/c_0^2), \quad t \ge 0,$$

where c_0 is a universal constant.

Proof. By the construction of $\mathcal{B}$ via Algorithm 1 and noting that $|\mathcal{B}| = B = J\lfloor n/m \rfloor$, we have

$$\widehat{F}_n(x) - F_{m,P}(x) = \frac{1}{JL} \sum_{l=1}^{L} \sum_{\pi \in \{\pi_1, \dots, \pi_J\}} \left[\widehat{F}_{\pi,n}^{(l)}(x) - F_{m,P}(x) \right], \tag{23}$$

where

$$\widehat{F}_{\pi,n}^{(l)}(x) := \frac{1}{\lfloor n/m \rfloor} \Bigg[\mathbb{I}\left\{ T_m^{(l)}(X_{\pi(1)}, \dots, X_{\pi(m)}) \le x \right\} +$$
$$\mathbb{I}\left\{ T_m^{(l)}(X_{\pi(m+1)}, \dots, X_{\pi(2m)}) \le x \right\} + \cdots + \mathbb{I}\left\{ T_m^{(l)}(X_{\pi((\lfloor n/m \rfloor - 1)m+1)}, \dots, X_{\pi(\lfloor n/m \rfloor m)}) \le x \right\} \Bigg].$$

The summands in the square brackets correspond to $\lfloor n/m \rfloor$ disjoint subsamples of data

$$(X_{\pi(1)}, \dots, X_{\pi(m)}), \quad (X_{\pi(m+1)}, \dots, X_{\pi(2m)}), \quad \dots \quad , (X_{\pi((\lfloor n/m \rfloor - 1)m+1)}, \dots, X_{\pi(\lfloor n/m \rfloor m)})$$

and they are therefore independent.

Fix π and l. By definition, we know $\mathbb{E}\,\widehat{F}_{\pi,n}^{(l)}(x) = F_{m,P}(x)$ and hence $\mathbb{E}\,\widehat{F}_n(x) = F_{m,P}(x)$. In view of $\widehat{F}_{\pi,n}^{(l)}$ as an empirical distribution function, the inequality due to Dvoretzky et al. (1956) and Massart (1990) states that

$$\mathbb{P}\left(\|\widehat{F}_{\pi,n}^{(l)} - F_{m,P}\|_\infty > t \right) \le 2\exp\left(-2\lfloor n/m \rfloor t^2\right), \quad t \ge 0.$$

This inequality does not require $F_{m,P}$ to be continuous; see Massart (1990, Comment 2(iii)). The inequality implies that $\|\widehat{F}_{\pi,n}^{(l)} - F_{m,P}\|_\infty$ is $c_0/(2\sqrt{\lfloor n/m \rfloor})$-sub-Gaussian for a universal constant $c_0 > 0$ (Vershynin, 2018, Proposition 2.5.2). From Eq. (23), it follows that

$$\|\widehat{F}_n - F_{m,P}\|_\infty \leq \frac{1}{JL} \sum_{l=1}^{L} \sum_{\pi \in \{\pi_1, \ldots, \pi_J\}} \|\widehat{F}_{\pi,n}^{(l)} - F_{m,P}\|_\infty,$$

where we observe that every $\|\widehat{F}_{\pi,n}^{(l)} - F_{m,P}\|_\infty$ is identically distributed. By Jensen's inequality, we have that the upper bound above is also $c_0/(2\sqrt{\lfloor n/m \rfloor})$-sub-Gaussian and hence so is $\|\widehat{F}_n - F_{m,P}\|_\infty$, from which the result follows. □

Recall that $\widehat{G}_n(x) := \mathbb{F}_{\{\widehat{S}_b\}}(x)$ is the natural subsampling estimation of G_P, the asymptotic distribution function of S_n. For $b = 1, \ldots, B$, $\widehat{S}_b$ is the result of applying S to the b-th row of $\widehat{\mathbf{H}}$. Let $G_{n,P}$ be the distribution function of S_n. We define

$$K_n := \sqrt{\frac{\lfloor n/m \rfloor}{\log(n/m)}}, \tag{24}$$

which is a positive sequence that tends to infinity.

Lemma C.2 (High probability bound on subsampling approximation)**.** *Let $T_n^{(1)}, \ldots, T_n^{(L)}$ be exchangeable statistics under $(X, \Omega) \sim P^n \times P_\Omega$. Given any distribution functions F and G, with probability at least $1 - 4(n/m)^{-2/c_0^2}$ (where c_0 is the universal constant from Lemma C.1), it holds that*

$$\|\widehat{F}_n - F\|_\infty \leq 1/K_n + \|F_{m,P} - F\|_\infty,$$
$$\|\widehat{G}_n - G\|_\infty \leq 1/K_n + \|G_{m,P} - G\|_\infty.$$

Proof. Taking $t = 1/K_n$ in Lemma C.1, by the triangle inequality, the first inequality above holds with probability at least $1 - 2(n/m)^{-2/c_0^2}$. By the definition of $\widehat{G}_n$, a tail bound of the same type as Lemma C.1 holds for $\|\widehat{G}_n - G_{m,P}\|_\infty$. Hence, the second inequality above also holds with probability at least $1 - 2(n/m)^{-2/c_0^2}$. The result follows from a union bound. □

Corollary C.1. *Let $T_n^{(1)}, \ldots, T_n^{(L)}$ be exchangeable statistics under $(X, \Omega) \sim P^n \times P_\Omega$ for $P \in \mathcal{P}_0$. Under Condition 1 and Assumption 1, we have*

$$\|\widehat{F}_n - F_0\|_\infty \to_p 0, \quad \|\widehat{G}_n - G_P\|_\infty \to_p 0$$

as $n \to \infty$.

Proof. Take $F = F_0$ and $G = G_P$ in Lemma C.2. Because F_0 and G_P are continuous, by van der Vaart (2000, Lem. 2.11) Condition 1 implies $\|F_{m,P} - F_0\|_\infty \to 0$ and Assumption 1 implies $\|G_{m,P} - G_P\|_\infty \to 0$. Therefore, with probability tending to one, the bounds in Lemma C.2 tend to zero, which yields the result. □

C.2 Rank transform

In this section, we show that the rank transform does not affect the validity of subsampling approximation under the null. Recall that $\widetilde{G}_n$ is the empirical distribution of $\{\widetilde{S}_b\}$, where $\widetilde{S}_b$ is the result of applying aggregation S to the b-th row of the rank-transformed matrix $\widetilde{\mathbf{H}}$.

Lemma C.3. *Given $x_1, \dots, x_n \in \mathbb{R}$, $x'_1, \dots, x'_n \in \mathbb{R}$ and $\varepsilon > 0$ such that $|x_i - x'_i| \le \varepsilon$ for all i, for every $x \in R$ it holds that*

$$|\mathbb{F}_{\{x_i\}}(x) - \mathbb{F}_{\{x'_i\}}(x)| \le \mathbb{F}_{\{x_i\}}(x+\varepsilon) - \mathbb{F}_{\{x_i\}}(x-\varepsilon).$$

Proof. As $|\mathbb{F}_{\{x_i\}}(x) - \mathbb{F}_{\{x'_i\}}(x)| = n^{-1}\sum_{i=1}^n (\mathbb{I}\{x_i \le x\} - \mathbb{I}\{x'_i \le x\})$, we know $|\mathbb{F}_{\{x_i\}}(x) - \mathbb{F}_{\{x'_i\}}(x)|$ is bounded above by the fraction of i such that x is between x_i and x'_i. Due to the fact that $|x_i - x'_i| \le \varepsilon$, this fraction is further bounded above by the fraction of x_i such that $x \in [x_i - \varepsilon, x_i + \varepsilon)$, or equivalently, $x_i \in (x - \varepsilon, x + \varepsilon]$, which is the right-hand side of the display above. □

Lemma C.4 (High probability bound on rank-transformed subsampling)**.** *Let $T_n^{(1)}, \dots, T_n^{(L)}$ be exchangeable statistics under $(X, \Omega) \sim P^n \times P_\Omega$ for $P \in \mathcal{P}_0$. Suppose S is chosen such that Condition 2 holds. Let*

$$\begin{aligned}\varepsilon_n^{F,P} &:= 1/K_n + \|F_{m,P} - F_0\|_\infty + 1/(2\,B_n\,L),\\ \varepsilon_n^{G,P} &:= 1/K_n + \|G_{m,P} - G_P\|_\infty,\end{aligned}$$

where $B_n = J\lfloor n/m \rfloor$. Then, under Condition 1 and Assumption 1, we have the following results.

(1) For $F_0 = \mathrm{unif}(0,1)$, with probability tending to one, it holds that

$$\|\widetilde{G}_n - G_P\|_\infty \le 2\, g_{\max,P}\, \varepsilon_n^{F,P} + 3\,\varepsilon_n^{G,P}.$$

(2) For $F_0 = \mathcal{N}(0,1)$, with probability tending to one, it holds that

$$\|\widetilde{G}_n - G_P\|_\infty \le \frac{2\, g_{\max,P}}{|\Phi^{-1}(\varepsilon_n^{F,P})|} + 4L\,\varepsilon_n^{F,P} + 3\,\varepsilon_n^{G,P} + \frac{1}{B_n}.$$

Proof. In what follows, let $\mathcal{A}_n$ be the event that the bounds in Lemma C.2 hold with $F = F_0$ and $G = G_P$. Note that $\mathbb{P}_P(\mathcal{A}_n) \to 1$.

(1) Under $F_0 = \mathrm{unif}(0,1)$, the rank transform Eq. (4) becomes

$$\widetilde{H}_{b,l} = \widehat{F}_n(\widehat{H}_{b,l}) - \frac{1}{2B_nL}, \quad b = 1, \dots, B_n, \quad l = 1, \dots, L.$$

On $\mathcal{A}_n$, we have

$$\sup_{x \in [0,1]} |\widehat{F}_n(x) - x| \le 1/K_n + \|F_{m,P} - F_0\|_\infty$$

and hence, by the rank transform,

$$\max_{b,l} |\widetilde{H}_{b,l} - \widehat{H}_{b,l}| \le \varepsilon_n^{F,P}.$$

By Condition 2, it implies

$$\max_b |\widetilde{S}_b - \widehat{S}_b| \le \varepsilon_n^{F,P}.$$

Recall that on $\mathcal{A}_n$, $\|\widehat{G}_n - G_P\|_\infty \le \varepsilon_n^{G,P}$. Note that $\widetilde{G}_n$ and $\widehat{G}_n$ are respectively the distribution functions of $\{\widetilde{S}_b\}$ and $\{\widehat{S}_b\}$. Applying Lemma C.3, we have that for every x,

$$\begin{aligned}|\widetilde{G}_n(x) - \widehat{G}_n(x)| &\le \widehat{G}_n(x + \varepsilon_n^{F,P}) - \widehat{G}_n(x - \varepsilon_n^{F,P})\\ &\le G_P(x + \varepsilon_n^{F,P}) - G_P(x - \varepsilon_n^{F,P}) + 2\varepsilon_n^{G,P}\\ &\le 2 g_{\max,P}\,\varepsilon_n^{F,P} + 2\,\varepsilon_n^{G,P} \quad \text{on } \mathcal{A}_n,\end{aligned}$$

where the last step follows from Assumption 1 and the mean value theorem. Note that the final bound above does not depend on x. Consequently, on $\mathcal{A}_n$, we have

$$\begin{aligned}\|\widetilde{G}_n - G_P\|_\infty &\le \|\widetilde{G}_n - \widehat{G}_n\|_\infty + \|\widehat{G}_n - G_P\|_\infty \\ &\le 2\, g_{\max,P}\, \varepsilon_n^{F,P} + 3\, \varepsilon_n^{G,P},\end{aligned}$$

which occurs with probability tending to one.

(2) Under $F_0 = \Phi$, the rank transform in Eq. (4) becomes

$$\widetilde{H}_{b,l} = \Phi^{-1}\left(\widehat{F}_n(\widehat{H}_{b,l}) - 1/(2B_nL)\right).$$

By a Taylor expansion of the outer function at $\Phi(\widehat{H}_{b,l})$,

$$\widetilde{H}_{b,l} = \widehat{H}_{b,l} + \frac{1}{\phi(\Phi^{-1}(\xi))}\left[\widehat{F}_n(\widehat{H}_{b,l}) - 1/(2B_nL) - \Phi(\widehat{H}_{b,l})\right] \tag{25}$$

for some ξ between $\widehat{F}_n(\widehat{H}_{b,l}) - 1/(2B_nL)$ and $\Phi(\widehat{H}_{b,l})$. Let

$$\Gamma_\varepsilon := \left\{(b,l) : \widehat{F}_n(\widehat{H}_{b,l}) - 1/(2B_nL) \in [0, 2\varepsilon_n^{F,P}) \cup (1 - 2\varepsilon_n^{F,P}, 1]\right\}.$$

Because $\{\widehat{F}_n(\widehat{H}_{b,l})\}$ is the set of normalised ranks, we have

$$|\Gamma_\varepsilon| \le 4\varepsilon_n^{F,P} B_n L + 1. \tag{26}$$

Observe that $\|\widehat{F}_n - 1/(2B_nL) - \Phi\|_\infty \le \varepsilon_n^{F,P}$ on $\mathcal{A}_n$. Then on $\mathcal{A}_n$, for every $(b,l) \in \Gamma_\varepsilon^c$, we know ξ in Eq. (25) is in $[\varepsilon_n^{F,P}, 1 - \varepsilon_n^{F,P}]$. By Mills' ratio (Lemma B.2), we have

$$\phi(\Phi^{-1}(\xi)) \ge \phi(\Phi^{-1}(1 - \varepsilon_{F,P})) > \varepsilon_n^{F,P} |\Phi^{-1}(\varepsilon_n^{F,P})|.$$

It then follows from Eq. (25) that, on $\mathcal{A}_n$,

$$|\widetilde{H}_{b,l} - \widehat{H}_{b,l}| \le \frac{1}{|\Phi^{-1}(\varepsilon_n^{F,P})|}, \quad (b,l) \in \Gamma_\varepsilon^c. \tag{27}$$

Define

$$\Xi_\varepsilon := \{b : (b,l) \in \Gamma_\varepsilon \text{ for some } l \in \{1, \dots, L\}\}.$$

By Condition 2, on $\mathcal{A}_n$, Eq. (27) implies

$$|\widetilde{S}_b - \widehat{S}_b| \le \delta_n := \frac{1}{|\Phi^{-1}(\varepsilon_n^{F,P})|}, \quad b \in \Xi_\varepsilon^c. \tag{28}$$

In light of $|\Xi_\varepsilon| \le |\Gamma_\varepsilon|$ and Eq. (26), observe that

$$\rho_\varepsilon := |\Xi_\varepsilon|/B_n \le 4\varepsilon_n^{F,P} L + 1/B_n. \tag{29}$$

We now bound the distance between the empirical distributions of the rank-transformed and untransformed statistics. We have

$$\begin{aligned}\|\widetilde{G}_n - \widehat{G}_n\|_\infty &= \left\|\mathbb{F}_{\{\widetilde{S}_b\}} - \mathbb{F}_{\{\widehat{S}_b\}}\right\|_\infty \\ &= \left\|(1-\rho_\varepsilon)\mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon^c}} + \rho_\varepsilon \mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon}} - (1-\rho_\varepsilon)\mathbb{F}_{\widehat{S}_{\Xi_\varepsilon^c}} - \rho_\varepsilon \mathbb{F}_{\widehat{S}_{\Xi_\varepsilon}}\right\|_\infty \\ &\le (1-\rho_\varepsilon)\left\|\mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon^c}} - \mathbb{F}_{\widehat{S}_{\Xi_\varepsilon^c}}\right\|_\infty + \rho_\varepsilon \left\|\mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon}} - \mathbb{F}_{\widehat{S}_{\Xi_\varepsilon}}\right\|_\infty \\ &\le (1-\rho_\varepsilon)\left\|\mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon^c}} - \mathbb{F}_{\widehat{S}_{\Xi_\varepsilon^c}}\right\|_\infty + \rho_\varepsilon,\end{aligned} \tag{30}$$

where the last step uses the fact that Kolmogorov distance is bounded by one. Given Eq. (28), by Lemma C.3, at any x, we have

$$\begin{aligned}\left|\mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon^c}}(x) - \mathbb{F}_{\widehat{S}_{\Xi_\varepsilon^c}}(x)\right| &\le \frac{\left|\{b \in \Xi_\varepsilon^c : \widehat{S}_b \in (x-\delta_n, x+\delta_n]\right|}{|\Xi_\varepsilon^c|} \\ &\le \frac{\left|\{b : \widehat{S}_b \in (x-\delta_n, x+\delta_n]\right|}{|\Xi_\varepsilon^c|} \\ &= \frac{\left|\{b : \widehat{S}_b \in (x-\delta_n, x+\delta_n]\right|}{B(1-\rho_\varepsilon)} \\ &= \frac{1}{1-\rho_\varepsilon}\left[\widehat{G}_{m,n,P}(x+\delta_n) - \widehat{G}_{m,n,P}(x-\delta_n)\right],\end{aligned}$$

where ρ_ε is defined in Eq. (29). Further, on $\mathcal{A}_n$, we also have $\|\widehat{G}_n - G_P\|_\infty \le \varepsilon_n^{G,P}$ and hence

$$\left|\mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon^c}}(x) - \mathbb{F}_{\widehat{S}_{\Xi_\varepsilon^c}}(x)\right| \le \frac{1}{1-\rho_\varepsilon}\left[G_P(x+\delta_n) - G_P(x-\delta_n) + 2\varepsilon_n^{G,P}\right].$$

Taking supremum over x and using Assumption 1, we have

$$\left\|\mathbb{F}_{\widetilde{S}_{\Xi_\varepsilon^c}} - \mathbb{F}_{\widehat{S}_{\Xi_\varepsilon^c}}\right\|_\infty \le \frac{1}{1-\rho_\varepsilon}(2g_{\max,P}\delta_n + 2\varepsilon_n^{G,P}).$$

Hence, by substituting the above, Eq. (28) and (29) into Eq. (30), we arrive at

$$\begin{aligned}\|\widetilde{G}_n - \widehat{G}_n\|_\infty &\le 2g_{\max,P}\delta_n + 2\varepsilon_n^{G,P} + 4L\varepsilon_n^{F,P} + 1/B_n \\ &= \frac{2g_{\max,P}}{|\Phi^{-1}(\varepsilon_n^{F,P})|} + 2\varepsilon_n^{G,P} + 4L\varepsilon_n^{F,P} + \frac{1}{B_n} \quad \text{on } \mathcal{A}_n.\end{aligned}$$

Finally, combining the above and $\|\widehat{G}_n - G_P\|_\infty \le \varepsilon_n^{G,P}$, we conclude that on $\mathcal{A}_n$,

$$\|\widetilde{G}_n - G_P\|_\infty \le \frac{2g_{\max,P}}{|\Phi^{-1}(\varepsilon_n^{F,P})|} + 4L\varepsilon_n^{F,P} + 3\varepsilon_n^{G,P} + 1/B_n,$$

whose probability tends to one.

□

Remark C.1. The normal case needs a separate proof because when S is Lipschitz, $(t_1, \ldots, t_L) \mapsto S(\Phi^{-1}(t_1), \ldots, \Phi^{-1}(t_L))$ need not be Lipschitz.

C.3 Proof of Theorem 1

Proof. Observe that with $m = \lfloor n/\log n \rfloor$ and $B_n = J\lfloor n/m \rfloor$, the bounds on $\|\widetilde{G}_n - G_P\|_\infty$ in Lemma C.4 tend to zero as $n \to \infty$. We have $\|\widetilde{G}_n - G_P\|_\infty \to_p 0$.

(i) This follows from the proof of (ii) by taking $\mathcal{P}_0 = \{P\}$.

(ii) Note first that by Lemma C.8 and van der Vaart (2000, Lem. 2.11), we have

$$\sup_{P \in \mathcal{P}_0} \|F_{n,P} - F_0\|_\infty \to 0, \quad \sup_{P \in \mathcal{P}_0} \|G_{n,P} - G_P\|_\infty \to 0.$$

By Lemma C.4 and the given assumptions, there exists a sequence $\gamma_n \to 0$ that does not depend on P, such that

$$\sup_{P\in\mathcal{P}_0} \varepsilon_n^{F,P} \le \gamma_n, \quad \sup_{P\in\mathcal{P}_0} \varepsilon_n^{G,P} \le \gamma_n,$$

where $\varepsilon_n^{F,P}$ and $\varepsilon_n^{G,P}$ are defined in Lemma C.4. Also, because the probability in Lemma C.2 does not depend on P, there exists sequences $\lambda_n \to 0$ and $\delta_n \to 0$, which do not depend on P, such that for every $P \in \mathcal{P}_0$,

$$\mathcal{D}_{n,P} := \{\|\widetilde{G}_n - G_P\|_\infty < \lambda_n\}$$

occurs with probability at least $1-\delta_n$. Let $\mathcal{B}_n := \left\{S_n > \widetilde{G}_n^{-1}(1-\alpha)\right\}$.

We first show that $\limsup_n \sup_{P\in\mathcal{P}_0} \mathbb{P}_P(\mathcal{B}_n) \le \alpha$. Observe that

$$1-\alpha \le \widetilde{G}_n\left(\widetilde{G}_n^{-1}(1-\alpha)\right) \le G_P\left(\widetilde{G}_n^{-1}(1-\alpha)\right) + \|\widetilde{G}_n - G_P\|_\infty$$

using Lemma C.6, which implies

$$G_P\left(\widetilde{G}_n^{-1}(1-\alpha)\right) \ge 1-\alpha-\|\widetilde{G}_n - G_P\|_\infty. \tag{31}$$

It follows that on $\mathcal{B}_n$,

$$G_P(S_n) \ge G_P\left(\widetilde{G}_n^{-1}(1-\alpha)\right) \ge 1-\alpha-\|\widetilde{G}_n - G_P\|_\infty.$$

Hence on $\mathcal{B}_n \cap \mathcal{D}_n$,

$$S_n > G_P^{-1}(1-\alpha-\lambda_n).$$

We have

$$\begin{aligned}
\mathbb{P}_P(\mathcal{B}_n) &\le \mathbb{P}_P(\mathcal{B}_n \cap \mathcal{D}_n) + \mathbb{P}_P(\mathcal{D}_n^c) \\
&\le \mathbb{P}_P\left\{S_n > G_P^{-1}(1-\alpha-\lambda_n)\right\} + \delta_n \\
&= 1 - G_{n,P}\left(G_P^{-1}(1-\alpha-\lambda_n)\right) + \delta_n \\
&\le 1 - G_P\left(G_P^{-1}(1-\alpha-\lambda_n)\right) + \|G_P - G_{n,P}\|_\infty + \delta_n \\
\text{(by Lemma C.6)} \quad &\le \alpha + \lambda_n + \|G_P - G_{n,P}\|_\infty + \delta_n,
\end{aligned}$$

where $G_{n,P}$ is the distribution function of S_n. Because λ_n and δ_n do not depend on P, it follows that

$$\sup_{P\in\mathcal{P}_0} \mathbb{P}_P(\mathcal{B}_n) \le \alpha + \lambda_n + \sup_{P\in\mathcal{P}_0} \|G_P - G_{n,P}\|_\infty + \delta_n,$$

and hence $\limsup_n \sup_{P\in\mathcal{P}_0} \mathbb{P}_P(\mathcal{B}_n) \le \alpha$ because $\lambda_n \to 0$, $\delta_n \to 0$ and $\sup_{P\in\mathcal{P}_0} \|G_P - G_{n,P}\|_\infty \to 0$.

We now show that $\liminf_n \inf_{P\in\mathcal{P}_0} \mathbb{P}_P(\mathcal{B}_n) \ge \alpha$, which will yield the final conclusion. Let

$$y_{n,P} := \sup\{x : G_P(x) \le 1-\alpha+\lambda_n + 1/B_n\} \tag{32}$$

and consider $\mathcal{F}_{n,P} := \{S_n > y_{n,P}\}$. Observe that $\mathcal{F}_{n,P}$ implies $G_P(S_n) - \lambda_n > 1-\alpha+1/B$. Then, $\mathcal{F}_{n,P} \cap \mathcal{D}_{n,P}$ implies $\widetilde{G}_n(S_n) > 1-\alpha+1/B_n$, which further leads to

$$S_n > \widetilde{G}_n^{-1}(1-\alpha)$$

because $\widetilde{G}_n$ is an empirical measure with B atoms. Hence, we have $\mathcal{F}_{n,P} \cap \mathcal{D}_{n,P} \subseteq \mathcal{B}_n \cap \mathcal{D}_{n,P}$ and therefore

$$\begin{aligned}
\mathbb{P}_P(\mathcal{B}_n) &\geq \mathbb{P}_P(\mathcal{B}_n \cap \mathcal{D}_{n,P}) \\
&\geq \mathbb{P}_P(\mathcal{F}_{n,P} \cap \mathcal{D}_{n,P}) \\
&\geq \mathbb{P}_P(\mathcal{F}_{n,P}) - \mathbb{P}_P(\mathcal{D}_{n,P}^c) \\
&= 1 - G_{n,P}(y_{n,P}) - \mathbb{P}_P(\mathcal{D}_{n,P}^c) \\
&\geq 1 - G_{n,P}(y_{n,P}) - \delta_n \\
&\geq 1 - G_P(y_{n,P}) - \|G_P - G_{n,P}\|_\infty - \delta_n \\
&\overset{(i)}{=} 1 - (1 - \alpha + \lambda_n + 1/B_n) - \|G_P - G_{n,P}\|_\infty - \delta_n \\
&= \alpha - \lambda_n - 1/B_n - \|G_P - G_{n,P}\|_\infty - \delta_n,
\end{aligned}$$

where (i) uses Eq. (32) and the continuity of G_P. Because sequences λ_n, B and δ_n do not depend on P, it then follows that

$$\inf_{P \in \mathcal{P}_0} \mathbb{P}_P(\mathcal{B}_n) \geq \alpha - \lambda_n - 1/B - \sup_{P \in \mathcal{P}_0} \|G_P - G_{n,P}\|_\infty - \varepsilon_n,$$

and hence $\liminf_n \inf_{P \in \mathcal{P}_0} \mathbb{P}_P(\mathcal{B}_n) \geq \alpha$ because $\lambda_n \to 0$, $B \to \infty$, $\delta_n \to 0$ and $\sup_{P \in \mathcal{P}_0} \|G_P - G_{n,P}\|_\infty \to 0$.

□

Remark C.2. The proof above needs no assumption on G_P being strictly increasing at $G_P^{-1}(1-\alpha)$; see also Politis et al. (1999, Remark 1.2.1).

C.4 Proof of Theorem 2

We first prove the following lemma.

Lemma C.5. *Under the assumptions of Theorem 2 and Assumption 2, it holds that* $\|\widetilde{Q}_n - Q_P\|_\infty \to_p 0$, *where* Q_P *is the distribution function of* $\max\{G_P^1(S^1), \ldots, G_P^W(S^W)\}$.

Proof. We show the result by relating R_n to

$$R_n^0 := \max(G_P^1(S_n^1), \ldots, G_P^W(S_n^1))$$

and relating $\widetilde{Q}_n := \mathbb{F}_{\{\widetilde{R}_b\}}$ to the subsampling estimate $\widehat{Q}_n := \mathbb{F}_{\{\widehat{R}_b\}}$ with $\{\widehat{R}_b\}$ computed as follows. Applying $S^1, \ldots, S^W$ to the b-th row of the untransformed matrix $\widehat{\mathbf{H}}$, we get

$$\widehat{S}_b^1 := S^1(\widehat{H}_{b,1}, \ldots, \widehat{H}_{b,L}), \quad \ldots, \quad \widehat{S}_b^W := S^W(\widehat{H}_{b,1}, \ldots, \widehat{H}_{b,L}).$$

Recall that for $w = 1, \ldots, W$, G_P^w is the asymptotic distribution function of S_n^w. We define

$$\widehat{R}_b := \max\left\{G_P^1(\widehat{S}_b^1), \ldots, G_P^W(\widehat{S}_b^W)\right\}, \quad b = 1, \ldots, B_n;$$

compare this with the definition of $\widetilde{R}_b$ in Algorithm 3.

First note that Q_P has a density that is bounded by W. Indeed, fix $1 \geq t > u \geq 0$ and set $\mathcal{T}_w := \{G_P^w(S_n^w) \leq t\}$ and $\mathcal{U}_w := \{G_P^w(S_n^w) \leq u\}$ for $w = 1, \ldots, W$. Then

$$\begin{aligned}
Q_P(t) - Q_P(u) &= \mathbb{P}_P\left(\cap_w \mathcal{T}_w\right) - \mathbb{P}_P(\cap_w \mathcal{U}_w) \\
&\leq \mathbb{P}_P\left\{(\cap_w \mathcal{T}_w) \setminus (\cap_w \mathcal{U}_w)\right\}.
\end{aligned}$$

Now

$$\begin{aligned}(\cap_w \mathcal{T}_w) \setminus (\cap_w \mathcal{U}_w) &= \cap_w \mathcal{T}_w \cap (\cap_w \mathcal{U}_w)^c \\ &= \cap_w \mathcal{T}_w \cap (\cup_w \mathcal{U}_w^c) \\ &= \cup_w (\cap_w \mathcal{T}_w \cap \mathcal{U}_w^c) \\ &\subseteq \cup_w \{u < G_P^w(S_n^w) \le t\}.\end{aligned}$$

Thus by a union bound,

$$Q_P(t) - Q_P(u) \le W(t-u),$$

proving the claim.

By the joint stability in Assumption 2, the distribution of R_n^0 converges to Q_P. By definition, $\widehat{Q}_n$ is the natural subsampling estimate of Q_P. Let

$$\mathcal{Q}_{n,P} := \left\{ \|\widehat{Q}_n - Q_P\|_\infty \le 1/K_n + \|Q_{n,P} - Q_P\|_\infty \right\},$$

where K_n is defined in Eq. (24) and $Q_{n,P}$ is the distribution function of R_n^0. By Assumption 2, $\|Q_{n,P} - Q_P\|_\infty \to 0$. By an argument similar to Lemma C.2, we know $\mathbb{P}_P(\mathcal{Q}_{n,P}) \to 1$. In what follows, let $\mathcal{E}_n^w$ be the event that the corresponding bound in Lemma C.4 on $\|\widetilde{G}_n^w - G_P^w\|$ holds, namely

$$\mathcal{E}_{n,P}^w := \left\{ \|\widetilde{G}_n^w - G_P^w\|_\infty \le \Delta_{n,P}^w \right\},$$

where

$$\Delta_{n,P}^w = \begin{cases} 2\, g_{\max,P}^w\, \varepsilon_n^{F,P} + 3\, \varepsilon_{G,P}^w, & F_0 = \mathrm{unif}(0,1) \\ 2\, g_{\max,P}^w / |\Phi^{-1}(\varepsilon_n^{F,P})| + 4L\, \varepsilon_n^{F,P} + 3\, \varepsilon_{G,P}^w + 1/B_n, & F_0 = \mathcal{N}(0,1) \end{cases} \tag{33}$$

with $\varepsilon_n^{F,P}$ and $\varepsilon_{G,P}^w$ given by Lemma C.4. From Lemma C.4, we know $\mathbb{P}_P(\mathcal{E}_{n,P}^w) \to 1$. We now prove $\|\widetilde{Q}_n - Q_P\|_\infty \to_p 0$.

We first show the $F_0 = \mathrm{unif}(0,1)$ case. Let $\mathcal{A}_{n,P}$ be the event that the bound in Lemma C.2 holds, which implies $\|\widehat{F}_n - F_0\|_\infty \le \varepsilon_n^{F,P}$; we have $\mathbb{P}_P(\mathcal{A}_{n,P}) \to 1$. Because Condition 2 holds for $S^1, \ldots, S^W$, we have that on $\mathcal{A}_{n,P}$,

$$|\widetilde{S}_b^w - \widehat{S}_b^w| \le \varepsilon_n^{F,P}, \quad b = 1, \ldots, B_n, \quad w = 1, \ldots, W.$$

Then, we have for $w = 1, \ldots, W$, on the event $\mathcal{A}_{n,P} \cap \mathcal{E}_{n,P}^w$,

$$\begin{aligned}|\widetilde{G}_n^w(\widetilde{S}_b^w) - G_P^w(\widehat{S}_b^w)| &\le |\widetilde{G}_n^w(\widetilde{S}_b^w) - G_P^w(\widetilde{S}_b^w)| + |G_P^w(\widetilde{S}_b^w) - G_P^w(\widehat{S}_b^w)| \\ &\le \Delta_{n,P}^w + g_{P,\max}^w\, \varepsilon_n^{F,P}, \quad b = 1, \ldots, B_n.\end{aligned}$$

Further, on $\mathcal{A}_{n,P} \cap (\cap_w \mathcal{E}_{n,P}^w)$, it holds that

$$|\widetilde{G}_n^w(\widetilde{S}_b^w) - G_P^w(\widehat{S}_b^w)| \le (\max_w \Delta_{n,P}^w) + (\max_w g_{P,\max}^w)\, \varepsilon_n^{F,P}, \quad b = 1, \ldots, B_n, \quad w = 1, \ldots, W,$$

and, because $\max(\cdot)$ is 1-Lipschitz in $\|\cdot\|_\infty$, that

$$|\widetilde{R}_b - \widehat{R}_b| \le (\max_w \Delta_{n,P}^w) + (\max_w g_{P,\max}^w)\, \varepsilon_n^{F,P}, \quad b = 1, \ldots, B_n.$$

Because $\widetilde{Q}_n := \mathbb{F}_{\{\widetilde{R}_b\}}$ and $\widehat{Q}_n := \mathbb{F}_{\{\widehat{R}_b\}}$, using Lemma C.3 one can show that on $\mathcal{A}_{n,P} \cap (\cap_w \mathcal{E}_{n,P}^w)$,

$$\|\widetilde{Q}_n - \widehat{Q}_n\|_\infty \le 2 q_{\max,P} \left[(\max_w \Delta_{n,P}^w) + (\max_w g_{P,\max}^w)\, \varepsilon_n^{F,P} \right] + 2\|\widehat{Q}_n - Q_P\|_\infty.$$

It follows that, on $\mathcal{Q}_{n,P} \cap \mathcal{A}_{n,P} \cap (\cap_w \mathcal{E}^w_{n,P})$,

$$\|\widetilde{Q}_n - Q_P\|_\infty \leq 2W \left[(\max_w \Delta^w_{n,P}) + (\max_w g^w_{P,\max})\, \varepsilon^{F,P}_n \right] + 3\,(1/K_n + \|Q_{n,P} - Q_P\|_\infty)\,. \quad (34)$$

The bound vanishes because $\|Q_{n,P} - Q_P\|_\infty \to 0$ (by Assumption 2), $\varepsilon^{F,P}_n \to 0$, $\max_w \varepsilon^w_{G,P} \to 0$ (by Lemma C.4, Condition 1 and Assumption 2), $\max_w \Delta^w_{n,P} \to 0$ (by Eq. (33)) and $K \to \infty$ (by definition). Since $\mathbb{P}_P(\mathcal{Q}_{n,P} \cap \mathcal{A}_{n,P} \cap (\cap_w \mathcal{E}^w_{n,P})) \to 1$ by the union bound, we conclude that $\|\widetilde{Q}_n - Q_P\|_\infty \to_p 0$.

The $F_0 = \mathcal{N}(0,1)$ case can be shown by an argument similar to the above and the argument used for proving the corresponding case in Lemma C.4. □

Proof of Theorem 2. We continue to use the notation for proving the last lemma.

(i) This follows from the proof of (ii) by taking $\mathcal{P}_0 = \{P\}$.

(ii) Note first that by Lemma C.8 and van der Vaart (2000, Lem. 2.11), we have

$$\sup_{P \in \mathcal{P}_0} \|F_{n,P} - F_0\|_\infty \to 0, \quad \max_w \sup_{P \in \mathcal{P}_0} \|G^w_{n,P} - G^w_P\|_\infty \to 0, \quad \sup_{P \in \mathcal{P}_0} \|Q_{n,P} - Q_P\|_\infty \to 0.$$

Now fix $P \in \mathcal{P}_0$. Observe that on the event $\cap_w \mathcal{E}^w_{n,P}$, we have

$$|R^0_n - R_n| \leq \max_w \Delta^w_{n,P},$$

where each $\Delta^w_{n,P}$ in Eq. (33) tends to zero and $\mathbb{P}_P(\cap_w \mathcal{E}^w_{n,P}) \to 1$. Now let P vary in $\mathcal{P}_0$. Under the given assumptions, observe that there exists a sequence $\delta_n \to 0$, which does not depend on P, such that $\sup_{P \in \mathcal{P}_0} \max_w \Delta^w_{n,P} \leq \delta_n$. Further, because the probability in Lemma C.2 does not depend on P, also observe that

$$\sup_{P \in \mathcal{P}_0} \mathbb{P}_P \left\{ \cup_w (\mathcal{E}^w_{n,P})^c \right\} \to 0$$

holds under the given assumptions; see also the proof of Lemma C.4. Hence, with $\delta_n \to 0$ that does not depend on P, we have

$$\sup_{P \in \mathcal{P}_0} \mathbb{P}_P \left(|R_n - R^0_n| > \delta_n \right) \to 0.$$

Also, observe from the proof of (1), e.g., Eq. (34) for the case of $F_0 = \text{unif}(0,1)$, that there exists sequence $\lambda_n \to 0$, which does not depend on P, such that

$$\sup_{P \in \mathcal{P}_0} \mathbb{P}_P \left(\|\widetilde{Q}_n - Q_P\|_\infty > \lambda_n \right) \to 0.$$

The result then follows from an argument similar to that for proving (ii) of Theorem 1.

□

C.5 Proofs of power results

Proposition C.1. *Let $(T_n^{(1)}, \dots, T_n^{(L)})$ be L exchangeable test statistics. Suppose $\phi_n(T_n^{(1)}, \dots, T_n^{(L)})$ is a most powerful at test level α for testing certain $P_0 \in \mathcal{P}_0$ against certain $P_1 \in \mathcal{P} \setminus \mathcal{P}_0$. Then ϕ_n equals a symmetric function of $(T_n^{(1)}, \dots, T_n^{(n)})$ P_0-almost everywhere.*

Proof. Let p_0 and p_1 be respectively the densities of P_0 and P_1 with respect to some dominating measure μ. By the Neyman–Pearson lemma (Lehmann and Romano, 2005, Theorem 3.2.1), ϕ_n must be equal to $\mathbb{I}\{p_1(T_n^{(1)}, \dots, T_n^{(L)}) > kp_0(T_n^{(1)}, \dots, T_n^{(L)})\}$ for some k, μ-almost everywhere. By exchangeability, p_1 and p_0 are symmetric in $(T_n^{(1)}, \dots, T_n^{(L)})$. The result follows. □

Proof of Theorem 3. We prove the result by showing that $\|\widetilde{G}_n - G_{P_0}\|_\infty \to_p 0$.

We first prove the result for the case when $F_0 = \mathrm{unif}(0,1)$. Recall that in subsampling, for each $b = 1, \dots, B_n$ and each $l = 1, \dots, L$, we have

$$\widehat{H}_{b,l} = T_m^{(l)}(X_{i_{1,b}}, \dots, X_{i_{m,b}}),$$

whose distribution function is F_{m,P_n} because data is drawn iid from P_n. With K_n defined in Eq. (24), let

$$\mathcal{A}_n := \left\{ \|\mathbb{F}_{\widehat{\mathbf{H}}} - F_m^{P_n}\|_\infty \le 1/K_n \right\}.$$

Using Lemma C.1, it holds that $\mathbb{P}_{P_n}(\mathcal{A}_n) \to 1$. Also recall from Algorithm 2 that

$$\widetilde{H}_{b,l} = F_0^{-1}\left(\mathbb{F}_{\widehat{\mathbf{H}}}(\widehat{H}_{b,l}) - 1/(2B_nL)\right) = \mathbb{F}_{\widehat{\mathbf{H}}}(\widehat{H}_{b,l}) - 1/(2B_nL), \quad b = 1, \dots, B_n,\ l = 1, \dots, L.$$

Let $\check{H}_{b,l} := F_{m,P_n}(\widehat{H}_{b,l})$ accordingly. From the definitions of $\widetilde{H}_{b,l}$, $\check{H}_{b,l}$ and $\mathcal{A}_n$, observe that on $\mathcal{A}_n$,

$$\max_{b,l} |\widetilde{H}_{b,l} - \check{H}_{b,l}| \le 1/K_n + 1/(2B_nL) =: \varepsilon_n,$$

where $\varepsilon_n \to 0$. Recall that $\widetilde{S}_b := S(\widetilde{H}_{b,1}, \dots \widetilde{H}_{b,L})$ for $b = 1, \dots, B_n$. Accordingly, let $\check{S}_b := S(\check{H}_{b,1}, \dots \check{H}_{b,L})$. By Condition 2 on S, it follows that on $\mathcal{A}_n$,

$$\max_b |\widetilde{S}_b - \check{S}_b| \le \varepsilon_n. \tag{35}$$

Because P_n converges in copula to P_0, observe that for every b, it holds that

$$\begin{aligned}(\check{H}_{b,1}, \dots, \check{H}_{b,L}) &= \left(F_{m,P_n}(T_m^{(1)}(X_{i_{1,b}}, \dots, X_{i_{m,b}})), \dots, F_{m,P_n}(T_m^{(L)}(X_{i_{1,b}}, \dots, X_{i_{m,b}}))\right) \\ &\to_d (C^{(1)}, \dots, C^{(L)}).\end{aligned}$$

By the continuous mapping theorem, it then follows that

$$\check{S}_b \to_d S(C^{(1)}, \dots, C^{(L)}) \sim G_{P_0}, \quad b = 1, \dots, B_n. \tag{36}$$

Let

$$\mathcal{B}_n := \left\{ \|\mathbb{F}_{\{\check{S}_b\}} - G_{P_0}\|_\infty \le 1/K_n + \|G_{m,P_n} - G_{P_0}\|_\infty =: \delta_n \right\},$$

where G_{m,P_n} is the distribution function of $S(F_{m,P_n}(T_m^{(1)}), \dots, F_{m,P_n}(T_m^{(L)}))$ under $(X, \Omega) \sim P_n^m \times P_\Omega$. By an argument similar to the proof of Lemma C.2, we have that $\mathbb{P}_{P_n}(\mathcal{B}_n) \to 1$. Further, in light of Eq. (36) where $\check{S}_b \sim G_{m,P_n}$ and the continuity of G_{P_0} by Assumption 1, we know $\|G_{m,P_n} - G_{P_0}\|_\infty \to 0$ and hence $\delta_n \to 0$.

Recall that $\widetilde{G}_n := \mathbb{F}_{\{\widetilde{S}_b\}}$. Now, by definition of $\mathcal{B}_n$ and Eq. (35), on $\mathcal{A}_n \cap \mathcal{B}_n$, using Lemma C.3, for every x it holds that

$$\begin{aligned}|\widetilde{G}_n(x) - \mathbb{F}_{\{\check{S}_b\}}(x)| &\le \mathbb{F}_{\{\check{S}_b\}}(x+\varepsilon_n) - \mathbb{F}_{\{\check{S}_b\}}(x-\varepsilon_n) \\ &\le G_{P_0}(x+\varepsilon_n) - G_{P_0}(x-\varepsilon_n) + 2\delta_n \\ &\le 2g_{\max,P_0}\varepsilon_n + 2\delta_n\end{aligned}$$

and consequently,

$$\|\widetilde{G}_n - G_{P_0}\|_\infty \le 2g_{\max,P_0}\varepsilon_n + 3\delta_n \to 0$$

since $\varepsilon_n \to 0$, $\delta_n \to 0$ and $g_{\max,P_0} < \infty$ (by Assumption 1). Because $\mathbb{P}_{P_n}(\mathcal{A}_n \cap \mathcal{B}_n) \to 1$ by a union bound, our desired result that $\|\widetilde{G}_n - G_{P_0}\|_\infty \to_p 0$ follows.

The proof for the case when $F_0 = \mathcal{N}(0,1)$ follows analogously using an argument similar to that employed to prove case (2) of Lemma C.4. □

C.6 Test after rejection sampling

Proof of Proposition 1. Let $\widehat{Q}_n$ be the distribution defined through $\mathrm{d}\widehat{Q}_n/\,\mathrm{d}P = \widehat{r}_n$. Note that $\widehat{Q}_n$ is random because $\widehat{r}_n$ is random. Let γ_n be a sequence of positive integers such that $\gamma_n/n \to 1/C$. Consider a triangular array consisting of rows $(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n)$ drawn iid from $\widehat{Q}_n$ for $n = 1, 2, \ldots$, i.e., the n-th row is drawn from $\widehat{Q}_n^{\gamma_n}$.

We first show that $T_{\gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n) \to_d T$. Let π_n be the total variation coupling between $(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n) \sim \widehat{Q}_n^{\gamma_n}$ and $(\widetilde{X}_1^*, \ldots, \widetilde{X}_{\gamma_n}^*) \sim Q^{\gamma_n}$, where $\widehat{Q}_n^{\gamma_n}$ and Q^{γ_n} are respectively the γ_n-fold product measures of $\widehat{Q}_n$ and Q. Under π_n it holds that

$$\begin{aligned}\mathbb{P}\left(T_{\gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n) \ne T_{\gamma_n}(\widetilde{X}_1^*, \ldots, \widetilde{X}_{\gamma_n}^*) \mid \widehat{r}_n\right) &\le \mathbb{P}\left((\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n) \ne (\widetilde{X}_1^*, \ldots, \widetilde{X}_{\gamma_n}^*) \mid \widehat{r}_n\right) \\ &= \mathrm{d}_{\mathrm{TV}}(\widehat{Q}_n^{\gamma_n},\, Q^{\gamma_n}) \\ &\le \gamma_n\, \mathrm{d}_{\mathrm{TV}}(\widehat{Q}_n, Q) = \gamma_n \int |\widehat{r}_n - r|\, \mathrm{d}P,\end{aligned}$$

where we used total variation's tensorisation property. Taking expectation over $\widehat{r}_n$, we have

$$\begin{aligned}\mathbb{P}\left(T_{\gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n) \ne T_{\gamma_n}(\widetilde{X}_1^*, \ldots, \widetilde{X}_{\gamma_n}^*)\right) &\le \gamma_n\, \mathbb{E} \int |\widehat{r}_n - r|\, \mathrm{d}P \\ &= (\gamma_n/n)\, n\, \mathbb{E} \int |\widehat{r}_n - r|\, \mathrm{d}P \to 0,\end{aligned}$$

where the last step used Eq. (16) and $\gamma_n/n \to 1/C$. Again, using the optimal coupling formulation of total variation, it follows that the total variation between the law of $T_{\gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n)$ and the law of $T_{\gamma_n}(\widetilde{X}_1^*, \ldots, \widetilde{X}_{\gamma_n}^*)$ tends to zero. Because total variation is a strong metric and $T_{\gamma_n}(\widetilde{X}_1^*, \ldots, \widetilde{X}_{\gamma_n}^*) \to_d T$ by our assumption, we have that $T_{\gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n) \to_d T$.

By the properties of rejection sampling (von Neumann, 1951), conditioned on $\Gamma_n = \gamma_n$, the accepted sample $\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n$ are iid from $\widehat{Q}_n$. Further, $\Gamma_n \sim \mathrm{binom}(n, 1/C)$ and hence $\Gamma_n/n \to_p 1/C$. Consider the sequence $\widetilde{T}_n := T_{\Gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\Gamma_n}^n)$. For every subsequence $n_i \nearrow \infty$, there is a further subsequence n_{i_k} such that $\Gamma_{n_{i_k}}/n_{i_k} \to 1/C$ almost surely. This completes the proof because by the result we showed earlier $\widetilde{T}_{n_{i_k}}$ has the same limit law for every such subsequence. □

Proposition C.2 (Parametric density ratio and simple test statistic)**.** *Consider the setting of Proposition 1 without requiring Eq. (16) to hold. Suppose $r(x) = r(x;\beta_0)$ holds for some parametric family $r(x;\beta)$ differentiable in β at β_0, such that $\|\beta-\beta_0\| \to 0$ implies $\int |r(x;\beta) - r(x;\beta_0)|^2 \,\mathrm{d}P(x) \to 0$. Further, let $\widehat{r}_n(x) := r(x;\widehat{\beta}_n)$, where $\widehat{\beta}_n$ is a consistent, asymptotically normal estimator of β_0 fitted on a separate sample of size n from P. Let f be a real-valued function of X such that $\mathbb{E}_P f^4(X) < \infty$. Define $T_n := n^{-1/2}\sum_{i=1}^n \{f(X_i) - \mathbb{E}_Q f(X)\}$ and $T_0 := 0$. Then we have*

$$T_{\Gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\Gamma_n}^n) \to_d \mathcal{N}\left(0,\, \mathrm{var}_Q f(X) + v_{\beta_0}^\top \Sigma v_{\beta_0}/C\right),$$

where Σ is the asymptotic covariance of $\widehat{\beta}_n$ and $v_{\beta_0} = \nabla_\beta \mathbb{E}_P f(X) r(X;\beta)\,|_{\beta=\beta_0}$.

Proof. Let $\widehat{Q}_n$ be the distribution defined through $\mathrm{d}\widehat{Q}_n/\,\mathrm{d}P = \widehat{r}_n$. Note that $\widehat{Q}_n$ is random because $\widehat{r}_n$ is random. Let γ_n be a sequence of positive integers such that $\gamma_n/n \to 1/C$. Consider a triangular array consisting of rows $(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n)$ drawn iid from $\widehat{Q}_n$ for $n = 1, 2, \ldots$, i.e., the n-th row is drawn iid from $\widehat{Q}_n$. By the proof of Proposition 1, it suffices to show that

$$\widetilde{T}_n := T_{\gamma_n}(\widetilde{X}_1^n, \ldots, \widetilde{X}_{\gamma_n}^n) \to_d \mathcal{N}\left(0,\, \mathrm{var}_Q f(X) + v_{\beta_0}^\top \Sigma v_{\beta_0}/C\right).$$

By Lévy's continuity theorem, to show the above, it suffices to show that the characteristic function of $\widetilde{T}_n$ converges to the characteristic function of the limit distribution pointwise. Using

$$T_n = \gamma_n^{-1/2} \sum_{i=1}^{\gamma_n} \left\{ f(\widetilde{X}_i^n) - \mathbb{E}_{\widehat{Q}_n} f(\widetilde{X}_1^n) \right\} + \sqrt{\gamma_n} \left\{ \mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X) \right\},$$

we can write the characteristic function of $\widetilde{T}_n$ as

$$\varphi_n(t) := \mathbb{E}\exp(it\widetilde{T}_n) = \mathbb{E}\left[A_n(t) \exp\left(it\sqrt{\gamma_n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\} \right) \right], \tag{37}$$

where

$$A_n(t) := \mathbb{E}\left[\exp\left(it\gamma_n^{-1/2} \sum_{i=1}^{\gamma_n} \{ f(\widetilde{X}_i^n) - \mathbb{E}_{\widehat{Q}_n} f(X) \} \right) \mid \{\widehat{Q}_n\}_{n=1}^\infty \right].$$

By the Lyapunov CLT, we have

$$\gamma_n^{-1/2} \sum_{i=1}^{\gamma_n} \frac{f(\widetilde{X}_i^n) - \mathbb{E}_{\widehat{Q}_n} f(X)}{\sqrt{\mathrm{var}_{\widehat{Q}_n} f(X)}} \mid \{\widehat{Q}_n\}_{n=1}^\infty \to_d \mathcal{N}(0,1). \tag{38}$$

To see this, note that $\mathbb{E}_P f^4(X) < \infty$ and $\widehat{r}_n \le C$ imply $\mathbb{E}_{\widehat{Q}_n} |f(X)|^3 = \int \widehat{r}_n |f^3| \,\mathrm{d}P < \infty$, which ensures $\mathbb{E}_{\widehat{Q}_n} |f(X) - \mathbb{E}_{\widehat{Q}_n} f(X)|^3 < \infty$. Further, observe that

$$\begin{aligned} |\,\mathbb{E}_{\widehat{Q}_n} f^2(X) - \mathbb{E}_Q f^2(X)| &\le \int |r(x;\widehat{\beta}_n) - r(x;\beta_0)| f^2(x) \,\mathrm{d}P(x) \\ &\le \left(\int |r(x;\widehat{\beta}_n) - r(x;\beta_0)|^2 \,\mathrm{d}P(x) \right)^{1/2} (\mathbb{E}_P f^4(X))^{1/2} \to_p 0 \end{aligned}$$

by $\widehat{\beta}_n \to_p \beta_0$ and our assumption on $r(x;\beta)$ and f. This implies $\mathrm{var}_{\widehat{Q}_n} f(X) \to_p \mathrm{var}_Q f(X)$. Let n' be an arbitrary subsequence of n. Then there exists a further subsequence n'' along

which $\mathrm{var}_{\widehat{Q}_{n''}} f(X) \to_{a.s.} \mathrm{var}_Q f(X)$. By Slutsky's theorem, along the subsequence n'', Eq. (38) implies

$$\gamma_{n''}^{-1/2} \sum_{i=1}^{\gamma_{n''}} \frac{f(\widetilde{X}_i^{n''}) - \mathbb{E}_{\widehat{Q}_{n''}} f(\widetilde{X}_1^{n''})}{\sqrt{\mathrm{var}_Q f(X)}} \mid \{\widehat{Q}_n\}_{n=1}^{\infty} \to_d \mathcal{N}(0,1)$$

and hence

$$\gamma_{n''}^{-1/2} \sum_{i=1}^{\gamma_{n''}} \left\{ f(\widetilde{X}_i^{n''}) - \mathbb{E}_{\widehat{Q}_{n''}} f(\widetilde{X}_1^{n''}) \right\} \mid \{\widehat{Q}_n\}_{n=1}^{\infty} \to_d \mathcal{N}(0, \mathrm{var}_Q f(X)).$$

It follows from Lévy's continuity theorem that at every t, $A_{n''}(t) \to_{a.s.} \exp\left\{-\mathrm{var}_Q f(X)\, t^2/2\right\}$ along n''. Because n' is an arbitrary subsequence of n, we conclude that

$$A_n(t) \to_p \exp\left\{-\mathrm{var}_Q f(X)\, t^2/2\right\}$$

along the original sequence n. Rewrite Eq. (37) as

$$\varphi_n(t) = \exp\left\{-\mathrm{var}_Q f(X)\, t^2/2\right\} \mathbb{E} \exp\left(it\sqrt{\gamma_n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\}\right) + \mathbb{E}\left\{\left[A_n(t) - \exp\left\{-\mathrm{var}_Q f(X)\, t^2/2\right\}\right] \exp\left(it\sqrt{\gamma_n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\}\right)\right\}.$$

For every t, $\left(A_n(t) - \exp\left\{-\mathrm{var}_Q f(X)\, t^2/2\right\}\right) \to_p 0$. Because the moduli of both $A_n(t)$ and $\exp\left(it\sqrt{\gamma_n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\}\right)$ are bounded by one, by DCT, we have

$$\mathbb{E}\left\{\left[A_n(t) - \exp\left\{-\mathrm{var}_Q f(X)\, t^2/2\right\}\right] \exp\left(it\sqrt{\gamma_n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\}\right)\right\} \to 0$$

and hence

$$\varphi_n(t) \to \exp\left\{-\mathrm{var}_Q f(X)\, t^2/2\right\} \lim_n \mathbb{E} \exp\left(it\sqrt{\gamma_n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\}\right).$$

Because $\sqrt{n}(\widehat{\beta}_n - \beta_0) \to_d \mathcal{N}(0, \Sigma)$, by the delta method, we have

$$\sqrt{n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\} \to_d \mathcal{N}(0, v_{\beta_0}^\top \Sigma v_{\beta_0})$$

and hence at every t,

$$\mathbb{E} \exp\left(it\sqrt{\gamma_n}\{\mathbb{E}_{\widehat{Q}_n} f(X) - \mathbb{E}_Q f(X)\}\right) \to \exp\left[-(v_{\beta_0}^\top \Sigma v_{\beta_0}/C)t^2/2\right].$$

Finally, at every t, $\varphi_n(t)$ converges to the characteristic function of $\mathcal{N}(0,\ \mathrm{var}_Q f(X) + v_{\beta_0}^\top \Sigma v_{\beta_0}/C)$ and this completes the proof. □

Remark C.3. In addition to the variance of $f(X)$ under Q, the estimation of the density ratio also contributes to the asymptotic variance of the CLT in Proposition C.2. While it may be difficult to form a self-normalised Z-statistic using the accepted sample, we expect other forms of normalisation, such as using a permutation test (as employed in our numerical studies), to behave properly. For permutations, the additional asymptotic variance equally affects T_{Γ_n} and its permutation counterparts.

C.7 Variance estimator from rank-transformed subsampling

Proof of Proposition 2. Fix $\varepsilon > 0$. Let $\Pi := \left\{(b,l) : \mathbb{F}_{\widehat{\theta}^{(l)}_{m,b}}(\widehat{\theta}^{(l)}_{m,b}) \in (\varepsilon/2, 1-\varepsilon/2)\right\}$. Observe that $|\Pi| > (1-\varepsilon)B_n L \nearrow \infty$. Let $\widetilde{\beta}$ be the least squares estimate of the slope from linearly regressing $\widetilde{H}_{b,l}$ on $\widehat{H}_{b,l} = \sqrt{m/L}(\widehat{\theta}^{(l)}_{m,b} - \theta_0)/\sigma$ and an intercept term using those points in Π. It suffices to show that $\widetilde{\beta} \to_p 1$. By inspecting the proof for case (2) of Lemma C.4, because $\varepsilon_n^{F,P} \to 0$, observe that when n is sufficiently large, we have $2\varepsilon_n^{F,P} < \varepsilon$ and, with probability tending to one, $|\widetilde{H}_{b,l} - \widehat{H}_{b,l}| \le 1/|\Phi^{-1}(\varepsilon_n^{F,P})|$ for every $(b,l) \in \Pi$. Let $\widehat{h}_0 := |\Pi|^{-1}\sum_{(b,l)\in\Pi}\widehat{H}_{b,l}$ and note that

$$\sum_{(b,l)\in\Pi}\widehat{H}_{b,l}(\widehat{H}_{b,l} - \widehat{h}_0) = \sum_{(b,l)\in\Pi}(\widehat{H}_{b,l} - \widehat{h}_0)^2.$$

Thus

$$\widetilde{\beta} - 1 = \frac{\sum_{(b,l)\in\Pi}(\widetilde{H}_{b,l} - \widehat{H}_{b,l})(\widehat{H}_{b,l} - \widehat{h}_0)}{\sum_{(b,l)\in\Pi}(\widehat{H}_{b,l} - \widehat{h}_0)^2} \le \frac{1}{|\Phi^{-1}(\varepsilon_n^{F,P})|}\frac{1}{\sqrt{|\Pi|^{-1}\sum_{(b,l)\in\Pi}(\widehat{H}_{b,l} - \widehat{h}_0)^2}}$$

with probability tending to one. Noting that $|\Phi^{-1}(\varepsilon_n^{F,P})| \to \infty$ and that $|\Pi|^{-1}\sum_{(b,l)\in\Pi}(\widehat{H}_{b,l} - \widehat{h}_0)^2$ tends to a positive constant by Eq. (18), it follows that $\widetilde{\beta} \to_p 1$. □

C.8 Auxiliary lemmas

The following standard result follows directly from the definition of F^{-1}.

Lemma C.6. *Given any distribution function F, we have $F^{-1}(F(x)) \le x$ and $F(F^{-1}(\alpha)) \ge \alpha$.*

C.9 Other results

Proposition C.3. *Suppose $T_n^{(1)}, \ldots, T_n^{(L)}$ are exchangeable and S satisfies Condition 2. Then, under Condition 1, the sequence (S_n) is uniformly tight.*

Proof. We know from Prohorov's theorem (van der Vaart, 2000, Thm. 2.4) and Condition 1 that the sequence $(T_n^{(1)})$ is uniformly tight, so given $\varepsilon > 0$, there exists $M \ge 0$ such that $\sup_n \mathbb{P}_P(|T_n^{(1)}| > M) < \varepsilon/L$.

Now from Condition 2 we have that $|S_n - S(0,\ldots,0)| \le \|T_n\|_\infty$. Thus,

$$\begin{aligned}\sup_n \mathbb{P}_P\left(|S_n| > M + |S(0,\ldots,0)|\right) &\le \sup_n \mathbb{P}_P\left(\|T_n\|_\infty > M\right) \\ &\le L \sup_n \mathbb{P}_P\left(|T_n^{(1)}| > M\right) < \varepsilon,\end{aligned}$$

applying a union bound and appealing to exchangeability in the final line. □

Lemma C.7. *Suppose that for all $P \in \mathcal{P}$, distribution functions $H_{n,P}$ converge uniformly to a continuous distribution function H_P, i.e., for all $x \in \mathbb{R}$,*

$$\sup_{P\in\mathcal{P}} |H_{n,P}(x) - H_P(x)| \to 0.$$

Assume that each H_P has density h_P and $\sup_{P\in\mathcal{P}}\|h_P\|_\infty =: h_{\max} < \infty$. Further, suppose $\{H_P : P \in \mathcal{P}\}$ is tight. Then we have

$$\sup_{P\in\mathcal{P}} \|H_{n,P} - H_P\|_\infty \to 0.$$

Proof. Given $\varepsilon > 0$, by tightness of $\{H_P : P \in \mathcal{P}\}$, there exists $M \geq 0$ such that $\sup_{P\in\mathcal{P}} H_P(-M) < \varepsilon$ and $\sup_{P\in\mathcal{P}}\{1 - H_P(M)\} < \varepsilon$. Next set $\delta := \varepsilon/h_{\max}$ and $x_j = \delta(j-1) - M$ for $j = 1, \ldots, \lceil 2M/\delta \rceil =: J - 1$. Additionally set $x_0 := -\infty$ and $x_J := \infty$. Fix $P \in \mathcal{P}$ and note that $H_P(x_j) - H_P(x_{j-1}) \leq \varepsilon$ for all j.

Now for each $x \in \mathbb{R}$, there exists j with $x_{j-1} \leq x \leq x_j$, and

$$H_{n,P}(x) - H_P(x) \leq H_{n,P}(x_j) - H_P(x_{j-1}) \leq H_{n,P}(x_j) - H_P(x_j) + \varepsilon$$
$$H_{n,P}(x) - H_P(x) \geq H_{n,P}(x_{j-1}) - H_P(x_j) \geq H_{n,P}(x_{j-1}) - H_P(x_{j-1}) - \varepsilon,$$

so

$$\|H_{n,P} - H_P\|_\infty \leq \max_{j=0,\ldots,J} |H_{n,P}(x_j) - H_P(x_j)| + \varepsilon.$$

Thus

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \|H_{n,P} - H_P\|_\infty \leq \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \max_{j=0,\ldots,J} |H_{n,P}(x_j) - H_P(x_j)| + \varepsilon = \varepsilon.$$

As ε was arbitrary, we have the desired result. □

Lemma C.8. *Let $G_{n,p}$ and G_P for $P \in \mathcal{P}_0$ be as in (3) of Theorem 1. Then*

$$\sup_{P\in\mathcal{P}_0} \|G_{n,P} - G_P\|_\infty \to 0.$$

Proof. By Lemma C.7 it suffices to show that for all $\varepsilon > 0$, there exists $M \geq 0$ such that $\sup_{P\in\mathcal{P}_0} G_P(-M) < \varepsilon$ and $\sup_{P\in\mathcal{P}_0}\{1 - G_P(M)\} < \varepsilon$.

We have that given $\varepsilon > 0$, there exists $M' \geq 0$ such that $F_0(-M') + F_0(M') < \varepsilon/L$. Then, arguing similarly to the proof of Proposition C.3, for all $P \in \mathcal{P}_0$,

$$\mathbb{P}_P(|S_n| > M + |S(0,\ldots,0)|) \leq L\,\mathbb{P}_P\left(|T_n^{(1)}| > M\right).$$

Taking suprema over $P \in \mathcal{P}_0$ and limits as $n \to \infty$,

$$\sup_{P\in\mathcal{P}_0} \{G_P(-M') + G_P(M')\} < \varepsilon,$$

as required. □

Proposition C.4. *Let $W_1, \ldots, W_n$ be iid from P. Suppose for $f(W;\beta_0)$ it holds that $\mathbb{E}\, f(W;\beta_0) = 0$, $\sigma^2 := \mathbb{E}\, f^2(W;\beta_0) < \infty$ and that $f(W;\beta)$ permits a second order expansion in β in a neighbourhood of β_0. Let $\widehat{\beta}$ be a consistent, asymptotically linear estimator of β_0 from $W_1, \ldots, W_n$. Then it holds that*

$$\frac{\sum_i f(W_i;\widehat{\beta})}{\sqrt{\sum_i f^2(W_i;\widehat{\beta})}} \to_d \mathcal{N}(0,1).$$

Proof. In view of

$$\frac{\sum_i f(W_i;\widehat{\beta})}{\sqrt{\sum_i f^2(W_i;\widehat{\beta})}} = \frac{n^{-1/2}\sum_i f(W_i;\widehat{\beta})}{\sqrt{n^{-1}\sum_i f^2(W_i;\widehat{\beta})}},$$

it suffices to show that $n^{-1}\sum_i f(W_i;\widehat{\beta})$ is a consistent, asymptotically linear estimator of zero, i.e.,

$$n^{-1/2}\sum_i f(W_i;\widehat{\beta}) = n^{-1/2}\sum_i \varphi(W_i) + o_p(1),$$

for some φ such that $\mathbb{E}\,\varphi(W) = 0$, $\mathbb{E}\,\varphi^2(W) < \infty$. Let $h(W) := (\partial f(W;\beta)/\partial\beta)|_{\beta=\beta_0}$. Using

$$f(W;\widehat{\beta}) = f(W;\beta_0) + h(W)^\top(\widehat{\beta} - \beta_0) + O(\|\widehat{\beta} - \beta_0\|^2),$$

we have

$$\begin{aligned}
n^{1/2}\sum_i f(W_i;\widehat{\beta}) &= n^{-1/2}\sum_i \left[f(W_i;\beta_0) + h(W_i)^\top(\widehat{\beta} - \beta_0) + O(\|\widehat{\beta} - \beta_0\|^2)\right] \\
&= n^{-1/2}\sum_i f(W_i;\beta_0) + \sqrt{n}(\widehat{\beta} - \beta_0)^\top\left[n^{-1}\sum_i h(W_i)\right] + O(n^{1/2}\|\widehat{\beta} - \beta_0\|^2) \\
&= n^{-1/2}\sum_i f(W_i;\beta_0) + \sqrt{n}\left[\mathbb{E}\,h^\top(W)\right](\widehat{\beta} - \beta_0) + O_p(n^{-1/2}) \\
&= n^{-1/2}\sum_i \left\{f(W_i;\beta_0) + \left[\mathbb{E}\,h^\top(W)\right]\gamma(W_i)\right\} + o_p(1),
\end{aligned}$$

where $\gamma(W)$ is the influence function for $\widehat{\beta}$. Hence, we have $\varphi(W) = f(W;\beta_0) + (\mathbb{E}\,h)^\top\gamma(W)$ and it is easy to check that $\mathbb{E}\,\varphi(W) = 0$ and $\mathbb{E}\,\varphi^2(W) < \infty$. □

D First-order power analysis

This section of the appendix is dedicated to proving Theorem 4, our main theorem for the first-order power behaviour of rank-transformed subsampling; we also prove Theorem D.2 for the first-order power behaviour of ordinary subsampling. We sketch the proof of our main theorem here. If we compare the copula of $(T_n^{(1)}, \ldots, T_n^{(L)})$ before and after the rank transform, the rank transform can be viewed as a random perturbation applied to each coordinate of the copula. It follows from the standard empirical process that the perturbation, when scaled by $\sqrt{n/m}$, converges to an average over exchangeable copies of a Brownian bridge (Lemma D.5). To propagate this random limit to the change in distribution of $S(T_n^{(1)}, \ldots, T_n^{(L)})$ after the rank transform, we rely on the functional delta method. As a crucial precursory step, in Appendix D.1 we establish the Hadamard differentiability of the distribution function of a monotone function of a random vector with respect to the coordinatewise perturbation of the random vector (Proposition D.1). This is achieved by firstly establishing the result for a single random variable (Lemma D.1) and then extending it to a random vector through a conditioning argument (Proposition D.1). In Appendix D.2, we fully state and prove Theorem 4 , which builds on Corollary D.1, itself a specialised version of Proposition D.1 under symmetry. Then, in Appendix D.3, we prove Theorem D.2 about the first-order behaviour of ordinary subsampling. In Appendix D.4, we check that the regularity assumptions required by Theorem 4 are met by cases considered in this paper when the underlying copula is Gaussian. We illustrate our theory along with numerical results using a simple example in Appendix D.5. Finally, various auxiliary results used through the arguments are stated and proved in Appendix D.6.

Notation For any real number a, we use the notation $a_+ = a\mathbb{I}\{a \geq 0\}$ and $a_- = -a\mathbb{I}\{a < 0\}$, and similarly for real-valued functions. Let $D[a,b]$ be the space of càdlàg functions on $[a,b]$ equipped with $\|\cdot\|_\infty$. For a function f, its modulus of continuity is $\Delta f(\delta) := \sup_{x,x':\|x-x'\|\leq\delta} |f(x) - f(x')|$ for $\delta \geq 0$. We use $a \vee b$ and $a \wedge b$ to denote $\max(a,b)$ and $\min(a,b)$, respectively. We use Id to denote the distribution function of $\mathrm{unif}(0,1)$. We use $\mathrm{d_{TV}}(X,Y)$ to denote the total variation distance between X's distribution and Y's distribution.

Equicontinuity Let $f_\gamma(x)$ be a class of real-valued functions defined on a Euclidean domain, indexed by γ. We say $\{f_\gamma(x) : \gamma \in \Gamma\}$ is equicontinuous at x_0 if for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|f_\gamma(x) - f_\gamma(x_0)| < \varepsilon$ holds for every $\gamma \in \Gamma$ whenever $\|x - x_0\| < \delta$.

D.1 Hadamard differentiability of additive coordinatewise perturbation

D.1.1 Univariate result

Lemma D.1. *Let X be a real-valued random variable that admits a density $f(x)$ with respect to the Lebesgue measure. Let A be a subset of $\mathbb{R}$. Suppose $f_{\max} := \sup_{x\in A} f(x) < \infty$. Suppose $f(x)$ is uniformly continuous on A. Let $\mathcal{H}$ be the set of $\mathbb{R} \to \mathbb{R}$ functions such that the function takes value zero outside A. Then, for any $t \to 0$, $h_t, h \in \mathcal{H}$ such that $\|h_t - h\|_\infty \to 0$ and h is uniformly continuous, it holds that*

$$\sup_{x\in A} \left| t^{-1} \left\{ \mathbb{P}\left(X + t h_t(X) \le x\right) - \mathbb{P}\left(X \le x\right) \right\} + f(x)\, h(x) \right| \to 0.$$

Proof. Without loss of generality, we assume $t \searrow 0$ and $\|h_t\| > 0$. Let $m_t := t\,\|h_t\|_\infty$, which is a positive sequence converging to zero by our assumption. In light of

$$\begin{aligned} \{X + t h_t(X) \le x\} = \{X \le x - m_t\} \cup \left(\{x - m_t < X \le x\} \cap \{X \le x - t h_t(X)\}\right) \\ \cup \left(\{x < X \le x + m_t\} \cap \{X \le x - t h_t(X)\}\right), \qquad (39) \\ \{X \le x\} = \{X \le x - m_t\} \cup \{x - m_t < X \le x\}, \end{aligned}$$

we can write

$$\begin{aligned} \frac{\mathbb{P}(X + t\, h_t(X) \le x) - \mathbb{P}(X \le x)}{t} = \frac{1}{t}\, \mathbb{P}\left(\{x < X \le x + m_t\} \cap \{X \le x - t h_t(X)\}\right) \\ - \frac{1}{t}\, \mathbb{P}\left(\{x - m_t < X \le x\} \cap \{X > x - t h_t(X)\}\right). \qquad (40) \end{aligned}$$

We claim that the first and second terms on the RHS converge to $f(x)\, h_-(x)$ and $-f(x)h_+(x)$ respectively, uniformly over $x \in A$. Using the fact that $h = h_+ - h_-$, combining the two yields $-f(x)h(x)$ as desired.

We now prove the claim for the first term; the argument for the second term is similar. For any $x \in A$, we have

$$\frac{1}{t}\, \mathbb{P}\left(\{x < X \le x + m_t\} \cap \{X \le x - t h_t(X)\}\right) = \frac{1}{t} \int_x^{x+m_t} \mathbb{I}\{h_t(y) \le (x - y)/t\} f(y)\, \mathrm{d}y.$$

By a change of variable $y = x + m_t u$ for $u \in (0, 1]$, we get

$$\begin{aligned}
&\frac{1}{t}\,\mathbb{P}\left(\{x < X \le x + m_t\} \cap \{X \le x - t h_t(X)\}\right) \\
&= \frac{1}{t}\int_0^1 \mathbb{I}\{h_t(x + m_t u) \le -(m_t/t)u\} f(x + m_t u)\, m_t \,\mathrm{d}u \\
&= \left(\frac{m_t}{t}\right)\int_0^1 \mathbb{I}\{\|h_t\|_\infty u \le -h_t(x + m_t u)\} f(x + m_t u)\,\mathrm{d}u \\
&= \|h_t\|_\infty \int_0^1 \mathbb{I}\{\|h_t\|_\infty u \le -h_t(x + m_t u)\} f(x + m_t u)\,\mathrm{d}u \\
&= \underbrace{\|h_t\|_\infty \int_0^1 \mathbb{I}\{\|h\|_\infty u \le -h(x)\} f(x)\,\mathrm{d}u}_{\text{(I)}} + \underbrace{\|h_t\|_\infty \int_0^1 \mathbb{I}\{\|h_t\|_\infty u \le -h_t(x + m_t u)\}\,[f(x + m_t u) - f(x)]\,\mathrm{d}u}_{\text{(II)}} \\
&\quad + \underbrace{\|h_t\|_\infty \int_0^1 \left[\mathbb{I}\{\|h_t\|_\infty u \le -h_t(x + m_t u)\} - \mathbb{I}\{\|h\|_\infty u \le -h(x)\}\right] f(x)\,\mathrm{d}u}_{\text{(III)}}.
\end{aligned}$$

The result follows from analyzing (I), (II) and (III) separately, detailed as below.

Term (I): When $\|h\|_\infty = 0$, observe that

$$|\text{(I)}| \le \|h_t\|_\infty f(x) \int_0^1 \mathrm{d}u = \|h_t\|_\infty f(x) \le \|h_t\|_\infty f_{\max} \to 0.$$

When $\|h\|_\infty > 0$, we have

$$\begin{aligned}
\text{(I)} = \|h_t\|_\infty f(x) \left(\int_0^{h_-(x)/\|h\|_\infty} \mathrm{d}u\right) &= f(x)\, h_-(x) \|h_t\|_\infty / \|h\|_\infty \\
&= f(x)\, h_-(x) + f(x)\, h_-(x)\left(\|h_t\|_\infty / \|h\|_\infty - 1\right),
\end{aligned}$$

where the second term uniformly converges to zero because h is bounded and $f(x) \le f_{\max}$. Thus, in either case, (I) converges to $f(x)\, h_-(x)$ uniformly over $x \in A$.

Term (II): We show that (II) converges to zero uniformly. Because $\|h_t\|_\infty \to \|h\|_\infty$ irrespective of x, it suffices to show

$$\underbrace{\int_0^1 \mathbb{I}\{\|h_t\|_\infty u \le -h_t(x + m_t u)\}\,|f(x + m_t u) - f(x)|\,\mathrm{d}u}_{\text{(II)}'} \to 0$$

uniformly over $x \in A$. If $x + m_t u \notin A$, $\text{(II)}' = 0$ by $h_t \in \mathcal{H}$; if $x + m_t u \in A$, we have

$$\text{(II)}' \le \int_0^1 |f(x + m_t u) - f(x)|\,\mathrm{d}u \le \sup_{x, x' \in A : |x - x'| \le m_t} |f(x) - f(x')| \to 0$$

by uniform continuity of f on A.

Term (III): We now show that (III) converges to zero uniformly. The case when $\|h\|_\infty = 0$ is straightforward: we have

$$|\text{(III)}| \le 2\|h_t\|_\infty f(x) \int_0^1 \mathrm{d}u \le 2 f_{\max} \|h_t\|_\infty \to 0.$$

Now we suppose $\|h\|_\infty > 0$. We have that

$$\begin{aligned}(\mathrm{III}) &= \|h_t\|_\infty f(x) \int_0^1 \left[\mathbb{I}\left\{u \le -\frac{h_t(x+m_t u)}{\|h_t\|_\infty}\right\} - \mathbb{I}\left\{u \le -\frac{h(x)}{\|h\|_\infty}\right\}\right] \mathrm{d}u \\ &= \|h_t\|_\infty f(x) \underbrace{\int_0^1 \left[\mathbb{I}\left\{u \le \left(\frac{h_t(x+m_t u)}{\|h_t\|_\infty}\right)_-\right\} - \mathbb{I}\left\{u \le \left(\frac{h(x)}{\|h\|_\infty}\right)_-\right\}\right] \mathrm{d}u}_{(\mathrm{III})'}.\end{aligned}$$

Because $|f(x)| \le f_{\max}$ for $x \in A$, it suffices to show that $(\mathrm{III})'$ uniformly converges to zero. Observe that

$$\begin{aligned}\left|\left(\frac{h_t(x)}{\|h_t\|_\infty}\right)_- - \left(\frac{h(x)}{\|h\|_\infty}\right)_-\right| &\le \left|\left(\frac{h_t(x)}{\|h_t\|_\infty}\right) - \left(\frac{h(x)}{\|h\|_\infty}\right)\right| \\ &\le \|h\|_\infty(1/\|h_t\|_\infty - 1/\|h\|_\infty) + \|h_t - h\|_\infty/\|h_t\|_\infty \to 0,\end{aligned}$$

uniformly over x. Therefore, because $|m_t u| \le m_t = t\|h_t\|_\infty \to 0$, given any $\varepsilon > 0$, there exists $T'_\varepsilon > 0$ such that for every $t < T'_\varepsilon$,

$$\left\|\left(\frac{h_t}{\|h_t\|_\infty}\right)_- - \left(\frac{h}{\|h\|_\infty}\right)_-\right\|_\infty < \varepsilon.$$

It follows that for $t < T'_\varepsilon$,

$$(\mathrm{III})' \le \int_0^1 \left[\mathbb{I}\left\{u \le \varepsilon + \left(\frac{h(x+m_t u)}{\|h\|_\infty}\right)_-\right\} - \mathbb{I}\left\{u \le \left(\frac{h(x)}{\|h\|_\infty}\right)_-\right\}\right] \mathrm{d}u.$$

Further, because h is uniformly continuous, there exists $T''_\varepsilon > 0$ such that $|h(x+m_t) - h(x)| < \varepsilon\|h\|_\infty$ for all $t < T''_\varepsilon$. Hence, for all $t < T'_\varepsilon \wedge T''_\varepsilon$, we have

$$(\mathrm{III})' \le \int_0^1 \left[\mathbb{I}\left\{u \le 2\varepsilon + \left(\frac{h(x)}{\|h\|_\infty}\right)_-\right\} - \mathbb{I}\left\{u \le \left(\frac{h(x)}{\|h\|_\infty}\right)_-\right\}\right] \mathrm{d}u \le 2\varepsilon.$$

An analogous lower bound can be derived in a similar way. Because $T'_\varepsilon, T''_\varepsilon$ do not depend on x and $\varepsilon > 0$ is arbitrary, we conclude that $(\mathrm{III})'$ converges to zero uniformly as desired.

□

D.1.2 Multivariate result

Proposition D.1 (Hadamard differentiability of additive coordinatewise perturbation)**.** *Let $R(x_1, \ldots, x_L) : [0,1]^L \to \mathbb{R} \cup \{\pm\infty\}$ be a function that is non-decreasing in each coordinate. Let $[a,b]$ be an interval in the range of R. For $l = 1, \ldots, L$, define*

$$R_l^{-1}(x_{-l}; r) := \sup\{x_l \in [0,1] : R(x) \le r\}, \quad (x_{-l}, r) \in [0,1]^{L-1} \times [a,b].$$

Let $X = (X_1, \ldots, X_L)$ be a random vector in $[0,1]^L$ with a distribution absolutely continuous with respect to the Lebesgue measure. Let $f(x_l \mid x_{-l}) := f(x_l \mid X_{-l} = x_{-l})$ for $l = 1, \ldots, L$ be the corresponding conditional densities with respect to the Lebesgue measure.

Let D_φ be the subset of $D[0,1]$ consisting of g such that $0 \le u + g(u) \le 1$ for $u \in [0,1]$ and $g(0) = g(1) = 0$. For $g \in D_\varphi$, consider the following map

$$\varphi : D_\varphi \to D[a,b], \quad \varphi(g) = r \mapsto \mathbb{P}\left(R(X_1 + g(X_1), \ldots, X_L + g(X_L)) \le r\right).$$

Then, under Assumptions D.1 and D.2 below, the map φ is Hadamard differentiable at $g = 0$ tangentially to any uniformly continuous $h \in D[0,1]$ satisfying $h(0) = h(1) = 0$ with derivative

$$\varphi'_{g=0}(h) = r \mapsto -\sum_{l=1}^{L} \mathbb{E}\left[f(R_l^{-1}(X_{-l}; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}; r))\right], \quad r \in [a,b].$$

Further, the derivative is continuous on $[a,b]$.

Assumption D.1. *For $l = 1, \ldots, L$ and Lebesgue almost every $x_{-l} \in [0,1]^{L-1}$, the map $x_l \mapsto f(x_l \mid x_{-l})$ is continuous on $[0,1]$.*

Assumption D.2. *There exists $\delta > 0$ such that for $l = 1, \ldots, L$, $\{x_l \mapsto f(x_l \mid x_{-l}) : x_{-l}\}$ is equicontinuous on $A_\delta^l := \{(x_l, x_{-l}) : x_l \in A^l_{\delta, x_{-l}},\ x_{-l} \in [0,1]^{L-1}\}$, where*

$$A^l_{\delta, x_{-l}} := \left[\left(R_l^{-1}(x_{-l} + \delta; a) - \delta\right) \vee 0,\ \left(R_l^{-1}(x_{-l} - \delta; b) + \delta\right) \wedge 1\right],$$

where $x_{-l} \pm \delta$ is applied coordinatewise.

The next corollary specialises Proposition D.1 to the case where R is a symmetric function and the random vector is exchangeable.

Corollary D.1 (Under symmetry)**.** *Let $R(c_1, \ldots, c_L) : [0,1]^L \to \mathbb{R} \cup \{\pm\infty\}$ be a symmetric function that is non-decreasing in each coordinate. Let $[a,b]$ be an interval in the range of R. Define*

$$R^{-1}(c_{-1}; r) := \sup\{c_1 \in [0,1] : R(c) \le r\}, \quad (c_{-1}, r) \in [0,1]^{L-1} \times [a,b].$$

Let $C = (C_1, \ldots, C_L)$ be an exchangeable random vector in $[0,1]^L$ with a distribution absolutely continuous with respect to the Lebesgue measure. Let $f(c_1 \mid c_{-1}) := f(c_1 \mid C_{-1} = c_{-1})$ be the corresponding conditional density with respect to the Lebesgue measure.

Let D_φ be the subset of $D[0,1]$ consisting of g such that $0 \le u + g(u) \le 1$ for $u \in [0,1]$ and $g(0) = g(1) = 0$. For $g \in D_\varphi$, consider the following map

$$\varphi : D_\varphi \to D[a,b], \quad \varphi(g) = r \mapsto \mathbb{P}\left(R(C_1 + g(C_1), \ldots, C_L + g(C_L)) \le r\right).$$

Then, under Assumptions D.3 and D.4 below, the map φ is Hadamard differentiable at $g = 0$ tangentially to any uniformly continuous $h \in D[0,1]$ satisfying $h(0) = h(1) = 0$ with derivative

$$\varphi'_{g=0}(h) = r \mapsto -L\, \mathbb{E}\left[f(R^{-1}(C_{-1}; r) \mid C_{-1})\, h(R^{-1}(C_{-1}; r))\right], \quad r \in [a,b].$$

Further, the derivative is continuous on $[a,b]$.

Assumption D.3. *For Lebesgue almost every $c_{-1} \in [0,1]^{L-1}$, the map $c_1 \mapsto f(c_1 \mid c_{-1})$ is continuous on $[0,1]$.*

Assumption D.4. *There exists $\delta > 0$ such that $\{c_1 \mapsto f(c_1 \mid c_{-1}) : c_{-1}\}$ is equicontinuous on $A_\delta := \{(c_1, c_{-1}) : c_1 \in A_{\delta, c_{-1}},\ c_{-1} \in [0,1]^{L-1}\}$, where*

$$A_{\delta, c_{-1}} := \left[\left(R^{-1}(c_{-1} + \delta; a) - \delta\right) \vee 0,\ \left(R^{-1}(c_{-1} - \delta; b) + \delta\right) \wedge 1\right],$$

where $c_{-1} \pm \delta$ is applied coordinatewise.

We now prove Proposition D.1, the key proposition that underpins the first-order power result. The proof relies several auxiliary results which can be found in Appendix D.6.

Proof of Proposition D.1. First note that the claimed $\varphi'_{g=0}(h)$ is linear in h. Further, we claim that the linear map is continuous because $\|\varphi'_{g=0}(h)\|_\infty \le L f_{\max} \|h\|_\infty$, where we define

$$f_{\max} := \max_{1\le l\le L} \sup_{x\in A^l_\delta} f(x_l \mid x_{-l}), \tag{41}$$

based on the region defined in Assumption D.2. Using Assumption D.2 and the fact that every $A^l_{\delta,x_{-l}} \subset [0,1]$ is compact, we have $f_{\max} < \infty$. To see the $f(R_l^{-1}(X_{-l};r) \mid X_{-l}) \le f_{\max}$, note that the monotonicity of R implies the following property of $A^l_{\delta,x_{-l}}$ $(l = 1,\dots,L)$:

$$x_l \in [0,1],\ |x_l - R_l^{-1}(x'_{-l};r)| \le \delta \text{ for some } x'_{-l}, r \text{ such that } \|x'_{-l} - x_{-l}\|_\infty \le \delta,\ r \in [a,b] \\ \implies x_l \in A^l_{\delta,x_{-l}}. \tag{42}$$

Clearly, $R_l^{-1}(x_{-l};r) \in A^l_{\delta,x_{-l}}$ for any $r \in [a,b]$.

Fix any uniformly continuous $h \in D[0,1]$ with $h(0) = h(1) = 0$. Consider $t \searrow 0$ and let h_t be a sequence in $D[0,1]$ such that $\|h_t - h\|_\infty \to 0$ and $th_t \in D_\varphi$, which implies $h_t(0) = h_t(1) = 0$. Without loss of generality, we can assume $\|h_t\|_\infty > 0$. Because $t\|h_t\|_\infty \to 0$, fix any $T_\delta > 0$ with the following property, which will be useful later:

$$t < T_\delta \implies t\|h_t\|_\infty < \delta. \tag{43}$$

We want to show that

$$\sup_{a\le r\le b} \Big| t^{-1} \{\mathbb{P}\left(R(X_1 + th_t(X_1),\dots,X_L + th_t(X_L)) \le r\right) - \mathbb{P}\left(R(X_1,\dots,X_L) \le r\right)\} \\ + \sum_{l=1}^{L} \mathbb{E}\left[f(R_l^{-1}(X_{-l};r) \mid X_{-l})\, h(R_l^{-1}(X_{-l};r))\}\right] \Big| \to 0.$$

In light of the following telescoping sum of L quotients

$$\begin{aligned}
&t^{-1} \{\mathbb{P}\left(R(X_1 + th_t(X_1),\dots,X_L + th_t(X_L)) \le r\right) - \mathbb{P}\left(R(X_1,\dots,X_L) \le r\right)\} \\
&= t^{-1}\big\{ \mathbb{P}\left(R(X_1 + th_t(X_1),\dots,X_L + th_t(X_L)) \le r\right) \\
&\qquad - \mathbb{P}\left(R(X_1, X_2 + th_t(X_2),\dots,X_L + th_t(X_L)) \le r\right) \big\} \\
&\quad + t^{-1}\big\{ \mathbb{P}\left(R(X_1, X_2 + th_t(X_2),\dots,X_L + th_t(X_L)) \le r\right) \\
&\qquad - \mathbb{P}\left(R(X_1, X_2, X_3 + th_t(X_3),\dots,X_L + th_t(X_L)) \le r\right) \big\} \\
&\qquad + \cdots + t^{-1}\big\{ \mathbb{P}\left(R(X_1,\dots,X_{L-1}, X_L + th_t(X_L)) \le r\right) - \mathbb{P}\left(R(X_1,\dots,X_{L-1}, X_L) \le r\right) \big\},
\end{aligned}$$

we prove our result by showing that, for $l = 1,\dots,L$,

$$\sup_{a\le r\le b} \Big| t^{-1}\big\{ \mathbb{P}\left(R(X_1,\dots,X_{l-1}, X_l + th_t(X_l),\dots,X_L + th_t(X_L)) \le r\right) \\ - \mathbb{P}\left(R(X_1,\dots,X_{l-1}, X_l, X_{l+1} + th_t(X_{l+1}),\dots,X_L + th_t(X_L)) \le r\right) \big\} \\ + \mathbb{E}\left[f(R_l^{-1}(X_{-l};r) \mid X_{-l})\, h(R_l^{-1}(X_{-l};r))\right] \Big| \to 0,$$

or, by conditioning on $X_{-l} := (X_1,\dots,X_{l-1},X_{l+1},\dots,X_L)$ and using the monotonicity of R, equivalently,

$$\sup_{a\le r\le b} \Big| \mathbb{E}\left[t^{-1}\left\{\mathbb{P}\left(X_l + th_t(X_l) \le R_l^{-1}(X^t_{-l};r) \mid X_{-l}\right) - \mathbb{P}\left(X_l \le R_l^{-1}(X^t_{-l};r) \mid X_{-l}\right)\right\}\right. \\ \left. + f(R_l^{-1}(X_{-l};r) \mid X_{-l})\, h(R_l^{-1}(X_{-l};r))\}\right] \Big| \to 0. \tag{44}$$

In above, we define

$$X_{-l}^t := (X_1, \ldots, X_{l-1}, X_{l+1} + th_t(X_{l+1}), \ldots, X_L + th_t(X_L)), \tag{45}$$

which is a measurable function of X_{-l} and satisfies

$$\|X_{-l}^t - X_-\|_\infty \le t\|h_t\|_\infty. \tag{46}$$

We prove our desired Eq. (44) by showing both (I) and (II) below converge to zero uniformly over $r \in [a, b]$:

$$\begin{aligned}\text{(I)} := \mathbb{E}\big[t^{-1}\big\{&\mathbb{P}\left(X_l + th_t(X_l) \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right) - \mathbb{P}\left(X_l \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right)\big\} \\ &+ f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}^t; r))\big],\end{aligned}$$

$$\text{(II)} := \mathbb{E}\left[f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}^t; r)) - f(R_l^{-1}(X_{-l}; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}; r))\right].$$

Bounding (I) We write $\text{(I)} = \mathbb{E}\, J_{t,r}^l(X_{-l})$ with

$$\begin{aligned}J_{t,r}^l(X_{-l}) := t^{-1}\big\{&\mathbb{P}\left(X_l + th_t(X_l) \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right) - \mathbb{P}\left(X_l \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right)\big\} \\ &+ f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}^t; r)).\end{aligned} \tag{47}$$

It holds that

$$\begin{aligned}\sup_{r\in[a,b]} |J_{t,r}^l(X_{-l})| \le \sup_{x_l\in[0,1]} \big|t^{-1}\{&\mathbb{P}\left(X_l + th_t(X_l) \le x_l \mid X_{-l}\right) - \mathbb{P}\left(X_l \le x_l \mid X_{-l}\right)\} \\ &+ f(x_l \mid X_{-l})\, h(x_l)\big| =: \widetilde{J}_t^l(X_{-l}).\end{aligned} \tag{48}$$

We claim that

$$\sup_{r\in[a,b]} |J_{t,r}^l(X_{-l})| \le \widetilde{J}_t^l(X_{-l}) \to 0 \quad \text{a.e.},$$

which follows from observing that $R_l^{-1}(X_{-l}^t; r)$ is a measurable function of X_{-l} and applying Lemma D.1 with $A = [0, 1]$ conditional on X_{-l}. To see that Lemma D.1 can be applied, note that for almost every x_{-l}, the map $x_l \mapsto f(x_l \mid x_{-l})$ is continuous on A (Assumption D.1), which, by compactness of A, further implies that $x_l \mapsto f(x_l \mid x_{-l})$ is uniformly continuous and is bounded.

In light of

$$\sup_{r\in[a,b]} |(\text{I})| = \sup_{r\in[a,b]} |\mathbb{E}\, J_{t,r}^l(X_{-l})| \le \mathbb{E} \sup_{r\in[a,b]} |J_{t,r}^l(X_{-l})|,$$

to show that LHS converges to zero through showing $\mathbb{E}\sup_{r\in[a,b]} |J_{t,r}^l(X_{-l})| \to 0$, by DCT, it remains to show that $\sup_{r\in[a,b]} |J_{t,r}^l(X_{-l})|$ is dominated by an integrable random variable for all sufficiently small t. We have

$$\begin{aligned}\sup_{r\in[a,b]} |J_{t,r}^l(X_{-l})| \le &\underbrace{\sup_{r\in[a,b]} t^{-1}\left|\mathbb{P}\left(X_l + th_t(X_l) \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right) - \mathbb{P}\left(X_l \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right)\right|}_{\text{(I-a)}} \\ &+ \underbrace{\sup_{r\in[a,b]} |f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}^t; r))|}_{\text{(I-b)}}.\end{aligned}$$

First, by Eqs. (42), (43) and (46), whenever $t < T_\delta$, we have $R_l^{-1}(X_{-l}^t; r) \in A_{\delta, x_{-l}}^l$ and hence (I-b) $\le f_{\max}\|h\|_\infty$ by Eq. (41). We now argue that (I-a) is also bounded when $t < T_\delta$. To see the upper bound, observe that

$$\begin{aligned}
&\sup_{r\in[a,b]} t^{-1}\left\{\mathbb{P}\left(X_l + t h_t(X_l) \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right) - \mathbb{P}\left(X_l \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right)\right\} \\
\le &\sup_{r\in[a,b]} t^{-1}\left\{\mathbb{P}\left(X_l \le R_l^{-1}(X_{-l}^t; r) + t\|h_t\|_\infty \mid X_{-l}\right) - \mathbb{P}\left(X_l \le R_l^{-1}(X_{-l}^t; r) \mid X_{-l}\right)\right\} \\
= &\sup_{r\in[a,b]} f(X_l^*(r) \mid X_{-l})\,\|h_t\|_\infty \le f_{\max}\|h_t\|_\infty, \quad \forall\, t < T_\delta,
\end{aligned}$$

where we used the mean value theorem and $X_l^*(r) \in [R_l^{-1}(X_{-l}^t; r), (R_l^{-1}(X_{-l}^t; r) + t\|h_t\|_\infty) \wedge 1]$ is a random variable that depends on X_{-l} and r. The final upper bound follows from the fact that $X_l^*(r) \in A_{\delta, X_{-l}}^l$ (see Assumption D.2) almost surely for any $r \in [a, b]$ and every $t < T_\delta$. By a similar argument, (I-a) is also lower bounded when $t < T_\delta$. Hence, we can apply DCT and conclude that (I) converges to zero uniformly over $r \in [a, b]$.

Bounding (II) We write (II) $= \mathbb{E}\, V_{t,r}^l(X_{-l})$ with

$$V_{t,r}^l(X_{-l}) := f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}^t; r)) - f(R_l^{-1}(X_{-l}; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}; r)). \tag{49}$$

We first show the pointwise convergence

$$\mathbb{E}\, V_{t,r}^l(X_{-l}) \to 0, \quad r \in [a, b]. \tag{50}$$

Fix any $r \in [a, b]$. Recall that whenever $t < T_\delta$, we have $R_l^{-1}(X_{-l}^t; r) \in A_{\delta, x_{-l}}^l$ and hence $|f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})| \le f_{\max}$, $|V_{t,r}^l(X_{-l})| \le 2 f_{\max}\|h\|_\infty$. By DCT, it remains to show that $V_{t,r}^l(X_{-l}) \to 0$ for almost every X_{-l}. By monotonicity of R, for every fixed $r \in [a, b]$, $x_{-l} \mapsto -R_l^{-1}(x_{-l}; r)$ is coordinatewise non-decreasing. By Proposition D.2, we conclude that for every $r \in [a, b]$, $x_{-l} \mapsto R_l^{-1}(x_{-l}; r)$ is continuous at Lebesgue almost every $x_{-l} \in [0, 1]^{L-1}$. Hence, together with the continuity of $x_l \mapsto f(x_l \mid X_{-l})$ (Assumption D.1) and the continuity of h, we have

$$f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l}) \to f(R_l^{-1}(X_{-l}; r) \mid X_{-l}), \quad h(R_l^{-1}(X_{-l}^t; r)) \to h(R_l^{-1}(X_{-l}; r)), \quad \text{a.e.}.$$

We now argue that Eq. (50) can be strengthened to uniform convergence over $r \in [a, b]$. By Lemma D.2, it suffices to show that, for K' a dense subset of $[a, b]$,

$$\left|\mathbb{E}\, V_{t,r_t}^l(X_{-l}) - \mathbb{E}\, V_{t,r}^l(X_{-l})\right| \to 0, \quad \text{for any } r \in K' \text{ and any } [a, b] \ni r_t \to r. \tag{51}$$

Fix an arbitrary $r \in K'$ and $r_t \to r$. We have

$$\begin{aligned}
&\left|\mathbb{E}\, V_{t,r_t}^l(X_{-l}) - \mathbb{E}\, V_{t,r}^l(X_{-l})\right| \\
\le &\underbrace{\mathbb{E}\left|f(R_l^{-1}(X_{-l}^t; r_t) \mid X_{-l})\, h(R_l^{-1}(X_{-l}^t; r_t)) - f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}^t; r))\right|}_{\text{(II-a)}} \\
&+ \underbrace{\mathbb{E}\left|f(R_l^{-1}(X_{-l}; r_t) \mid X_{-l})\, h(R_l^{-1}(X_{-l}; r_t)) - f(R_l^{-1}(X_{-l}; r) \mid X_{-l})\, h(R_l^{-1}(X_{-l}; r))\right|}_{\text{(II-b)}}.
\end{aligned} \tag{52}$$

First, we claim that (II-b) tends to zero, which follows from $R_l^{-1}(X_{-l}; r_t) \to R_l^{-1}(X_{-l}; r)$ for a.e. X_{-l} (implied by Proposition D.2 and monotonicity of R_l^{-1}; see also the proof of Lemma D.3) and DCT. Then, for $t < T_\delta$, note that (II-a) is further bounded by

$$\begin{aligned}\text{(II-a)} \le \|h\|_\infty \, \mathbb{E}\left|f(R_l^{-1}(X_{-l}^t; r_t) \mid X_{-l}) - f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\right| \\ + f_{\max} \, \mathbb{E}\left|h(R_l^{-1}(X_{-l}^t; r_t)) - h(R_l^{-1}(X_{-l}^t; r))\right|.\end{aligned}$$

We shall prove Eq. (51) by showing that

$$\mathbb{E}\left|f(R_l^{-1}(X_{-l}^t; r_t) \mid X_{-l}) - f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\right| \to 0, \tag{53}$$

$$\mathbb{E}\left|h(R_l^{-1}(X_{-l}^t; r_t)) - h(R_l^{-1}(X_{-l}^t; r))\right| \to 0. \tag{54}$$

We now show Eq. (53). Recall that X_{-l}^t is given by Eq. (45) and the expectation in Eq. (53) is taken over X_{-l}. For any $\varepsilon > 0$, let

$$\mathcal{E}_{\varepsilon,t} := \{|R_l^{-1}(X_{-l}^t; r_t) - R_l^{-1}(X_{-l}^t; r)| > \varepsilon\}.$$

Considering the integral over $\mathcal{E}_{\varepsilon,t}$ and its complement, for $t < T_\delta$, we have

$$\begin{aligned}\mathbb{E}\left|f(R_l^{-1}(X_{-l}^t; r_t) \mid X_{-l}) - f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\right| &\le 2f_{\max}\,\mathbb{P}(\mathcal{E}_{\varepsilon,t}) + \mathbb{E}\left[\Delta f_l(\varepsilon \mid X_{-l}) \mid \mathcal{E}_{\varepsilon,t}^c\right]\mathbb{P}(\mathcal{E}_{\varepsilon,t}^c) \\ &\le 2f_{\max}\,\mathbb{P}(\mathcal{E}_{\varepsilon,t}) + \mathbb{E}\left[\Delta f_l(\varepsilon \mid X_{-l}) \mid \mathcal{E}_{\varepsilon,t}^c\right],\end{aligned}$$

where $\Delta f_l(\cdot \mid x_{-l})$ is the modulus of continuity of $x_l \mapsto f(x_l \mid x_{-l})$ on $A_{\delta,x_{-l}}^l$. Applying the reverse Fatou's lemma, we derive

$$\begin{aligned}&\limsup \mathbb{E}\left|f(R_l^{-1}(X_{-l}^t; r_t) \mid X_{-l}) - f(R_l^{-1}(X_{-l}^t; r) \mid X_{-l})\right| \\ &\le 2f_{\max} \limsup \mathbb{P}(\mathcal{E}_{\varepsilon,t}) + \limsup \mathbb{E}\left[\Delta f_l(\varepsilon \mid X_{-l}) \mid \mathcal{E}_{\varepsilon,t}^c\right] \\ &= \limsup \mathbb{E}\left[\Delta f_l(\varepsilon \mid X_{-l}) \mid \mathcal{E}_{\varepsilon,t}^c\right] \\ &\le \Delta f_l(\varepsilon), \quad \text{for all sufficiently small } \varepsilon > 0.\end{aligned}$$

For the penultimate step, we used Lemma D.3 to conclude $\mathbb{P}(\mathcal{E}_{\varepsilon,t}) \to 0$. For the final step, we used the equicontinuity of f on A_δ^l (Assumption D.2) to conclude that $\Delta f_l(\varepsilon \mid x_{-l})$ for $x_{-l} \in [0,1]^{L-1}$ admits an upper bound $\Delta f_l(\varepsilon)$ for all sufficiently small $\varepsilon > 0$. Because $\Delta f_l(\varepsilon)$ can be made arbitrarily small by choosing $\varepsilon \searrow 0$, Eq. (53) is proven.

Finally, Eq. (54) can be shown by a similar argument, which together with Eq. (53) proves Eq. (51). To complete the proof, observe that the continuity of the derivative $\varphi'_{g=0}(h)$ on $[a,b]$ follows from the fact that (II-b) in Eq. (52) tends to zero. □

D.2 First-order behaviour of rank-transformed subsampling

For technical reasons, Theorem 4 is established for $\widetilde{G}_n$ computed from a variant of Algorithm 2 that has two independent copies of the data, X and X'. The variant first uses X to define the rank transform, and then uses X' to approximate the distribution of the rank-transformed statistic, whereas the original Algorithm 2 uses the same X for both purposes. In this variant described below, the randomness pertaining to the first and the second subsampling are independent and this greatly simplifies our analysis.

First subsampling (for defining the rank transform) Recall that we choose $B = J\lfloor n/m \rfloor$ for a fixed integer J. Also recall that $\widehat{\mathbf{H}} = (\widehat{H}_{b,l})$ is a $B \times L$ matrix consisting of rows

$$\widehat{\mathbf{H}}_{b,\cdot} := \left(T_m^{(1)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}}), \quad \ldots \quad , T_m^{(L)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}})\right), \quad b = 1, \ldots, B.$$

Suppose the B rows are arranged into J blocks of $\lfloor n/m \rfloor$ rows so that $b = (j-1)\lfloor n/m \rfloor + 1, \ldots, j\lfloor n/m \rfloor$ consists of $\lfloor n/m \rfloor$ independent subsamples. Let $\mathbb{F}_{\widehat{\mathbf{H}}}$ be the empirical distribution function of $\{\widehat{H}_{b,l} : b = 1, \ldots, B,\ l = 1, \ldots, L\}$. The rank-transform is the following map from T_m to $\widetilde{T}_m$:

$$\widetilde{T}_m = F_0^{-1}\left(\frac{BL}{BL+1}\mathbb{F}_{\widehat{\mathbf{H}}}(T_m) + \frac{1/2}{BL+1}\right), \tag{55}$$

where the normalised rank is bounded away from zero and one by $(1/2)/(BL+1)$.

Second subsampling (for estimating the distribution of rank-transformed statistic) We study the version of the procedure where the second subsampling is performed on X', an independent copy of the data. The rank-transformed matrix $\widetilde{\mathbf{H}} = (\widetilde{H}_{b,l})$ consists of rows

$$\widetilde{\mathbf{H}}_{b,\cdot} := \left(\widetilde{T}_m^{(1)}(X'_{i_{1,b}}, \ldots, X'_{i_{m,b}}), \quad \ldots \quad , \widetilde{T}_m^{(L)}(X'_{i_{1,b}}, \ldots, X'_{i_{m,b}})\right), \quad b = 1, \ldots, B,$$

where $\widetilde{T}_m^{(l)}(X'_{i_{1,b}}, \ldots, X'_{i_{m,b}})$ is the image of $T_m^{(l)}(X'_{i_{1,b}}, \ldots, X'_{i_{m,b}})$ under the rank transform. Passing them through the aggregation function S, we get

$$\widetilde{S}_b := S(\widetilde{H}_{b,1}, \ldots, \widetilde{H}_{b,L}), \quad b = 1, \ldots, B$$

Let $\widetilde{G}_n(x) := \mathbb{F}_{\{\widetilde{S}_b\}}(x)$. For $\alpha \in (0,1)$, the rank-transformed critical value is $\widetilde{G}_n^{-1}(1-\alpha)$. We compare this critical value to the oracle critical value $G_{P_0}^{-1}(1-\alpha)$, where G is the asymptotic null distribution function of S.

The full version of Theorem 4 is stated as follows.

Theorem D.1 (Full statement of Theorem 4)**.** *Suppose Condition 1 holds and $(T_n^{(1)}, \ldots, T_n^{(L)})$ is exchangeable. Consider a sequence $P_n \in \mathcal{P}$ that converges in copula to some $P_0 \in \mathcal{P}_0$ in the sense of Definition 1 such that*

$$\mathrm{d}_{TV}(U_m, C) = o(\sqrt{m/n}). \tag{56}$$

Suppose the distribution of C is absolutely continuous with respect to the Lebesgue measure. Let S be a symmetric aggregation function that is non-decreasing in each coordinate. Let $\widetilde{G}_n$ denote the rank-transformed subsampling distribution function obtained with a variant of Algorithm 2 that uses two independent copies of the data under $(X, X', \Omega) \sim P_n^n \times P_n^n \times P_\Omega$.

Suppose Assumption 1 holds and fix $\alpha \in (0,1)$ such that the density G'_{P_0} is strictly positive and continuous in a neighbourhood of $G_{P_0}^{-1}(1-\alpha)$. With $c := (c_1, \ldots, c_L)$, define

$$R(c_1, \ldots, c_L) := S(F_0^{-1}(c_1), \ldots, F_0^{-1}(c_L)), \quad c \in [0,1]^L.$$

Then, under Assumptions D.3 and D.4, it holds that

$$\mathbb{E}\left[-M \vee \sqrt{n/m}\,\left(\widetilde{G}_n^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha)\right) \wedge M\right] \to 0. \tag{57}$$

Further, let G_{n,P_n} be the distribution function of $S(T_n^{(1)}, \ldots, T_n^{(L)})$ under $P_n^n \times P_\Omega$. Suppose $\|G_{n,P_n} - G_{alt}\|_\infty = o(\sqrt{m/n})$ holds for some distribution function G_{alt} that is differentiable at $G_{P_0}^{-1}(1-\alpha)$. Then, under the same set of assumptions, for any $M > 0$, we also have

$$\mathbb{E}\left[-M \vee \sqrt{n/m}\,\left(G_{n,P_n}(\widetilde{G}_n^{-1}(1-\alpha)) - G_{n,P_n}(G_{P_0}^{-1}(1-\alpha))\right) \wedge M\right] \to 0. \tag{58}$$

Proof. By Lemma D.4, we prove Eq. (57) by showing given any subsequence n_k of n, n_k admits a further subsequence n_ι such that $\sqrt{n_\iota/m_\iota}\left(\widetilde{G}_{n_\iota}^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha)\right)$ converges to a symmetric law. Similarly, we establish Eq. (58) by showing that

$$\sqrt{n/m}\left(G_{n,P_n}(G_{P_0}^{-1}(1-\alpha)) - G_{n,P_n}(\widetilde{G}_n^{-1}(1-\alpha))\right)$$

converges to a symmetric limit along the subsequence indexed by ι. We will first prove Eq. (57), after which Eq. (58) follows easily. In what follows, let n_k be an arbitrary subsequence of n.

First subsampling Let

$$U_{b,l} := F_{m,P_n}(\widehat{H}_{b,l}) = F_{m,P_n}\left(T(X_{i_{1,b}},\dots,X_{i_{m,b}};\Omega^{(b,l)})\right), \quad b=1,\dots,B,\ l=1,\dots,L.$$

By construction, $\widehat{H}_{b,l} = F_{m,P_n}^{-1}(U_{b,l})$. Because $\widehat{H}_{b,l} = F_{m,P_n}^{-1}(U_{b,l}) \le x$ iff $U_{b,l} \le F_{m,P_n}(x)$ for every x (van der Vaart, 2000, Lemma 21.1), $\mathbb{F}_{\widehat{\mathbf{H}}}$ can be written as

$$\mathbb{F}_{\widehat{\mathbf{H}}}(x) = \mathbb{F}_{\mathbf{U},n} \circ F_{m,P_n}(x). \tag{59}$$

With reindexing

$$U_{i,j,l} := U_{(j-1)\lfloor n/m\rfloor+i,\,l}, \quad i=1,\dots,\lfloor n/m\rfloor,\ j=1,\dots,J,\ l=1,\dots,L,$$

observe that for each (j,l), $U_{1,j,l},\dots,U_{\lfloor n/m\rfloor,j,l}$ are iid with a common distribution function, which we denote as $F_{U_m^{(1)},P_n}$. Using Eq. (56), we have

$$\|F_{U_m^{(1)},P_n} - \mathrm{Id}\,\|_\infty \le \mathrm{d_{TV}}(U_m^{(1)}, C_1) \le \mathrm{d_{TV}}(U_m, C) = o(\sqrt{m/n}).$$

Applying Lemma D.5 to the subsequence n_k, we conclude that n_k admits a further a subsequence n_j such that

$$\sqrt{n_j/m_j}\left(\mathbb{F}_{\mathbf{U},n_j} - \mathrm{Id}\right) \rightsquigarrow \xi_1 \quad \text{in } D[0,1], \tag{60}$$

where ξ_1 is almost surely a uniformly continuous with $\xi_1(0) = \xi_1(1) = 0$ and satisfies $\xi_1 =_d -\xi_1$.

Second subsampling, population version Define

$$g_n(u) := (\mathbb{F}_{\mathbf{U},n} - \mathrm{Id})(u) - \frac{1}{BL+1}\mathbb{F}_{\mathbf{U},n}(u) + \frac{1/2}{BL+1}, \tag{61}$$

which is a random element in $D_\varphi \subset D[0,1]$ (see Corollary D.1 for definition). For $l=1,\dots,L$, the rank transform can be written as $\widetilde{T}_m^{(l)} = F_0^{-1}(\widetilde{U}_m^{(l)})$ with

$$\begin{aligned}
\widetilde{U}_m^{(l)} &= \frac{BL}{BL+1}\,\mathbb{F}_{\widehat{\mathbf{H}}}(T_m^{(l)}) + \frac{1/2}{BL+1} \\
&\overset{(i)}{=} \frac{BL}{BL+1}\,\mathbb{F}_{\mathbf{U},n} \circ F_{m,P_n} \circ F_{m,P_n}^{-1}(U_m^{(l)}) + \frac{1/2}{BL+1} \\
&\overset{(ii)}{=} \frac{BL}{BL+1}\,\mathbb{F}_{\mathbf{U},n}(U_m^{(l)}) + \frac{1/2}{BL+1} \\
&= U_m^{(l)} + g_n(U_m^{(l)}),
\end{aligned}$$

where (i) uses Eqs. (6) and (59), (ii) uses the fact that $U_m^{(l)}$ is in the range of F_{m,P_n} by construction (van der Vaart, 2000, Lemma 21.1). In above, by the independence between X and X', we

know $T_m^{(l)}$ (and hence $U_m^{(l)}$) is independent of $\mathbb{F}_{\mathbf{U},n}$ (and hence g_n). Further, the rank-transformed aggregated statistic

$$\widetilde{S}_m = S(\widetilde{T}_m^{(1)}, \ldots, \widetilde{T}_m^{(L)}) = R(\widetilde{U}_m^{(1)}, \ldots, \widetilde{U}_m^{(L)}),$$

has its population distribution given by

$$\begin{aligned}\widetilde{G}_{\mathbb{F}_{\mathbf{U},n}}(r) &:= \mathbb{P}\left(R(\widetilde{U}_m^{(1)}, \ldots, \widetilde{U}_m^{(L)}) \leq r \mid \mathbb{F}_{\mathbf{U},n}\right) \\ &= \mathbb{P}\left(R(U_m^{(1)} + g_n(U_m^{(1)}), \ldots, U_m^{(L)} + g_n(U_m^{(L)})) \leq r \mid \mathbb{F}_{\mathbf{U},n}\right).\end{aligned}$$

Observe that G_{P_0} is simply the distribution function of $S(F_0^{-1}(C_1), \ldots, F_0^{-1}(C_L))$ under Condition 1 and Assumption 1. Using the definition of $R(\cdot)$, we can write

$$G_{P_0}(r) = \mathbb{P}\left(R(C_1, \ldots, C_L) \leq r\right).$$

Using Eq. (60) and $B = J\lfloor n/m \rfloor$, we have

$$\sqrt{n_j/m_j}(g_{n_j} - 0) \rightsquigarrow \xi_1. \tag{62}$$

By our assumption that S is symmetric and monotone, R is symmetric and non-decreasing in each coordinate. Also, by our assumption, C has a distribution that is absolutely continuous w.r.t. the Lebesgue measure. Given Assumptions D.3 and D.4, we can apply Corollary D.1 and conclude that the map

$$\varphi : D_\varphi \to D[a,b], \quad \varphi(g) = r \mapsto \mathbb{P}\left(R(C_1 + g(C_1), \ldots, C_L + g(C_L)) \leq r\right)$$

is Hadamard differentiable at $g = 0$ tangentially to any uniformly continuously $h \in D[0,1]$ with $h(0) = h(1) = 0$. These conditions are satisfied almost surely by ξ_1 in Eq. (62). Applying the functional delta method (van der Vaart, 2000, Theorem 20.8) along the subsequence n_j, we have the following weak convergence in $D[a,b]$:

$$\begin{aligned}r \mapsto \sqrt{n_j/m_j}\left\{\mathbb{P}(R\left(C_1 + g_{n_j}(C_1), \ldots, C_L + g_{n_j}(C_L)\right) \leq r) - G_{P_0}(r)\right\} \\ \rightsquigarrow \eta(\xi_1) := \left(r \mapsto -L\,\mathbb{E}\left[f(R^{-1}(C_{-1}; r) \mid C_{-1})\,\xi_1(R^{-1}(C_{-1}; r)) \mid \xi_1\right]\right),\end{aligned} \tag{63}$$

where the notation $\mathbb{E}[\cdot \mid \xi_1]$ highlights that the expectation is only taken over C. In addition, by Corollary D.1, $\eta(\xi_1)$ is almost surely a continuous function of r on $[a,b]$.

Note that the total variation bound in Eq. (56) implies

$$\begin{aligned}\sup_r \Big| \mathbb{P}\left(R(U_m^{(1)} + g_n(U_m^{(1)}), \ldots, U_m^{(L)} + g_n(U_m^{(L)})) \leq r \mid \mathbb{F}_{\mathbf{U},n}\right) \\ - \mathbb{P}\left(R(C_1 + g_n(C_1), \ldots, C_L + g_n(C_L)) \leq r\right) \Big| = o(\sqrt{m/n})\end{aligned}$$

and it follows that

$$\sqrt{n_j/m_j}\left(\widetilde{G}_{\mathbb{F}_{\mathbf{U},n}} - G_{P_0}\right) \rightsquigarrow \eta(\xi_1) \quad \text{in } D[a,b], \tag{64}$$

where $\eta(\xi_1) =_d -\eta(\xi_1)$ because $\xi_1 =_d -\xi_1$ and $\eta(\xi_1)$ is linear in ξ_1.

Second subsampling, sample version Now consider the second subsample $\{\widetilde{S}_b : b = 1, \dots, B\}$ with $\widetilde{S}_b = S(\widetilde{H}_{b,1}, \dots, \widetilde{H}_{b,L})$. For every $\widetilde{S}_b$, draw a random U'_b independently according to

$$U'_b \mid \widetilde{S}_b,\ \mathbb{F}_{\mathbf{U},n} \sim \mathrm{unif}[\widetilde{G}_{\mathbb{F}_{\mathbf{U},n}}(S_b-),\ \widetilde{G}_{\mathbb{F}_{\mathbf{U},n}}(S_b)].$$

Then, by construction,

$$U'_b \sim \mathrm{unif}(0,1), \quad \widetilde{S}_b = \widetilde{G}^{-1}_{\mathbb{F}_{\mathbf{U},n}}(U'_b),$$

and consequently, the rank-transformed subsampling critical value can be expressed as

$$\widetilde{G}^{-1}_n(1-\alpha) = \widetilde{G}^{-1}_{\mathbb{F}_{\mathbf{U},n}} \circ \mathbb{F}^{-1}_{\mathbf{U}',n}(1-\alpha), \quad \alpha \in (0,1), \tag{65}$$

where we use $\mathbb{F}_{\mathbf{U}',n}$ to denote the empirical distribution function of $\{U'_b\}$. If we reindex $U'_{i,j} := U'_b$ for $b = (j-1)\lfloor n/m \rfloor + i$, we see that $U'_{1,j}, \dots, U'_{\lfloor n/m \rfloor, j}$ are iid unif$(0,1)$ for every j irrespective of $\mathbb{F}_{\mathbf{U},n}$. It follows from Lemma D.5 and Eq. (64) that there exists a further subsequence n_ι of n_j, such that

$$\sqrt{n_\iota/m_\iota}\left(\mathbb{F}_{\mathbf{U}',n_\iota} - \mathrm{Id},\ \widetilde{G}_{\mathbb{F}_{\mathbf{U},n_\iota}} - G_{P_0}\right) \rightsquigarrow (\xi_2,\ \eta(\xi_1)) \quad \text{in } D[0,1] \times D[a,b], \tag{66}$$

where ξ_2 is a random element in $D[0,1]$ such that $\xi_2 =_d -\xi_2$ and ξ_2 is uniformly continuous almost surely. Further, because the second subsampling is performed on a separate sample X', we have $\xi_1 \perp\!\!\!\perp \xi_2$.

Using the first limit in Eq. (66), by van der Vaart (2000, Lemma 21.3) we have

$$\sqrt{n_\iota/m_\iota}\left(\mathbb{F}^{-1}_{\mathbf{U}',n_\iota}(1-\alpha) - (1-\alpha)\right) \to_d -\xi_2(1-\alpha) \tag{67}$$

and consequently,

$$\sqrt{n_\iota/m_\iota}\left(G^{-1}(\mathbb{F}^{-1}_{\mathbf{U}',n_\iota}(1-\alpha)) - G^{-1}_{P_0}(1-\alpha)\right) \to_d \frac{-\xi_2(1-\alpha)}{g_{P_0}(G^{-1}_{P_0}(1-\alpha))} \tag{68}$$

by the delta method.

Further, by our assumption, because the density $g_{P_0} := G'_{P_0}$ is strictly positive and continuous in a neighbourhood $[a,b] \supset [G^{-1}_{P_0}(1-\alpha) - \delta, G^{-1}_{P_0}(1-\alpha) + \delta]$ for some $\delta > 0$, using van der Vaart (2000, Lemma 21.4) and the functional delta method, from the second limit in Eq. (66) we derive

$$\sqrt{n_\iota/m_\iota}\,(\widetilde{G}^{-1}_{\mathbb{F}_{\mathbf{U},n_\iota}} - G^{-1}_{P_0}) \rightsquigarrow -(\eta(\xi_1)/g_{P_0}) \circ G^{-1}_{P_0} \quad \text{in } \ell^\infty[1-\alpha-\delta', 1-\alpha+\delta'] \tag{69}$$

for some $\delta' > 0$.

In fact, Eqs. (67) to (69) hold jointly in light of Eq. (66), which implies

$$\left(\sqrt{n_\iota/m_\iota}\,(\widetilde{G}^{-1}_{\mathbb{F}_{\mathbf{U},n_\iota}} - G^{-1}_{P_0}),\ \mathbb{F}^{-1}_{\mathbf{U}',n_\iota}(1-\alpha)\right) \rightsquigarrow \left(-(\eta(\xi_1)/g_{P_0}) \circ G^{-1}_{P_0},\ 1-\alpha\right)$$

in $\ell^\infty[1-\alpha-\delta', 1-\alpha+\delta'] \times \mathbb{R}$. Observe that $(h,z) \mapsto h(z)$, as a $\ell^\infty \times \mathbb{R} \to \mathbb{R}$ map, is continuous when h is continuous at z. Recall from Eq. (63) that $\eta(\xi_1)$ is almost surely a continuous function of r on $[a,b]$, which verifies the continuity of $-(\eta(\xi_1)/g_{P_0}) \circ G^{-1}_{P_0}$ at $1-\alpha$. Hence, we can apply the continuous mapping theorem (van der Vaart, 2000, Theorem 18.11) and conclude

$$\sqrt{n_\iota/m_\iota}\,(\widetilde{G}^{-1}_{\mathbb{F}_{\mathbf{U},n_\iota}} - G^{-1}_{P_0}) \circ \mathbb{F}^{-1}_{\mathbf{U}',n_\iota}(1-\alpha) \to_d -(\eta(\xi_1)/g_{P_0}) \circ G^{-1}_{P_0}(1-\alpha). \tag{70}$$

Finally, from Eqs. (65), (68) and (70) it follows that

$$\begin{aligned}
&\sqrt{n_\iota/m_\iota}\,\left(\widetilde{G}^{-1}_{n_\iota}(1-\alpha) - G^{-1}_{P_0}(1-\alpha)\right)\\
&= \sqrt{n_\iota/m_\iota}\,\left(\widetilde{G}^{-1}_{\mathbb{F}_{\mathbf{U},n_\iota}} - G^{-1}_{P_0}\right) \circ \mathbb{F}^{-1}_{\mathbf{U}',n_\iota}(1-\alpha) + \sqrt{n_\iota/m_\iota}\,\left(G^{-1}_{P_0}(\mathbb{F}^{-1}_{\mathbf{U}',n_\iota}(1-\alpha)) - G^{-1}_{P_0}(1-\alpha)\right)\\
&\to_d -(\eta(\xi_1)/g_{P_0}) \circ G^{-1}_{P_0}(1-\alpha) - \frac{\xi_2(1-\alpha)}{g_{P_0}(G^{-1}_{P_0}(1-\alpha))}.
\end{aligned} \tag{71}$$

Observe that the limit is symmetric because $\eta(\xi_1) =_d -\eta(\xi_1)$, $\xi_2 =_d -\xi_2$ and $\xi_1 \perp\!\!\!\perp \xi_2$. As this limit holds for n_ι, which is a subsequence of n_j and hence of n_k, this establishes our first result Eq. (57).

We now show Eq. (58) under the additional assumption that $\|G_{n,P_n} - G_{\text{alt}}\|_\infty = o(\sqrt{m/n})$ for a distribution function G_{alt} that is differentiable at $G^{-1}_{P_0}(1-\alpha)$. Again, by Lemma D.4, it suffices to show that

$$\sqrt{n_\iota/m_\iota}\,\left(G_{n_\iota,P_{n_\iota}}(G^{-1}_{P_0}(1-\alpha)) - G_{n_\iota,P_{n_\iota}}(\widetilde{G}^{-1}_{n_\iota}(1-\alpha))\right)$$

converges to a symmetric limit law. This simply follows from

$$\begin{aligned}
&\sqrt{n_\iota/m_\iota}\,\left(G_{n_\iota,P_{n_\iota}}(G^{-1}_{P_0}(1-\alpha)) - G_{n_\iota,P_{n_\iota}}(\widetilde{G}^{-1}_{n_\iota}(1-\alpha))\right)\\
&= \sqrt{n_\iota/m_\iota}\,\left(G_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) - G_{\text{alt}}(\widetilde{G}^{-1}_{n_\iota}(1-\alpha))\right) + O_p\left(\sqrt{n_\iota/m_\iota}\|G_{n_\iota,P_{n_\iota}} - G_{\text{alt}}\|_\infty\right)\\
&= \sqrt{n_\iota/m_\iota}\,\left(G_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) - G_{\text{alt}}(\widetilde{G}^{-1}_{n_\iota}(1-\alpha))\right) + o_p(1)
\end{aligned}$$

and an application of the delta method based on Eq. (71). □

D.3 First-order behaviour of ordinary subsampling

The result below formalises the first-order bias in approximating the oracle critical value with ordinary subsampling and the ensuing power loss; see also Appendix D.5 for numerical demonstration of the bias.

Theorem D.2 (First-order behaviour of ordinary subsampling)**.** *Consider a sequence $P_n \in \mathcal{P}$. Let G_{m,P_n} be the distribution function of $S(T^{(1)}_m, \dots, T^{(L)}_m)$ under $P^m_n \times P_\Omega$. Fix $\alpha \in (0,1)$ such that*

$$g_{m,P_n}(G^{-1}_{m,P_n}(x_n)) \to c > 0, \quad \text{for every sequence } x_n \to 1-\alpha, \tag{72}$$

where $g_{m,P_n} := G'_{m,P_n}$. Let $\widehat{G}_n := \mathbb{F}_{\{\widehat{S}_b : b=1,\dots,B\}}$ be the ordinary subsampling (without the rank transform; see Section 2.2) estimate of G_{P_0}. Then, for every $M > 0$, it holds that

$$\mathbb{E}\left[-M \vee \sqrt{n/m}\,\left(\widehat{G}^{-1}_n(1-\alpha) - G^{-1}_{m,P_n}(1-\alpha)\right) \wedge M\right] \to 0, \tag{73}$$

where the expectation is taken under $P^n_n \times P_\Omega$.

Further, suppose that for some $\beta \in (0, 1/2]$, we have $(n/m)^\beta(G^{-1}_{m,P_n}(1-\alpha) - G^{-1}_{P_0}(1-\alpha)) \to \tau > 0$. Suppose $\|G_{n,P_n} - G_{alt}\|_\infty = o((m/n)^\beta)$ holds for some distribution function G_{alt} continuously differentiable at $G^{-1}_{P_0}(1-\alpha)$. Then, under the same set of assumptions, for any $M > 0$, we also have

$$\begin{aligned}
\mathbb{E}\Big\{ - M \vee \Big[(n/m)^\beta\,\left\{G_{n,P_n}(\widehat{G}^{-1}_n(1-\alpha)) - G_{n,P_n}(G^{-1}_{P_0}(1-\alpha))\right\}\\
- \tau\, G'_{alt}(G^{-1}_{P_0}(1-\alpha))\Big] \wedge M\Big\} \to 0.
\end{aligned} \tag{74}$$

Proof. Using Lemma D.4, we prove Eq. (73) by showing that given any subsequence n_k of n, n_k admits a further subsequence n_ι such that $\sqrt{n_\iota/m_\iota}\left(\widehat{G}_{n_\iota}^{-1}(1-\alpha) - G_{m_\iota,P_{n_\iota}}^{-1}(1-\alpha)\right)$ converges to a symmetric law. In what follows, let n_k be an arbitrary subsequence of n.

With $\widehat{U}_{b,n} := G_{m,P_n}(\widehat{S}_{b,n})$ for $b = 1,\dots,B$, we have $\widehat{S}_{b,n} = G_{m,P_n}^{-1}(\widehat{U}_{b,n})$ and

$$\widehat{G}_n^{-1}(1-\alpha) = G_{m,P_n}^{-1} \circ \mathbb{F}_{\{\widehat{U}_{b,n}\}}^{-1}(1-\alpha), \tag{75}$$

where $\mathbb{F}_{\{\widehat{U}_{b,n}\}}$ is the empirical distribution of $\{\widehat{U}_{b,n}\}$. Further, if we reindex $\widehat{U}'_{i,j,n} := \widehat{U}_{b,n}$ for $b = (j-1)\lfloor n/m \rfloor + i$, we see that $\widehat{U}'_{1,j,n},\dots,U'_{\lfloor n/m\rfloor,j,n}$ are iid unif$(0,1)$ for every $j = 1,\dots,J$. By Lemma D.5, there exists a further subsequence n_ι of n_k such that

$$\sqrt{n_\iota/m_\iota}\left(\mathbb{F}_{\{\widehat{U}_{b,n_\iota}\}} - \mathrm{Id}\right) \rightsquigarrow \xi \quad \text{in } D[0,1],$$

where $\xi =_d -\xi$. It then follows from van der Vaart (2000, Lemma 21.3) that

$$\sqrt{n_\iota/m_\iota}\left(\mathbb{F}_{\{\widehat{U}_{b,n_\iota}\}}^{-1}(1-\alpha) - (1-\alpha)\right) \to_d -\xi(1-\alpha), \tag{76}$$

where the limit law is symmetric.

Using $(G_{m,P_n}^{-1})' = 1/\left(g_{m,P_n} \circ G_{m,P_n}^{-1}\right)$ and the mean value theorem, from Eq. (75) we derive

$$\sqrt{n_\iota/m_\iota}\left(\widehat{G}_{n_\iota}^{-1}(1-\alpha) - G_{m_\iota,P_{n_\iota}}^{-1}(1-\alpha)\right) = C_{n_\iota}^{-1}\sqrt{n_\iota/m_\iota}\left[\mathbb{F}_{\{\widehat{U}_{b,n_\iota}\}}^{-1}(1-\alpha) - (1-\alpha)\right]$$

for $C_{n_\iota} := g_{m_\iota,P_{n_\iota}} \circ G_{m_\iota,P_{n_\iota}}^{-1}(\zeta_{n_\iota})$ and some $\zeta_{n_\iota} \to_p 1-\alpha$. By our assumption Eq. (72), we have $C_{n_\iota} \to_p c > 0$ by van der Vaart (2000, Theorem 18.11) and hence

$$\sqrt{n_\iota/m_\iota}\left(\widehat{G}_{n_\iota}^{-1}(1-\alpha) - G_{m_\iota,P_{n_\iota}}^{-1}(1-\alpha)\right) \to_d -c^{-1}\,\xi(1-\alpha), \tag{77}$$

where the limit law is symmetric. Because n_ι is a subsequence of n_k, this proves Eq. (73).

Finally, we show Eq. (74) under the given additional assumptions. Let N be the neighbourhood of $G_{P_0}^{-1}(1-\alpha)$ on which G_{alt} is differentiable. By the mean value theorem, we have that

$$G_{\mathrm{alt}} \circ \widehat{G}_{n_\iota}^{-1}(1-\alpha) - G_{\mathrm{alt}} \circ G_{m_\iota,P_{n_\iota}}^{-1}(1-\alpha) = G'_{\mathrm{alt}}(\xi_n)\left(\widehat{G}_{n_\iota}^{-1}(1-\alpha) - G_{m_\iota,P_{n_\iota}}^{-1}(1-\alpha)\right) \tag{78}$$

on the sequence of events $\Omega_n := \{\xi_n \in N\}$, where ξ_n is a random variable lying between $\widehat{G}_{n_\iota}^{-1}(1-\alpha)$ and $G_{m_\iota,P_{n_\iota}}^{-1}(1-\alpha)$. By Eq. (77) and as

$$(n/m)^\beta(G_{m,P_n}^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha)) \to \tau > 0, \tag{79}$$

we have that $\xi_n \to_p G_{P_0}^{-1}(1-\alpha)$. Thus $\mathbb{P}(\Omega_n) \to 1$ and by the continuous mapping theorem, $G'_{\mathrm{alt}}(\xi_n) \to_p G'_{\mathrm{alt}} \circ G_{P_0}^{-1}(1-\alpha)$. Thus, multiplying both sides of Eq. (78) by $\sqrt{n_\iota/m_\iota}$, using Slutsky's lemma and Eq. (77) we have

$$\sqrt{\frac{n_\iota}{m_\iota}}\left(G_{\mathrm{alt}} \circ \widehat{G}_{n_\iota}^{-1}(1-\alpha) - G_{\mathrm{alt}} \circ G_{m_\iota,P_{n_\iota}}^{-1}(1-\alpha)\right) \to_d -c^{-1}G'_{\mathrm{alt}} \circ G_{P_0}^{-1}(1-\alpha)\,\xi(1-\alpha). \tag{80}$$

Next, note that by our assumption Eq. (79) and the mean value theorem, we have that for all n sufficiently large,

$$(n/m)^\beta \left\{ G_{\text{alt}}(G^{-1}_{m,P_n}(1-\alpha)) - G_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) \right\} = (n/m)^\beta \{ G^{-1}_{m,P_n}(1-\alpha) - G^{-1}_{P_0}(1-\alpha) \} G'_{\text{alt}}(\zeta_n) \\ \to \tau\, G'_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)),$$

where ζ_n lies between $G^{-1}_{P_0}(1-\alpha)$ and $G^{-1}_{m,P_n}(1-\alpha)$. Using the fact that $\|G_{n,P_n} - G_{\text{alt}}\|_\infty = o((m/n)^\beta)$, we have

$$\begin{aligned}
&(n/m)^\beta \left\{ G_{n,P_n}(\widehat{G}_n^{-1}(1-\alpha)) - G_{n,P_n}(G^{-1}_{P_0}(1-\alpha)) \right\} - \tau\, G'_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) \\
&= (n/m)^\beta \left\{ G_{\text{alt}}(\widehat{G}_n^{-1}(1-\alpha)) - G_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) \right\} - \tau\, G'_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) + o(1) \\
&= (n/m)^\beta \left\{ G_{\text{alt}}(\widehat{G}_n^{-1}(1-\alpha)) - G_{\text{alt}}(G^{-1}_{m,P_n}(1-\alpha)) \right\} \\
&\quad + (n/m)^\beta \left\{ G_{\text{alt}}(G^{-1}_{m,P_n}(1-\alpha)) - G_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) \right\} - \tau\, G'_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) + o(1) \\
&= (n/m)^\beta \left\{ G_{\text{alt}}(\widehat{G}_n^{-1}(1-\alpha)) - G_{\text{alt}}(G^{-1}_{m,P_n}(1-\alpha)) \right\} + o_p(1).
\end{aligned}$$

Using Eq. (80), along the subsequence n_ι, we have

$$(n_\iota/m_\iota)^\beta \left\{ G_{n_\iota,P_{n_\iota}}(\widehat{G}_{n_\iota}^{-1}(1-\alpha)) - G_{n_\iota,P_{n_\iota}}(G^{-1}_{P_0}(1-\alpha)) \right\} - \tau\, G'_{\text{alt}}(G^{-1}_{P_0}(1-\alpha)) \\ \to_d \begin{cases} -c^{-1} G'_{\text{alt}} \circ G^{-1}_{P_0}(1-\alpha)\, \xi(1-\alpha), & \beta = 1/2 \\ 0, & \beta < 1/2. \end{cases}$$

The result then follows from applying Lemma D.4. □

D.4 Checking assumptions

In this section, we check that for choices of F_0 and S considered in this paper, Assumptions D.3 and D.4 are satisfied by a Gaussian copula. Consider a symmetric, Gaussian copula with a non-negative correlation (see Fig. D.1). Observe that $c \mapsto f(c_1 \mid c_{-1})$ is continuous except for $(0,\dots,0)$ and $(1,\dots,1)$, the two points where f diverges. This verifies Assumption D.3. Further, for R^{-1} corresponding to choices of F_0 and S considered in this paper (detailed below), it suffices to consider a region defined in Assumption D.4 that is bounded away from $(0,\dots,0)$ and $(1,\dots,1)$. Because $c \mapsto f(c_1 \mid c_{-1})$ is uniformly continuous on this compact region, Assumption D.4 holds.

We give R^{-1} corresponding to choices of F_0 and S considered in this paper. In terms of p-values ($F_0 = \text{Id}$), the average p-value and the minimum p-value correspond to

$$R^{-1}_{\text{avg}}(c_{-1}; r) = 0 \vee \left(Lr - \sum_{i>1} c_i \right) \wedge 1, \quad R^{-1}_{\text{min}}(c_{-1}; r) = \begin{cases} 1, & \min(c_{-1}) \le r \\ r, & \min(c_{-1}) > r \end{cases}.$$

For aggregating Z-scores ($F_0 = \Phi$), we have

$$R^{-1}_{\text{avg}}(c_{-1}; r) = \Phi\left(Lr - \sum_{i>1} \Phi^{-1}(c_i) \right), \quad R^{-1}_{\text{min}}(c_{-1}; r) = \begin{cases} 1, & \min(c_{-1}) \le \Phi(r) \\ \Phi(r), & \min(c_{-1}) > \Phi(r) \end{cases}.$$

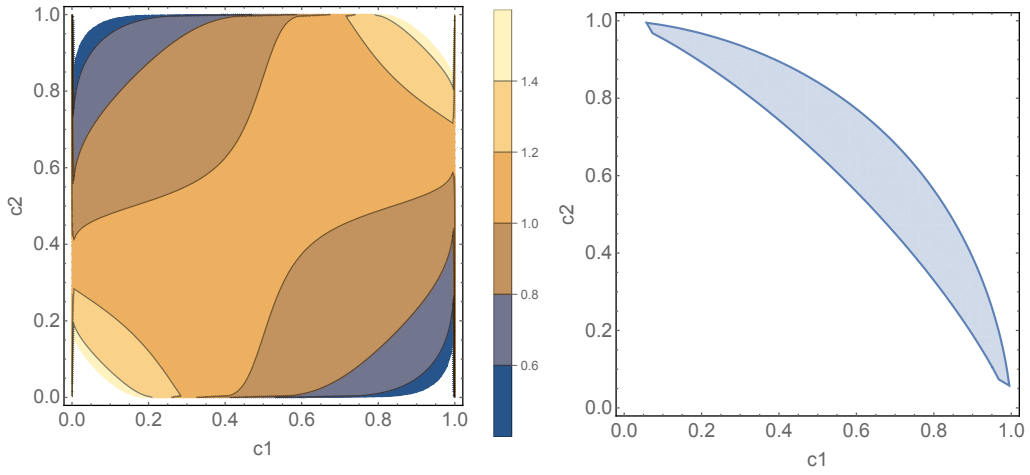

Figure D.1: Left: $f(c_1 \mid c_2)$ under Gaussian copula with $L = 2$ and $\rho = 0.2$, where the function tends to ∞ at $(0, 0)$ and $(1, 1)$. Right: region defined in Assumption D.4, on which $f(c_1 \mid c_2)$ is continuous.

D.5 An illustrative example

In this section, we consider a simple toy example proposed by an anonymous referee to illustrate an application of our theory and the degree to which it matches numerical results. The numerical results also indicate that the conclusion of Theorem 4 should reasonably hold for the original Algorithm 2, although it is proved for a variant of the algorithm that uses two independent copies of the data.

Example D.1. Let $X_1, \ldots, X_n$ be drawn iid from $\mathcal{N}(\mu, 1)$. Suppose that we want to test $H_0 : \mu = 0$ against $H_1 : \mu > 0$. We consider the single-split test statistic

$$T_n^{(1)}(X) := (np)^{-1/2} \sum_{i \in I^{(1)}} X_i,$$

where I is a subset of size $\lfloor np \rfloor$ drawn uniformly at random from $\{1, \ldots, n\}$. Clearly, $T_n^{(1)} \to_d \mathcal{N}(0, 1)$ under H_0.

Let us consider the aggregated statistic $S_n := L^{-1} \sum_l T_n^{(l)}$. Under local alternatives $P_n = P_{\mu_n}$ with $\mu_n = cn^{-1/2}$ and $c > 0$, the aggregated statistic satisfies

$$S_n \to_d \mathcal{N}(c\sqrt{p}, p + (1-p)/L). \tag{81}$$

It follows that the oracle critical value is

$$C_\alpha := G_{P_0}^{-1}(1-\alpha) = z_\alpha \sqrt{p + (1-p)/L},$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$. Meanwhile, we have $G_{m,P_n} \approx \mathcal{N}(c\sqrt{p}\sqrt{m/n}, p + (1-p)/L)$ and hence

$$G_{m,P_n}^{-1}(1-\alpha) \approx c\sqrt{p}\sqrt{m/n} + G_{P_0}^{-1}(1-\alpha).$$

Roughly speaking, by taking M to be a large constant, Theorems 4 and D.2 imply

$$\mathbb{E}\sqrt{n/m}\,(\widehat{C}_\alpha(\text{rank}) - C_\alpha) = 0,$$
$$\mathbb{E}\sqrt{n/m}\,(\widehat{C}_\alpha(\text{ord. sub.}) - C_\alpha) = c\sqrt{p}.$$

These results agree with our simulations plotted in the top and middle panels of Fig. D.2: the top panel is from the two-sample variant of Algorithm 2 and the middle panel is from the original Algorithm 2. A similar result also holds for other choices of S; see the bottom panel of the same figure.

Moreover, we can compute the power functions as follows. Let G_{n,P_n} be the distribution function of S_n. Given that the convergence rate in Eq. (81) should be faster than $\sqrt{m/n}$ (Berry–Esseen), the power of the oracle test is

$$\begin{aligned}\mathrm{pow}(\text{oracle}) = 1 - G_{n,P_n}(G_P^{-1}(1-\alpha)) &= 1 - \Phi\left(z_\alpha - c\sqrt{\frac{p}{p+(1-p)/L}}\right) + o(\sqrt{m/n}) \\ \text{(when } L \text{ is large)} \quad &\approx \boxed{1 - \Phi(z_\alpha - c) + o(\sqrt{m/n}).}\end{aligned}$$

Using Theorem 4 (and taking M to be a large constant), the power of rank-transformed subsampling is

$$\begin{aligned}\mathrm{pow}(\text{rank}) &= 1 - \mathbb{E}\, G_{n,P_n}(\widetilde{G}_n^{-1}(1-\alpha)) \\ &= \mathrm{pow}(\text{oracle}) + o(\sqrt{m/n}) \\ &= 1 - \Phi\left(z_\alpha - c\sqrt{\frac{p}{p+(1-p)/L}}\right) + o(\sqrt{m/n}) \\ \text{(when } L \text{ is large)} \quad &\approx \boxed{1 - \Phi(z_\alpha - c) + o(\sqrt{m/n}).}\end{aligned}$$

By Theorem D.2, where we use $\tau = \lim \sqrt{n/m}(G_{m,P_n}^{-1}(1-\alpha) - G_{P_0}^{-1}(1-\alpha)) = c\sqrt{p}$ and $G_{\mathrm{alt}} = \mathcal{N}(c\sqrt{p}, p + (1-p)/L)$, the power of ordinary subsampling is

$$\begin{aligned}\mathrm{pow}(\text{ord. sub.}) &= 1 - \mathbb{E}\, G_{n,P_n}(\widehat{G}_n^{-1}(1-\alpha)) \\ &= 1 - G_{n,P_n}(G_P^{-1}(1-\alpha)) - \sqrt{m/n}\,\tau\, G'_{\mathrm{alt}}(G_P^{-1}(1-\alpha)) + o(\sqrt{m/n}) \\ &= 1 - \Phi\left(z_\alpha - c\sqrt{\frac{p}{p+(1-p)/L}}\right) \\ &\quad - \sqrt{m/n}\,\phi\left(z_\alpha - c\sqrt{\frac{p}{p+(1-p)/L}}\right) c\sqrt{\frac{p}{p+(1-p)/L}} + o(\sqrt{m/n}) \\ \text{(when } L \text{ is large)} \quad &\approx \boxed{1 - \Phi(z_\alpha - c) - \sqrt{m/n}\,\phi(z_\alpha - c)\, c + o(\sqrt{m/n}).}\end{aligned}$$

Note that $\phi(z_\alpha - c)\, c$ is increasing in c for $c \in (0, z_\alpha)$.

D.6 Auxiliary results

Lemma D.2. *Let $f_n(x)$ be a sequence of real-valued functions on a compact, Euclidean set K. Suppose $f_n \to f$ pointwise on K. Suppose that for a dense subset $K' \subseteq K$,*

$$|f_n(x_n) - f_n(x)| \to 0, \quad \text{for every } x \in K' \text{ and every sequence } (x_n)_{n\in\mathbb{N}} \subseteq K \text{ with } x_n \to x. \tag{82}$$

Then, $f_n \to f$ uniformly on K.

Proof. Without loss of generality, by subtracting f from f_n we can assume $f = 0$. We prove uniform convergence by showing that for any $\varepsilon > 0$, there exists $N'(\varepsilon) > 0$ such that $|f_n(x)| < \varepsilon$ for every $x \in K$ and every $n > N'(\varepsilon)$. First, as $f_n \to 0$ pointwise on K, for every $x \in K$, there

exists $M(x,\varepsilon) > 0$ such that $|f_n(x)| < \varepsilon$ for every $n > M(x,\varepsilon)$. Second, observe that condition Eq. (82) is equivalent to

$$\forall x \in K', \varepsilon > 0, \quad \exists\, \delta(x,\varepsilon) > 0, N(x,\varepsilon) > 0 \ \text{ s.t.,} \\ |f_n(x') - f_n(x)| < \varepsilon, \quad \forall\, x' \in K,\, n : \|x' - x\| < \delta(x,\varepsilon),\ n > N(x,\varepsilon).$$

Consider $\cup_{x\in K'} B_{\delta(x,\varepsilon/2)}(x)$, which is an open cover of K as K' is dense in K. Because K is compact, it must admit a finite subcover of K, which we denote as $\cup_{i\in I} B_{\delta(x_i,\varepsilon/2)}(x_i)$ with every $x_i \in K$ and $|I| < \infty$. Hence, for every $x \in K$, there exists $\widehat{x} \in \{x_i : i \in I\}$ such that $x \in B_{\delta(\widehat{x},\varepsilon/2)}(\widehat{x})$. Then, for every $n > \max_{i\in I} N(x_i,\varepsilon/2) \vee \max_{i\in I} M(x_i,\varepsilon/2)$, we have

$$|f_n(x)| \le |f_n(x) - f_n(\widehat{x})| + |f_n(\widehat{x})| \le \varepsilon/2 + \varepsilon/2 = \varepsilon, \quad \forall x \in K,$$

which completes the proof. □

Proposition D.2. *The set of discontinuity points of a coordinatewise non-decreasing function $g : \prod_{k=1}^d [a_k, b_k] \to \mathbb{R}$ has Lebesgue measure zero, where $a_k < b_k$ for $k = 1, \ldots, d$.*

Proof. It suffices to show that for all $\varepsilon > 0$,

$$A_\varepsilon := \bigcap_{\delta>0} \left\{ x : \sup_{y\in B_\delta(x)} |g(y) - g(x)| \ge \varepsilon \right\}$$

is a Lebesgue null set, where $B_\delta(x)$ denotes the ℓ_2-norm closed ball of radius δ centred at x. Indeed, if x is a discontinuity point, then there exists some $n \in \mathbb{N}$ such that $x \in A_{1/k}$ for all integer $k \ge n$.

Let $\varepsilon > 0$ be given. Now for $j \in \mathbb{Z}$, let $D_j := g^{-1}((-\infty, \varepsilon j])$ and denote by ∂D_j its boundary. We claim that ∂D_j is null. To see this, take $x \in \partial D_j$ and note that as f is coordinatewise non-decreasing, $C := \prod_{k=1}^d [a_k, x_k) \cup \prod_{k=1}^d (x_k, b_k]$ is disjoint from ∂D_j. But then, for all $r > 0$, there exists y such that $B_{cr}(y) \subset B_r(x) \cap (\partial D_j)^c$, so for all $r > 0$,

$$\frac{\mu(\partial D_j \cap B_r(x))}{\mu(B_r(x))} = 1 - \frac{\mu((\partial D_j)^c \cap B_r(x))}{\mu(B_r(x))} \le 1 - \frac{\mu(B_{cr}(y))}{\mu(B_r(x))} = 1 - c^d.$$

Thus each $x \in \partial D_j$ has Lebesgue density (if it exists) strictly less than 1. However, by the Lebesgue density theorem (Mattila, 1995, Cor. 2.14), almost all $x \in \partial D_j$ have Lebesgue density

$$\lim_{r \searrow 0} \frac{\mu(\partial D_j \cap B_r(x))}{\mu(B_r(x))} = 1,$$

so ∂D_j must be a null set. Thus $D := \cup_j \partial D_j$ is also a null set.

We now claim that $D \supseteq A_\varepsilon$, which will give the result. Take $x \in D^c$, so $x \in \mathrm{int}(D_j)$, for some j; we take the minimal such j. Then for some $\delta > 0$, $B_\delta(x) \subset D_j \setminus \overline{D}_{j-1}$, where the overline denotes the closure. Thus for any $y \in B_\delta(x)$, we have

$$\varepsilon(j-1) < g(x),\, g(y) \le \varepsilon j,$$

so $|g(y) - g(x)| < \varepsilon$ and hence $x \notin A_\varepsilon$, which completes the proof. □

Lemma D.3. *Under the assumptions of Proposition D.1, there exists a dense subset K' of $[a,b]$, such that for any $\varepsilon > 0$, $r \in K'$ and $[a,b] \ni r_t \to r$ as $t \to 0$, it holds that*

$$\mathbb{P}\left(|R_l^{-1}(X_{-l}^t; r_t) - R_l^{-1}(X_{-l}^t; r)| > \varepsilon\right) \to 0.$$

Proof. By monotonicity of R, $R_l^{-1}(x_{-l}; r)$ is coordinatewise non-decreasing in $(-x_{-l}, r)$. From Proposition D.2, it follows that $(x_{-l}, r) \mapsto R_l^{-1}(x_{-l}; r)$ is continuous at almost every $(x_{-l}, r) \in [0,1]^{L-1} \times [a,b]$. Stated equivalently, for every $r \in K'$, where K' equals $[a,b]$ excluding a null set, and almost every $x_{-l} \in [0,1]^{L-1}$ (where the null set excluded can depend on r), R_l^{-1} is continuous at (x_{-l}, r). Clearly, K' is dense in $[a,b]$.

Fix $\varepsilon > 0$, $r \in K'$ and $[a,b] \ni r_t \to r$. By the reverse Fatou's lemma, we have

$$\begin{aligned}\limsup \mathbb{P}\left(|R_l^{-1}(X_{-l}^t; r_t) - R_l^{-1}(X_{-l}^t; r)| > \varepsilon\right) &\le \mathbb{P}\left(\{|R_l^{-1}(X_{-l}^t; r_t) - R_l^{-1}(X_{-l}^t; r)| > \varepsilon\} \text{ i.o.}\right) \\ &= \mathbb{P}\left(\limsup |R_l^{-1}(X_{-l}^t; r_t) - R_l^{-1}(X_{-l}^t; r)| > \varepsilon\right).\end{aligned}$$

By the fact that $\|x_{-l}^t - x_{-l}\|_\infty \le t\|h_t\|_\infty \to 0$ and the continuity of R_l^{-1} argued above, for Lebesgue a.e. x_{-l}, we have $R_l^{-1}(x_{-l}^t; r_t) \to R_l^{-1}(x_{-l}; r)$, $R_l^{-1}(x_{-l}^t; r) \to R_l^{-1}(x_{-l}; r)$ and hence

$$\limsup |R_l^{-1}(x_{-l}^t; r_t) - R_l^{-1}(x_{-l}^t; r)| = \lim |R_l^{-1}(x_{-l}^t; r_t) - R_l^{-1}(x_{-l}^t; r)| = 0.$$

Finally, because X_{-l} has a distribution absolutely continuous w.r.t. the Lebesgue measure, the result follows from $\mathbb{P}\left(\limsup |R_l^{-1}(X_{-l}^t; r_t) - R_l^{-1}(X_{-l}^t; r)| > \varepsilon\right) = 0$. □

Lemma D.4. *Let X_n be a sequence of real-valued random variables. If for every subsequence n_k, there is a further subsequence $n_{k(j)}$ such that $X_{n_{k(j)}}$ converges to a symmetric law (the law may depend on the subsequence), then we have*

$$\mathbb{E}[-M \vee X_n \wedge M] \to 0, \quad \text{for every } M > 0.$$

Proof. Fix any $M > 0$. By our assumption, for every subsequence n_k, there exists a further subsequence $n_{k(j)}$ such that $-M \vee X_{n_{k(j)}} \wedge M$ converges to a symmetric law. Because it is bounded, we have $\mathbb{E}[-M \vee X_{n_{k(j)}} \wedge M] \to 0$. The result follows because the limit does not depend on the subsequence. □

Lemma D.5. *Let $\{V_{i,k} : i = 1, \dots, n,\ k = 1, \dots, K,\}$ be random variables such that*

$$V_{1,k}, \dots, V_{n,k} \overset{iid}{\sim} F_n, \quad k = 1, \dots, K,$$

where F_n is a distribution function satisfying

$$\sqrt{n}\|F_n - \mathrm{Id}\,\|_\infty \to 0.$$

Let $\mathbb{V}_n$ be the empirical distribution function of $\{V_{i,k} : i = 1, \dots, n,\ k = 1, \dots, K\}$. Then, the empirical process $\sqrt{n}(\mathbb{V}_n - \mathrm{Id})$, as a random bounded function on $[0,1]$, is asymptotically tight and asymptotically measurable. Further, there exists a subsequence n_j along which

$$\sqrt{n_j}(\mathbb{V}_{n_j} - \mathrm{Id}) \rightsquigarrow \xi \quad \text{in } D[0,1],$$

where $\xi =_d -\xi$ and ξ is almost surely uniformly continuous (the distribution of ξ can depend on the subsequence) with $\xi(0) = \xi(1) = 0$.

Proof. Let $\mathbb{V}_{n,k}$ be the empirical distribution function of $\{V_{i,k} : i = 1, \dots, n\}$. For each k, it holds that $(V_{1,k}, \dots, V_{n,k}) =_d (F_n^{-1}(U_1), \dots, F_n^{-1}(U_n))$ for iid $U_1, \dots, U_n \sim \mathrm{unif}(0,1)$. Therefore, we have

$$\mathbb{V}_{n,k}(x) =_d n^{-1} \sum_i \mathbb{I}\{F_n^{-1}(U_i) \le x\} = n^{-1} \sum_i \mathbb{I}\{U_i \le F_n(x)\} = \mathbb{U}_n \circ F_n(x),$$

where $\mathbb{U}_n$ is the empirical distribution function of $\{U_i\}$. It follows that

$$\sqrt{n}(\mathbb{V}_{n,k} - \mathrm{Id}) =_d \sqrt{n}(\mathbb{U}_n - \mathrm{Id}) + \sqrt{n}(F_n - \mathrm{Id}) + \sqrt{n}(\mathbb{U}_n - \mathrm{Id})(F_n - \mathrm{Id}).$$

The empirical process $\sqrt{n}(\mathbb{U}_n - \mathrm{Id})$ converges weakly to a standard Brownian bridge. By our assumption, $\sqrt{n}\|F_n - \mathrm{Id}\,\|_\infty \to 0$. By continuous mapping theorem, $\|\sqrt{n}(\mathbb{U}_n - \mathrm{Id})(F_n - \mathrm{Id})\|_\infty \to_p 0$. Therefore, we conclude that $\sqrt{n}(\mathbb{V}_{n,k} - \mathrm{Id})$ converges weakly to a standard Brownian bridge. It then follows that the sequence $\sqrt{n}(\mathbb{V}_{n,k} - \mathrm{Id})$ is asymptotically tight and asymptotically measurable. Therefore,

$$\sqrt{n}(\mathbb{V}_{n,1} - \mathrm{Id}, \ldots, \mathbb{V}_{n,K} - \mathrm{Id}),$$

as a random bounded $[0,1] \to \mathbb{R}^K$ function, is also asymptotically tight and asymptotically measurable. By Prohorov's theorem (van der Vaart, 2000, 18.12), there exists a subsequence n_j such that

$$\sqrt{n_j}(\mathbb{V}_{n_j,1} - \mathrm{Id}, \ldots, \mathbb{V}_{n_j,K} - \mathrm{Id}) \rightsquigarrow (\xi_1, \ldots, \xi_K),$$

where every ξ_k is a standard Brownian bridge. By the continuous mapping theorem, their average

$$\sqrt{n_j}(\mathbb{V}_{n_j} - \mathrm{Id}) \rightsquigarrow (\xi_1 + \cdots + \xi_K)/K =: \xi.$$

Because $\xi_k =_d -\xi_k$ for each k, we know $\xi =_d -\xi$. Because each ξ_k is uniformly continuous and satisfies $\xi_k(0) = \xi_k(1) = 0$ almost surely (van der Vaart, 2000, §19.1), ξ also almost surely satisfies these properties. □

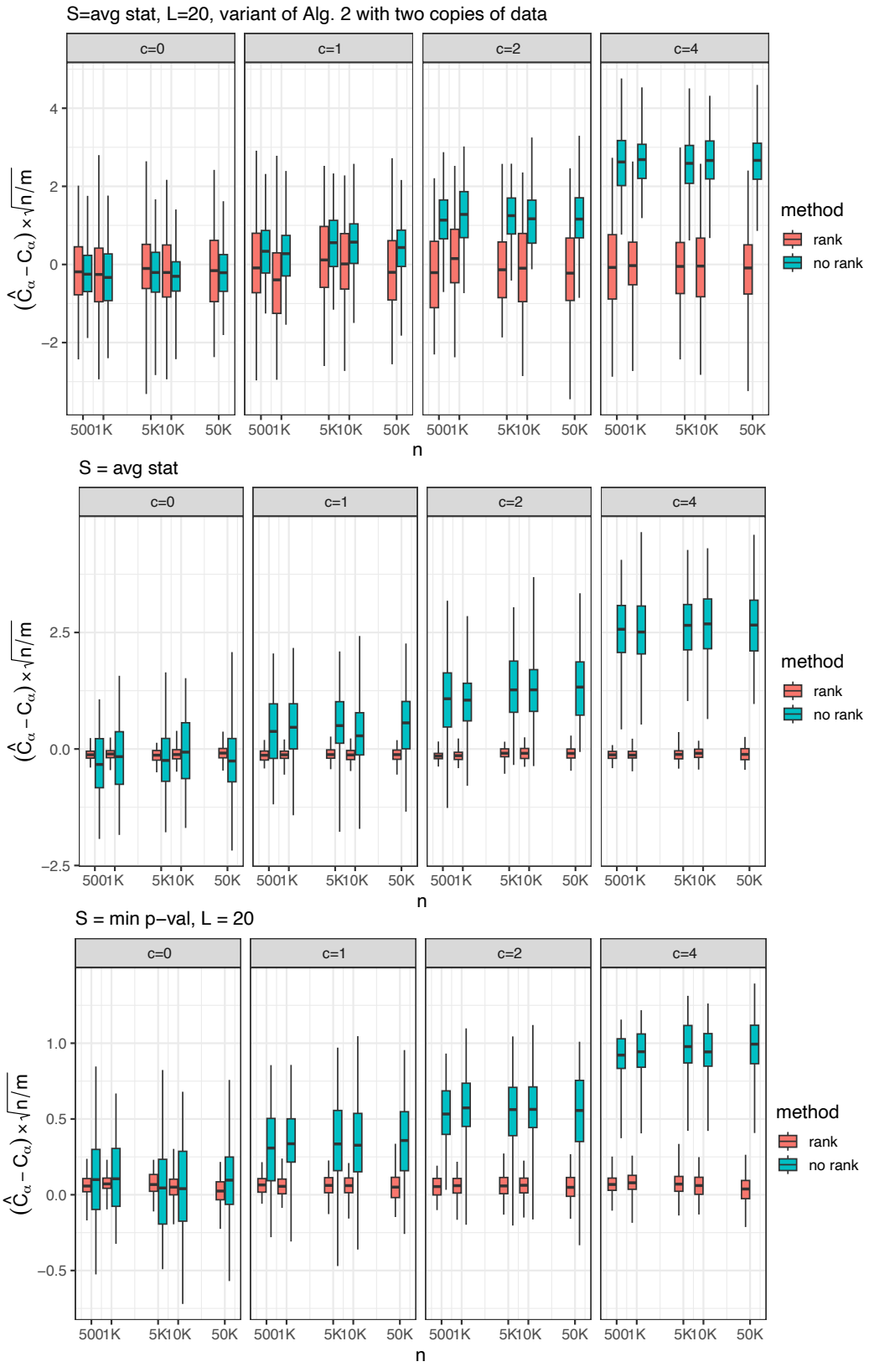

Figure D.2: Numerical results for Example D.1: first-order bias in approximating the oracle critical value under $\mu = cn^{-1/2}$ ('`rank`': rank-transformed subsampling, '`no rank`': ordinary subsampling). The thick middle line in the box plots represents the mean value of the first-order bias $\sqrt{n/m}(\widehat{C}_\alpha - C_\alpha)$. Top: the variant of Algorithm 2 that uses two independent copies of X with $S = \text{avg}$; Middle: Algorithm 2 with $S = \text{avg}$; Bottom: Algorithm 2 with S being (one minus) the minimum p-value.

E Calibration of cross-fitted DML

We first provide details for the numerical example in Section 4.3. In the partially linear model, we let $X \sim \mathcal{N}(0,1)$ and $s(X) := 4\sqrt{|X|}$. The heteroscedastic errors are drawn as

$$V = s(X)\,\varepsilon_V, \quad \xi = s(X)\,\varepsilon_\xi$$

with $\varepsilon_V \sim \mathsf{Ga}(0.5,1) - 0.5$ and $\varepsilon_V \sim \mathsf{Ga}(0.3,1) - 0.3$. The regression functions are chosen as

$$\begin{aligned} m_0(x) &= x + \cos(x) + \exp(x)/(1+\exp(x)), \\ g_0(x) &= \left[-10x + 3\cos(4x)x^2/(1+\exp(x/6))\right]/10. \end{aligned}$$

Below is an example of performing DML when one of the two (parametric) nuisance models is misspecified. Because of the double robustness of the debiased score, the resulting cross-fitted DML estimator is still consistent and in fact asymptotically normal. However, unlike the scenario discussed in Section 4.3 where the per-fold estimators $\widehat{\theta}^{(1)}, \ldots, \widehat{\theta}^{(L)}$ are asymptotically uncorrelated (so correlation is only a finite-sample phenomenon), because of the misspecification, $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ in the example below are negatively correlated even in large sample. Hence, the standard plugin estimator for the asymptotic variance (see Eq. (19)) is inconsistent. Nevertheless, the estimator based on rank-transformed subsampling is still consistent.

Example E.1 (DML under misspecification)**.** Consider the following specification of the partially linear model considered in Section 4.3:

$$\begin{aligned} D &= \beta_1 X + \beta_2 X^2 + V, \quad \mathbb{E}[V \mid X] = 0, \\ Y &= \theta_0 D + \gamma_1 X + \gamma_2 X^2 + \xi, \quad \mathbb{E}[\xi \mid D, Y] = 0, \end{aligned}$$

where θ_0 is the parameter of interest. For simplicity, we assume $X, \varepsilon_X, \varepsilon_Y$ are drawn independently from $\mathcal{N}(0,1)$. The DML estimator of θ_0, as the solution to the empirical Robinson's score function, is the least squares coefficient of regressing the residual $r_Y := Y - \widehat{E}[Y \mid X]$ on the residual $r_D := D - \widehat{E}[D \mid X]$, where the estimation of regression functions and the final least squares are performed on two separate parts of the sample.

Suppose that $\widehat{E}[D \mid X]$ is well-specified but $\widehat{E}[Y \mid X]$ is misspecified as $\eta_1 X$ (missing X^2), whereas the true model is $\eta_1 X + \eta_2 X^2$ with $\eta_1 = \theta_0\beta_1 + \gamma_1$ and $\eta_2 = \theta_0\beta_2 + \gamma_2$. Consider the DML estimator $\widehat{\theta}_{\mathrm{DML}} = (\widehat{\theta}^{(1)} + \widehat{\theta}^{(2)})/2$ for the case of $L = 2$ folds with

$$\widehat{\theta}^{(l)} = \frac{\sum_{i \in I^{(l)}} r_{Y_i} r_{D_i}}{\sum_{i \in I^{(-l)}} r_{D_i}^2}, \quad l = 1, 2,$$

where $(I^{(1)}, I^{(2)})$ is a random split of the sample into two parts of equal size. By an asymptotic linear expansion of both numerator and denominator and applying the delta method, it can be shown that $\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}$ admit the following asymptotic expansion:

$$\begin{aligned} \sqrt{n/2}\,(\widehat{\theta}^{(1)} - \theta_0) &= \frac{1}{\sqrt{n/2}} \sum_{i \in I^{(1)}} \left(\eta_2 X_i^2 V_i + V_i \xi_i\right) - \frac{1}{\sqrt{n/2}} \sum_{i \in I^{(2)}} \eta_2 X_i^2 V_i + o_p(1), \\ \sqrt{n/2}\,(\widehat{\theta}^{(2)} - \theta_0) &= \frac{1}{\sqrt{n/2}} \sum_{i \in I^{(2)}} \left(\eta_2 X_i^2 V_i + V_i \xi_i\right) - \frac{1}{\sqrt{n/2}} \sum_{i \in I^{(1)}} \eta_2 X_i^2 V_i + o_p(1). \end{aligned}$$

Applying CLT to the RHS, we derive

$$\sqrt{n/2} \begin{pmatrix} \widehat{\theta}^{(1)} - \theta_0 \\ \widehat{\theta}^{(2)} - \theta_0 \end{pmatrix} \to_d \mathcal{N}\left(0, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right),$$

where

$$\sigma^2 = 6\eta_2^2 + 1, \quad \rho = -6\eta_2^2/(1 + 6\eta_2^2). \tag{83}$$

Note that $-1 < \rho < 0$ whenever $\eta_2 \neq 0$, i.e., when $\widehat{\mathbb{E}}[Y \mid X]$ is misspecified. It then follows that

$$\sqrt{n}\,(\widehat{\theta}_{\mathrm{DML}} - \theta_0) \to_d \mathcal{N}\left(0, (1+\rho)\sigma^2\right).$$

In the below, we take $(\beta_1, \beta_2) = (1, 1/2)$, $(\gamma_1, \gamma_2) = (1, 1)$, $\theta_0 = 1$ and let $V, \xi \sim \mathcal{N}(0, 1)$. Asymptotic values of correlation and variance, along with values under finite sample, are compared to their estimates in Table E.1. As can be seen from the table, the rank-transformed subsampling estimates are consistent while the standard DML plugin estimate is too large. See also Table E.2 for the coverage of confidence intervals under $L = 2$ and $L = 5$.

Table E.1: Correlation, standard deviation, their rank-transformed subsampling estimates, as well as the standard plugin estimate (via Eq. (19)) under model misspecification ($L = 2$). The $n = \infty$ row refers to asymptotic values computed from Eq. (83).

n	ρ	$\widetilde{\rho}$	$\mathrm{sd}\,\sqrt{n}(\widehat{\theta}_{\mathrm{DML}} - \theta_0)$	$\widehat{\sigma}_{\mathrm{ls}}\sqrt{1+\widetilde{\rho}}$	$\widehat{\mathrm{sd}}_{\mathrm{plugin}}\sqrt{n}(\widehat{\theta}_{\mathrm{DML}} - \theta_0)$
500	-0.82	-0.53	1.63	2.85	3.14
1000	-0.86	-0.66	1.39	2.34	3.14
5000	-0.92	-0.85	1.08	1.51	3.12
10000	-0.92	-0.89	1.06	1.30	3.12
50000	-0.93	-0.92	1.05	1.05	3.12
∞	-0.93		1		

Table E.2: Coverage of nominal 95% confidence intervals under model misspecification (brackets: median width of intervals multiplied by $\sqrt{n}$)

	$L = 2$		$L = 5$	
n	Eq. (21)	DML	Eq. (21)	DML
500	0.994 [11.6]	0.999 [12.3]	0.988 [6.1]	1.000 [12.2]
1000	0.999 [9.5]	1.000 [12.3]	0.975 [5.3]	1.000 [12.2]
5000	0.996 [6.0]	1.000 [12.2]	0.968 [4.2]	1.000 [12.2]
10000	0.990 [5.1]	1.000 [12.2]	0.956 [3.9]	1.000 [12.2]
50000	0.961 [4.1]	1.000 [12.2]	0.941 [3.8]	1.000 [12.2]

F On the bootstrap

A curious reader may wonder if the rank transform can be applied to the bootstrap instead of subsampling to achieve the same goal. The short answer is no, if we want to maintain type-I error control under minimal assumptions. In contrast to subsampling, which only requires minimal assumptions (the existence of a non-degenerate limit law) to be consistent, the standard n-out-of-n bootstrap also requires the regularity of a "root", namely a function of both the test statistic and parameters of the underlying distribution such that its distribution is locally pivotal; see Beran (1997) and Politis et al. (1999, §1.6). This is considerably stronger than our Condition 1 and Assumption 1 required for pointwise level control, which only concern the behaviour of

the test statistic (instead of a root) under each fixed null (instead of neighbourhoods of each null). In fact, for multiple-split, hunt-and-test procedures considered in this paper, it is often unclear whether such a regular root (as an L-dimensional vector) exists, and if so, how to construct one. Consider the test for the null of multivariate unimodality in Section 4.1.2, for example. The existence of such a regular root would mean that were data to come from a (slightly) non-unimodal distribution, a certain transformation could be applied to data to restore its unimodality — such a transformation seems rather hard to construct. Thus, we use subsampling to avoid these limitations; replacing it with the bootstrap can fail to control the type-I error unless $(T_n^{(1)}, \ldots, T_n^{(L)})$ itself is a regular root.

G Relation to prepivoting

In resampling inference, ranks also come up when describing prepivoting (Beran, 1987, 1988), which refers to the technique of applying the probability integral transform to a statistic using its bootstrap null distribution, i.e., turning a statistic into its bootstrap p-value. Prepivoting is most useful when the asymptotic null distribution of the statistic contains unknown parameters. To achieve improvement for such cases, prepivoting must be iterated twice or more times with the nested bootstrap, before comparing the final transformed pivot to unif$(0, 1)$. Prepivoting reduces the dependency of the sampling distribution of the statistic on the underlying data distribution and can offer higher-order refinements to the bootstrap. We argue that our use of ranks is rather different: (1) while prepivoting improves level control, our rank transform is designed to improve power; (2) the rank transform is applied to the subsample statistics (instead of the test statistic) and applied only once; (3) we use ranks to enforce the marginals of a multivariate statistic but prepivoting is only applicable to a univariate statistic.

H Additional numerical results

H.1 Derandomization of Example 2

Fig. H.1 shows probability of non-replication (i.e. two applications of the same test on the same data leads to one acceptance and one rejection) for the numerical experiment in Section 4.1.1.

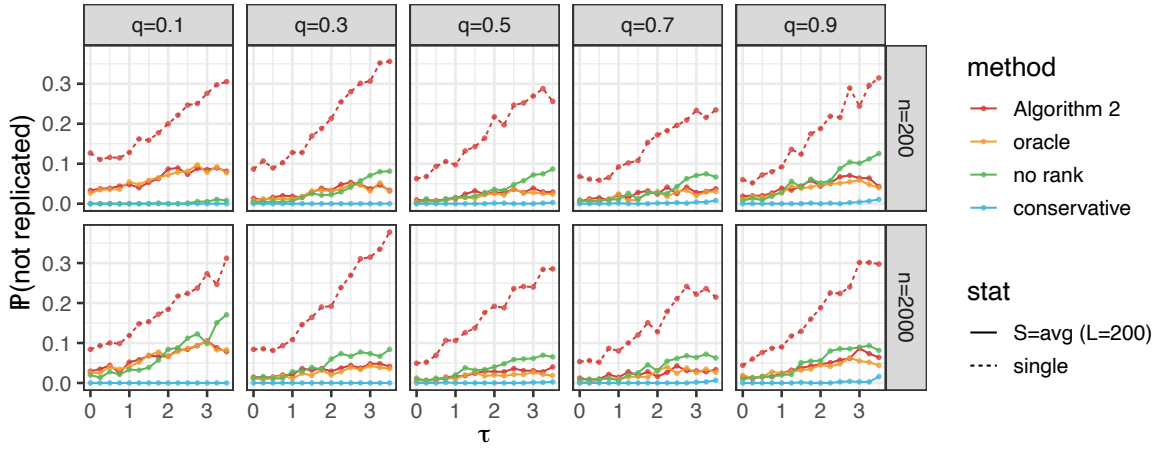

Figure H.1: Probability of non-replication for testing $\mu = \mathbf{0}$ in Example 2 at level $\alpha = 0.05$, where $\mu = \tau n^{-1/2} v_1$. Tests based on $S_n = (T_n^{(1)} + \cdots + T_n^{(L)})/L$ (solid line) significantly reduces the chance of non-replication from using the single-split test.

H.2 Goodness-of-fit testing for quantile regression

In this section, we include additional numerical results for goodness-of-fit testing of quantile regression considered in Section 4.1.3. Fig. H.2 reports the results when the non-linear function in Eq. (13) is chosen as $\eta(X) = 2\exp(-1 - X_2 - X_3)$.

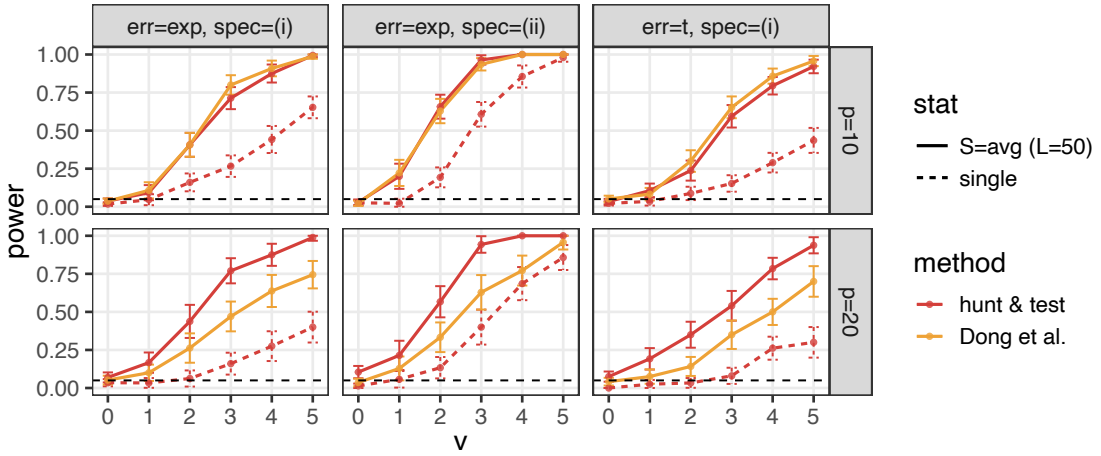

Figure H.2: Testing goodness-of-fit of a quantile regression model $q_{0.5}(X) = \beta_0 + \beta^\top X$: power (95% CI) at level $\alpha = 0.05$ (dashed horizontal) under $n = 1000$. The model is well-specified if and only if $v = 0$. The non-linear function in Eq. (13) is $\eta(X) = 2\exp(-1 - X_2 - X_3)$.

H.3 Testing generalised conditional independence in a more challenging setting

Consider changing the specification of Y in Section 4.2 to

$$Y = -A_2 + \beta_{H,Y}^\top H + (1 - A_1)\varepsilon_Y + A_1\xi_\nu, \quad \nu \in (0, \infty], \tag{84}$$

where ξ_ν is an independent t_ν-distributed random variable. The null hypothesis holds when $\nu = \infty$ and hence $\xi_\nu =_d \varepsilon_Y$. Because $\mathrm{cov}_Q(A_1, Y) = 0$ under the alternative, using it as the statistic leads to trivial power. Instead, we employ the maximum mean discrepancy (MMD) statistic (Gretton et al., 2012) between $Y \mid A_1 = 0$ and $Y \mid A_1 = 1$ under Q, with the Gaussian kernel and bandwidth chosen by the median heuristic (Gretton et al., 2012, §8), which is able to detect distributional discrepancy beyond the mean. Under the null hypothesis, the asymptotic distribution of the MMD statistic depends on unknown parameters (Gretton et al., 2012, Theorem 12) that are very difficult to handle under IPW. Nevertheless, the aggregated post-resampling permutation tests just works out of the box; see Fig. H.3.

H.4 Comparison with conservative p-value merging

In this section, we include numerical results that compare our method to various conservative, deterministic merging methods for non-independent p-values introduced by Vovk and Wang (2020). In what follows, `single` refers to the p-value resulting from a single application of a randomized test (e.g., one single data split). Labels `avg` and `min` respectively refer to subsampling based inference for the average p-value and the minimum p-value. Conservatively merged p-values are labelled as `M.xx` for the following merging functions $p = f(p_1, \ldots, p_L)$:

Arithmetic (`M.arith`) $p = 2\,(p_1 + \cdots + p_L)/L$,

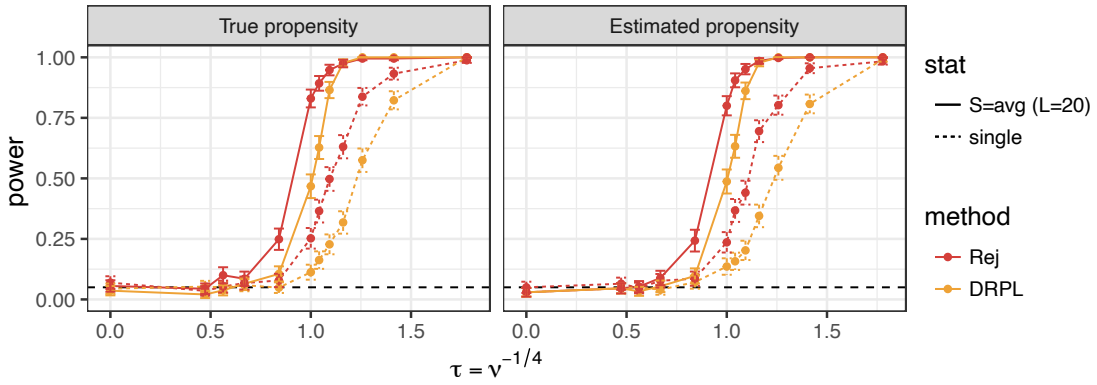

Figure H.3: Power (95% CI) for testing no individual direct effect of A_1 on Y ($\mathcal{G}$ of Fig. 9) under the more difficult setting in Appendix H.3, for which tests based on $|\operatorname{cov}_Q(A_1, Y)|$ cannot detect the effect. The null hypothesis corresponds to $\tau = 0$. Here 'single' refers to post-rejection-sampling (`Rej`) and post-DRPL (`DRPL`) permutation test of the MMD statistic; 'avg' refers to the corresponding aggregated test based on the average of 20 p-values, calibrated by Algorithm 2. Sample size is $n = 1000$ and level is 0.05 (dashed horizontal). See also Fig. H.7 in the appendix.

Geometric (`M.geom`) $p = e\,(p_1 \dots p_L)^{1/L}$,

Bonferroni (`M.Bonf`) $p = L \min(p_1, \dots, p_L)$,

Compound Bonferroni-geometric (`M.Bonf-geom`) $p = 2\min\{L \min(p_1, \dots, p_L), e\,(p_1 \dots p_L)^{1/L}\}$.

These merging functions are precise or asymptotically precise, which roughly mean that the multiplicative constants in their definitions cannot be improved in general; see Vovk and Wang (2020).

In the below, Fig. H.4 shows results for detecting multivariate unimodality considered in Section 4.1.2; Fig. H.5 shows results for goodness-of-fit testing for quantile regression considered in Section 4.1.3; Figs. H.6 and H.7 show results for testing generalised conditional independence in the two settings considered in Section 4.2.

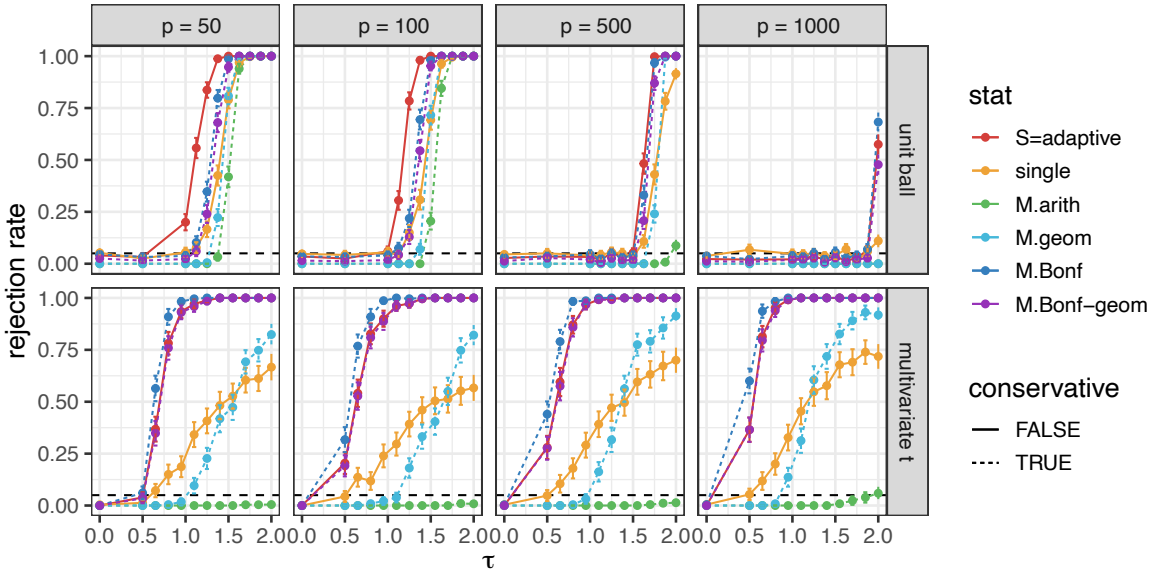

Figure H.4: Supplement to Fig. 6: comparison of single-split p-value (`single`), subsampling p-value (`S=adaptive`) and conservative p-value merging functions (`M.xx`, dashed curves) for testing multivariate unimodality in p dimensions considered in Section 4.1.2. The null hypothesis holds when $\tau = 0$ and the level is 0.05. Subsampling and conservative merging are based on $L = 50$ exchangeable p-values.

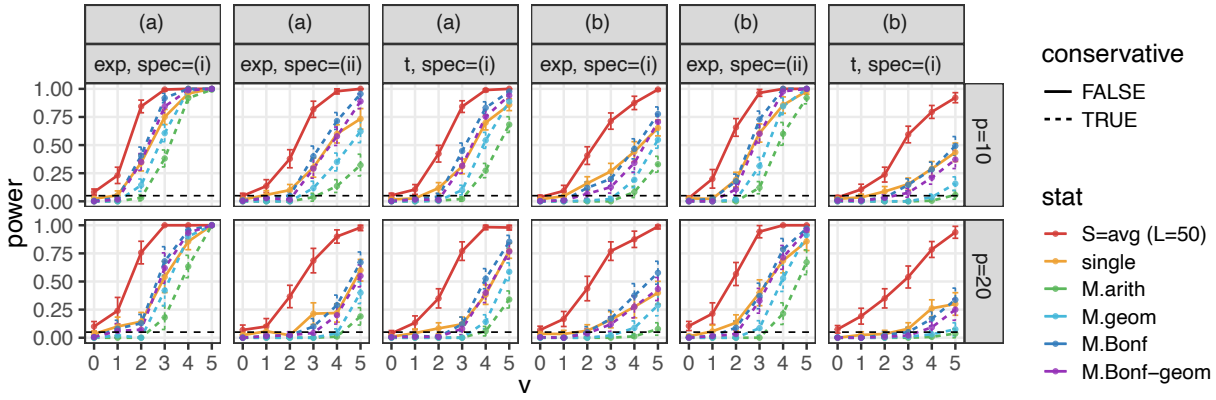

Figure H.5: Supplement to Figs. 8 and H.2: comparison of single-split p-value (`single`), subsampling p-value (`S=avg`) and conservative p-value merging functions (`M.xx`, dashed curves) for goodness-of-fit testing of quantile regression considered in Section 4.1.3. The null hypothesis holds when $v = 0$ and the level is 0.05. Subsampling and conservative merging are based on $L = 50$ exchangeable p-values. The non-linear function in Eq. (13) is: (a) $\eta(X) = 4\sqrt{X_1^2 + X_2^2}$, (b) $\eta(X) = 2\exp(-1 - X_2 - X_3)$.

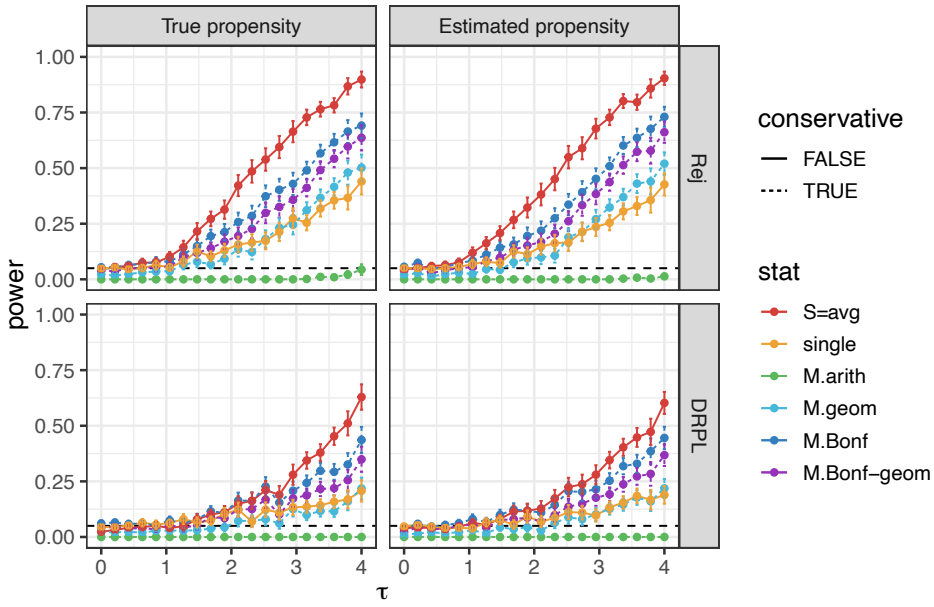

Figure H.6: Supplement to Fig. 10: comparison of single-split p-value (`single`), subsampling averaged p-value (`avg`) and conservative p-value merging functions (`M.xx`, dashed curves) for testing a generalised conditional independence; see Section 4.2 for the setting. The null hypothesis holds when $\tau = 0$ and the level is 0.05. Subsampling and conservative merging are based on $L = 20$ exchangeable p-values.

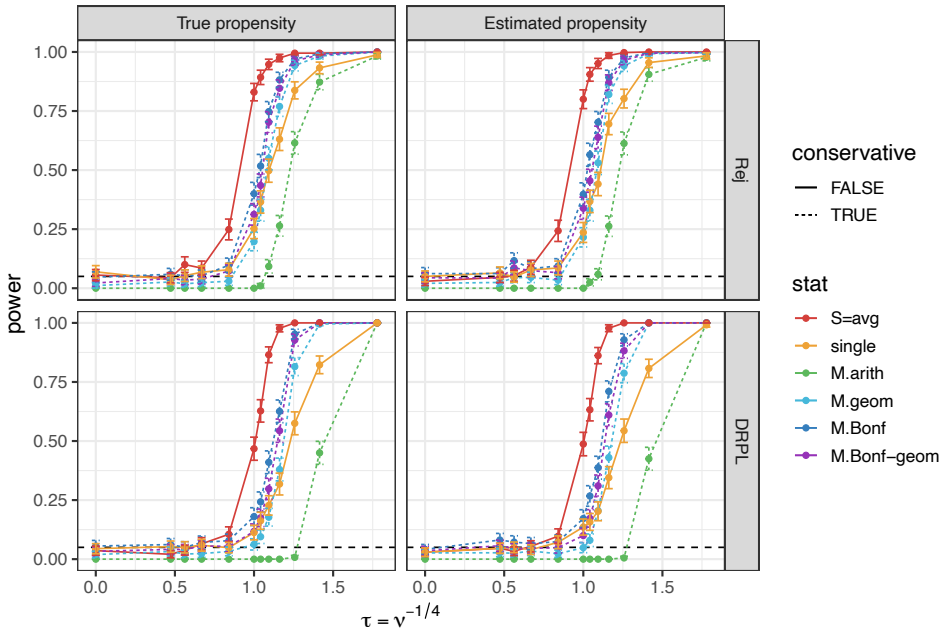

Figure H.7: Supplement to Fig. H.3: comparison of single-split p-value (`single`), subsampling averaged p-value (`avg`) and conservative p-value merging functions (`M.xx`, dashed curves) for testing a generalised conditional independence under the more difficult setting Eq. (84); see Section 4.2. The null hypothesis holds when $\tau = 0$ and the level is 0.05. Subsampling and conservative merging are based on $L = 20$ exchangeable p-values.