Testing Independence Under Biased Sampling

Yaniv Tenzer*, Micha Mandel, and Or Zuk
Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

**ABSTRACT**
Testing for dependence between pairs of random variables is a fundamental problem in statistics. In some applications, data are subject to selection bias that can create spurious dependence. An important example is truncation models, in which observed pairs are restricted to a specific subset of the $X$-$Y$ plane. Standard tests for independence are not suitable in such cases, and alternative tests that take the selection bias into account are required. Here, we generalize the notion of quasi-independence with respect to the sampling mechanism, and study the problem of detecting any deviations from it. We develop two tests statistics motivated by the classic Hoeffding's statistic, and use two approaches to compute their distribution under the null: (i) a bootstrap-based approach, and (ii) a permutation-test with nonuniform probability of permutations. We also handle an important application to the case of censoring with truncation, by estimating the biased sampling mechanism from the data. We prove the validity of the tests, and show, using simulations, that they improve power compared to competing methods for important special cases. The tests are applied to four datasets, two that are subject to truncation, with and without censoring, and two to bias mechanisms related to length bias.

1. Introduction
Testing independence of two random variables $X, Y$ is a fundamental statistical problem. Classical methods have focused on testing linear (Pearson's correlation coefficient) or monotone (Spearman's correlation, Kendall's tau) dependence, while other works focus on developing methods to capture complex dependencies (e.g., using Pearson's Chi-squared test). This classical problem keeps drawing attention from scholars with recent approaches focusing on omnibus tests employing computer-intensive methods (Gretton et al. 2008; Székely et al. 2009; Heller, Heller, and Gorfine 2012; Heller et al. 2016).

A more challenging task is testing independence of two random variables when data is obtained through a general biased sampling mechanism. The most familiar example is that of truncation, where observations are restricted to a certain "observable region." All cross-sectional samples are subject to some form of truncation, and the standard way of analyzing such data is to assume independence between the truncation mechanism and the variables of interest (see Chiu et al. 2018 for more discussion and references). The results of the analysis can be highly biased if the independence assumption is violated. Yet, another important problem that exploits tests for truncated data is testing the Markov assumption in the Illness-Death model (Rodríguez-Girondo and de Uña-Álvarez 2012). The problem of quasi-independence was first dealt with in the framework of contingency tables (e.g., Goodman (1968)) and was studied more recently in the framework of survival analysis where data are restricted by the condition $X \leq Y$ (see Tsai (1990) and the discussion below). Bickel and Ritov (1991) considered general biased regression models for which independence is equivalent to a zero regression coefficient. Another common framework where the problem naturally arises is in cross-sectional sampling designs; Section 7 provides several examples. Biased sampling in general, and truncation in particular, may imply dependence in the sample that does not exist in the population. This fact was acknowledged more than a century ago by Elderton et al. (1913) who studied the correlation between an intellectually disabled child's place in the family and the size of that family. They noticed that "the size of the family must be as great or greater than the imbecile's place in it... and there would certainly be correlation, if we proceeded to find it by the usual product moment method, but such correlation is, or clearly may be, wholly spurious."

In this article, we study the problem of detecting dependency from general biased samples. We are given a sample of $n$ independent and identically distributed (iid) observations $(x_i, y_i)_{i=1}^n$ drawn from a joint distribution having a density $F_{X,Y}(dx, dy) \propto w(x, y)F_{XY}(dx, dy)$, where $w$ is a known nonnegative function having a positive finite expectation with respect to $F_{XY}$. Here, $F_{XY}$ is the joint distribution of the pair $(X, Y)$ in the population, and $F_{X,Y}^w$ is the distribution of observed pairs, tilted by $w$, the sampling mechanism. The aim is to test the null hypothesis $H_0: F_{XY}(x, y) = F_X(x)F_Y(y)$ for all $(x, y)$. However, because $F_{XY}$ is not identifiable on $\{(x, y) : w(x, y) = 0\}$, the goal is restricted to testing quasi-independence defined as $F_{XY}(dx, dy) = F_X(dx)F_Y(dy)$ for all $(x, y)\in \{(x', y') : w(x', y') > 0\}$, for some functions $F_X, F_Y$ (Tsai 1990).
Testing quasi-independence under a biased-sampling regime is challenging and previous works mainly focused on simple truncation models. Tsai (1990) considered the problem of testing quasi-independence under left-truncation (and right censoring) based on the conditional Kendall's tau correlation coefficient. Efron and Petrosian (1999) and Martin and Betensky (2005) extended this method to the settings of double-truncation. Chen and Liu (2007) suggested an importance sampling algorithm to estimate the P-value under truncation models. Emura and Wang (2010) constructed a log-rank type statistic for the left-truncation setting. Chen, Tsai, and Chao (1996) proposed a conditional version of Pearson's product-moment correlation. Efron and Petrosian (1999) and Martin and Betensky (2006) focused on a scan statistic similar to the ones suggested by Heller et al. (2016). The test statistic is quasi-independent and performs much better than the procedure has in simulations and applied to real-life datasets in Sections 6 and 7, respectively. Section 8 completes the article with a discussion.

2. Preliminaries

Let \( F_{XY} \) be a bivariate distribution function with a density \( f_{XY} \) and univariate marginals \( f_X, f_Y \). We consider \( n \) independent pairs \((X_i, Y_i) \sim F_{XY}(x, y)\) of scalar continuous random variables, sampled from the joint density

\[
f_{X,Y}^{(w)}(x, y) = w(x, y)f_{XY}(x, y) / \mathbb{E}_{f_{XY}}\{w(X, Y)\},
\]

where \( w : \mathbb{R}^2 \to \mathbb{R}^+ \) is a nonnegative weight function such that \( 0 < \mathbb{E}_{f_{XY}}\{w(X, Y)\} < \infty \). The marginals of the observed data are denoted by \( f_X^{(w)}(x) = \int_{-\infty}^{\infty} f_{XY}^{(w)}(x, y) dy \) and \( f_Y^{(w)}(y) = \int_{-\infty}^{\infty} f_{XY}^{(w)}(x, y) dx \). The weighted independent density is defined as \( f_{XY}^{(w)}(x, y) = w(x, y) f_X(x) f_Y(y) / \mathbb{E}_{f_{XY}}\{w(X, Y)\} \), with the corresponding weighted distribution \( F_{X,Y}^{(w)} \).

An important special case is that of a truncated sample in which

\[
w(x, y) = 1_{A}(x, y) = \begin{cases} 1, & (x, y) \in A \\ 0, & \text{otherwise,} \end{cases}
\]

for some set \( A \subset \mathbb{R}^2 \). This special form of \( w(x, y) \) arises frequently in practice and was previously investigated by several authors, as discussed in Section 1.

A strongly related concept to our problem is that of quasi-independence in truncation models (Tsai 1990), which can be naturally extended to a general weight function \( w(x, y) \):

**Definition 1.** (quasi-dependence) We say that the joint distribution \( F_{XY}^{(w)} \) is quasi-independent with respect to the weight function \( w \), if there exist density functions \( f_X \) and \( f_Y \), such that

\[
F_{XY}^{(w)}(x, y) \propto \int_{-\infty}^{x} \int_{-\infty}^{y} w(s, t) f_X(s) f_Y(t) ds dt, \quad \forall x, y \in \mathbb{R}.
\]

Otherwise, we say that \( F_{XY}^{(w)} \) is quasi-dependent with respect to the weight function \( w \).

Based on the sample \( D = \{(x_i, y_i)\}^n_{i=1} \), we aim at performing the following hypothesis testing for quasi-dependence:

\[
H_0: \ F_{XY}^{(w)} \ \text{is quasi-independent} \quad H_1: \ \text{otherwise}
\]

**Remark 1.** Quasi-dependence implies dependence. When \( w \) is strictly positive, quasi-dependence is simply dependence of \( X \) and \( Y \).

**Remark 2.** Quasi-independence does not imply dependence. If \( w(x, y) = 0 \) for some \((x, y) \in \mathbb{R}^2 \), it is possible to have quasi-independence without independence (and then we must have either \( \bar{F}_X \neq F_X \) or \( \bar{F}_Y \neq F_Y \)). For the important case of \( w(x, y) = 1(x, y) \), Cheng, Hall, and Yang (2007) discussed the identifiability problem and its implications.
We denote by $D_x, D_y$ the unordered samples comprising of $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$, respectively. For convenience, we often keep the indices of the original data, but not the coupling between $x_i, y_i$ and to this end we use the unordered sample $(x, D_y)$—that is, we keep the original ordering of the $x_i$'s but only the marginal empirical distribution of the $y_i$'s.

3. Permutation Test

3.1. The Distribution of Permutations

Under biased-sampling, different permutations are not equally likely under the null model, thus should not be uniformly sampled as in standard permutation tests. Therefore, the sampling mechanism should account for the discrepancy in weight of distinct permutations based on the data. Let $\pi(y)$ be the vector $y$ rearranged according to a permutation $\pi$, that is, $\pi(y) = (y_{\pi(1)}, \ldots, y_{\pi(n)})$, and let $\pi(D)$ be the permuted sample: $\pi(D) \equiv ((x_1, y_{\pi(1)}), \ldots, (x_n, y_{\pi(n)}))$. For a sample $D = \{(x_i, y_i)\}_{i=1}^n$, let $W \in \mathbb{R}_{n \times n}$ be a weight-matrix defined by $W(i,j) \equiv w(x_i, y_j)$. Let $S_n$ be the set of all permutations of $n$ elements, and for $\pi \in S_n$, consider the probability

$$P_{XY}(\pi) = \frac{1}{\text{per}(W)} \prod_{i=1}^n W(i, \pi(i)), \quad (5)$$

where $\text{per}(W) = \sum_{\pi \in S_n} \prod_{i=1}^n W(i, \pi(i))$ is the normalizing constant, given by the permanent of the matrix $W$.

**Claim 1.** Under $H_0$, the probability $P_{XY}(\pi)$ represents the probability of observing permuted datasets conditional on the marginal sets, that is, $P_{XY}(\pi(D)) = P_{\pi}(\pi \mid x, D_y)$, where $P_0$ denotes the probability under quasi-independence.

**Proof.** For a general weighted model, we have

$$P(\pi(D) \mid x, D_y) = \frac{\prod_{i=1}^n f_{XY}(w)(x_i, y_{\pi(i)})}{\sum_{\pi' \in S_n} \prod_{i=1}^n f_{XY}(w)(x_i, y_{\pi'(i)})}. \quad (6)$$

Under the null, $f_{XY}(w)(x, y) \propto w(x, y)f_X(x)f_Y(y)$, hence

$$P_0(\pi(D) \mid x, D_y) = \frac{\prod_{i=1}^n f_X(x_i)f_Y(y_{\pi(i)})W(i, \pi(i))}{\sum_{\pi' \in S_n} \prod_{i=1}^n f_X(x_i)f_Y(y_{\pi'(i)})W(i, \pi'(i))} = P_{XY}(\pi). \quad \square$$

When $w(x, y)$ is a truncation function, $P_{XY}(\pi)$ is simply the uniform distribution over the set of valid permutations, that is, permutations $\pi$ yielding permuted datasets $\pi(D)$ which are consistent with the truncation. The next lemma shows that permuted data points drawn from Equation (6) follow the distribution of $n$ independent copies of $(X, Y) \sim f_{XY}(w)^n$.

**Lemma 3.1.** Let $\mathcal{D} \sim [f_{XY}(w)]^n$ and conditionally on $\mathcal{D}$ let $\pi$ be a permutation having the conditional probability law given in Equation (6). Then $\pi(\mathcal{D}) \sim [f_{XY}(w)]^n$.

**Proof.** Using Equation (6) and the law of total probability, we have

$$f_\pi(D) = f_{XY}(w) f_X(x_1) f_Y(y_1) \cdots f_X(x_n) f_Y(y_n)$$

$$\quad = \sum_{\pi \in S_n} \prod_{i=1}^n f_{XY}(w)(x_i, y_i) \prod_{i=1}^n f_{XY}(w)(x_i, y_{\pi(i)})$$

$$\quad = \prod_{i=1}^n f_{XY}(w)(x_i, y_i). \quad (7)$$

Recalling that under the null, Equation (6) reduces to (5) (by Claim 1), we have

**Corollary 1.** Under the null, permuted data points drawn according to Equation (5) follow the distribution of $n$ independent copies of $(X, Y) \sim [F_X F_Y]^{\text{permuted}}$.

3.2. The Weighted-Permutation Test of Independence

Let $T(D)$ be any test statistic. The permutation test consists of comparing $T(D)$ to its null distribution over all permuted samples $\pi(D)$, and calculating the $P$-value by the proportions of permutations $\pi_i$ with test statistic $T(\pi_i(D))$ exceeding $T(D)$. In practice, a large number of permutations, $B$, is sampled, and the $P$-value is approximated by

$$P_{\text{value}} = \frac{1}{B + 1} \sum_{i=0}^B I(T(\pi_i(D)) \geq T(D)), \quad (8)$$

where $\pi_0$ is the identity permutation corresponding to $T(D)$. We formalize our weighted-permutation test of independence as shown in Algorithm 1.

**Algorithm 1** Weighted Permutation Test of Quasi-Independence

**Input:** $D$ - sample, $w(x, y)$ - bias function, $T : \mathbb{R}^n \rightarrow \mathbb{R}$ - a test statistic

**Parameters:** $B$ - number of permutations

1. Generate $B$ permutations $\pi_1, \ldots, \pi_B \sim P_{XY}$.
2. Compute the test statistic $T_0 \equiv T(D)$.
3. for $i = 1$ to $B$ do
4. Compute the test-statistic for the permuted dataset, $T_i \equiv T(\pi_i(D))$.
5. **Output:** $P_{\text{value}} \equiv \frac{1}{B + 1} \sum_{i=0}^B I(T_i \geq T_0)$.

**Corollary 1** assures that under the null distribution $P_{\text{value}} \sim U[0, \frac{1}{B + 1}], \frac{B}{B + 1}]$, the type-1 error probability of the weighted permutation test is at most $\alpha$. (The addition of 1 to the denominator and numerator in stage 6, that is, including the original sample, is necessary to ensure Type 1 error below $\alpha$, but can be neglected in practice for large $B$.)

In order to implement Algorithm 1, a method to sample weighted permutations $\pi_i \sim P_{XY}$ is required. An MCMC algorithm that generates such permutations is discussed next.

3.3. Sampling Permutations Using MCMC

The case of sampling uniformly from a restricted set of permutations, that is, permutations $\pi$ with $P_{XY}(\pi) > 0$, was considered
by Diaconis, Graham, and Holmes (2001). Their algorithm deals with the important special case of truncation with a 0/1 weight function. To enable sampling from a general distribution, we use the Metropolis-Hasting (MH) algorithm (Metropolis et al. 1953; Hastings 1970). Let \( \pi_t = (\pi_t(1), \ldots, \pi_t(n)) \) be the permutation at step \( t \). Define the neighbors of \( \pi_t \) to be all permutations obtained from \( \pi_t \) by a single swap, that is,

\[
\text{Neig}(\pi_t) \equiv \{ \pi^{\leftrightarrow}_i \} = \{ \pi_t(1), \ldots, \pi_t(i), \ldots, \pi_t(n) \}, \quad \forall i < j.
\]

We then proceed according to the standard MH algorithm: at each iteration we sample a permutation uniformly from this set \( \pi^{\leftrightarrow}_i \sim U(\text{Neig}(\pi_t)) \), as well as generate a uniform random number \( u \) on \([0, 1]\). Finally, we accept the new permutation only if

\[
u \leq \frac{P(\pi^{\leftrightarrow}_i)}{P(\pi_t)} = \frac{W(i, \pi_t(j))W(i, \pi_t(i))}{W(i, \pi_t(j))W(j, \pi_t(j))}.
\]

A similar algorithm was suggested by Efron and Petrosian (1999) for doubly truncated data. However, for truncated data the weights are all 0 or 1, making the problem much simpler. Algorithm 2 describes our MCMC approach step-by-step.

**Algorithm 2** MCMC for Biased Sampling of Permutations

**Input:** \( D \) - sample, \( w(x, y) \) - bias function

**Parameters:** \( B \) - number of permutations, \( M_0 \) - 'burn-in' number of steps, \( M \) - number of steps between two permutations.

1. Compute \( \text{Var} (i, j) = w(x_i, y_j) \), \( \forall i, j = 1, \ldots, n \).
2. Set \( \pi_0 \) the identity permutation \( \pi_0(i) = i \).
3. for \( t = 0 \) to \( M_0 + BM - 1 \) do
4. Sample \( \pi^{\leftrightarrow}_i \sim U(\text{Neig}(\pi_t)) \), and \( u \sim U[0, 1] \).
5. if \( u \leq \frac{W(i, \pi_t(j))W(i, \pi_t(i))}{W(i, \pi_t(j))W(j, \pi_t(j))} \) then
6. set \( \pi_{t+1} \leftarrow \pi^{\leftrightarrow}_i \).
7. else
8. set \( \pi_{t+1} \leftarrow \pi_t \).
9. Output: The resulting \( B \) permutations \( \pi_{M_0}, \pi_{M_0 + M}, \ldots, \pi_{M_0 + BM} \).

### 4. Importance Sampling

Due to the difficulty of sampling directly from the weighted permutations distribution \( P_W \), we propose here another approach: sample permutations \( \pi_1, \ldots, \pi_B \) according to an importance probability law \( P_{IS} \) such that \( P_{IS}(\pi) > 0 \) whenever \( P_W(\pi) > 0 \), and calculate the \( P \)-value by

\[
P_{\text{value}} = \frac{\sum_{i=0}^{B} P_{IS}(\pi_i) \mathbb{1}_{\{T_{\pi_i} \geq T\}}}{\sum_{i=0}^{B} P_{IS}(\pi_i)}.
\]

(9)

The unknown term \( \text{per}(W) \) appearing in \( P_{IS} \) (see Equation (5)) is canceled in the above equation, thus enabling us to compute the \( P \)-value even when \( P_W \) is known only up to a normalizing constant. Chen and Liu (2007) suggest this importance sampling algorithm for statistical inference under truncated data. It is based on a simple sequential method to generate permutations under \( P_{IS} \). Kou and McCullagh (2009) have generalized one of the approaches proposed in Chen and Liu (2007) to estimate the permanent of general weight functions. We have derived several sequential importance sampling approaches, similar to those of Chen and Liu (2007) and Kou and McCullagh (2009), that are applicable for general weight functions \( w \), and investigated their performances in testing.

Harrison (2012) showed that for any test statistic satisfying mild invariance properties, the test that includes the identity permutation in the \( P \)-value calculation in Equation (9) controls the Type-1 error at level \( \alpha \) (see his Theorem 1). This result applies directly to our case by considering our approach as testing conditionally on the data \( D \).

Although the correction above ensures validity, the importance sampling approach can perform very poorly if the importance distribution \( P_{IS} \) is far from \( P_W \), for example when \( P_{IS} \) is taken to be the uniform distribution over \( S_n \), because in such cases the \( B \) sampled permutations have very low probability under \( P_W \). It is thus challenging to suggest a distribution \( P_{IS} \) that is easy to calculate and sample from, and yet close enough to \( P_W \) for general \( W \)—see supplementary materials, Section D for more details.

### 4. Bootstrap-Based Test

#### 4.1. The Bootstrap Algorithm

The permutation test bypasses the need to estimate the (unbiased) marginal distributions \( f_X, f_Y \), which can be difficult and even impossible when \( w \) vanishes on part of the support of \( f_{XY} \). Nevertheless, when we can estimate the univariate marginals consistently from the data, a bootstrap test is a viable alternative. Briefly, we generate independent samples from the estimated null distribution and compute the test statistic for each such sample. Then we reject the null hypothesis if the observed test statistic is greater than the \( 1 - \alpha \) quantile of the resulting bootstrap distribution. The test is summarized by Algorithm 3.

**Algorithm 3** Bootstrap-Based Test of Quasi-Independence

**Input:** \( D \) - sample, \( w(x, y) \) - bias function, \( T : \mathbb{R}^{2n} \rightarrow \mathbb{R} \) - a test statistic

**Parameters:** \( B \) - number of bootstrap samples

1. Estimate the marginals \( \hat{F}_X, \hat{F}_Y \).
2. Compute the test statistic \( T_0 = T(D) \).
3. for \( i = 1 \) to \( B \) do
4. Generate a bootstrap sample \( D_i \) by sampling with replacement \( n \) i.i.d. examples from \( [\hat{F}_X, \hat{F}_Y]^{(w)} \).
5. Estimate the marginals \( \hat{F}_{X,i}, \hat{F}_{Y,i} \) of the bootstrap sample \( D_i \).
6. Compute the test statistic \( T_i = T(D_i) \).
7. Output: \( P_{\text{value}} \equiv \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}_{\{T_i \geq T\}} \).

As an alternative of estimating the marginal distributions, samples can be drawn under the null from the unbiased conditional distribution of \( X \) given the observed \( Y \) values \( y_1, \ldots, y_t \); Efron and Petrosian (1999) apply this approach to doubly truncated data.

#### 4.2. Estimating the Marginal Distributions

The next challenge is implementing step 1 of Algorithm 3, namely estimating the univariate marginals, given a known bias function \( w(x, y) \). Naturally, under a bias-sampling regime, the underlying marginals may not be identifiable, unless additional modeling assumptions, either on \( f_{XY} \) or \( w(x, y) \), are made.
4.2.1. Case 1: Estimating the Marginal Distributions Under Quasi-independence

For a valid test, it is enough to estimate $F_X$ and $F_Y$ in the observable region under the null hypothesis of quasi-independence. A general algorithm for estimation of the marginal densities in Equation (3) under quasi-independence is developed next. Bickel and Ritov (1991) provided a somewhat similar algorithm for the case where $X$ is discrete, which reduces to the selection bias model of Vardi (1985).

Let $\tilde{X} \sim \tilde{F}_X, \tilde{Y} \sim \tilde{F}_Y$, where $\tilde{F}_X, \tilde{F}_Y$ are the cumulative distribution functions of $f_X, f_Y$ in Equation (3). Under quasi-independence, the density of $X$ can be written as $f_X(x) = \mathbb{E}[w(x, \tilde{Y}) f_Y(x)/\mathbb{E}[w(\tilde{X}, \tilde{Y})]]$. Thus, an estimate of $\tilde{F}_Y$ yields an estimate for $\mathbb{E}[w(x, \tilde{Y})]$, which can be used to build an inverse weighting estimate for $\tilde{F}_X$:

$$\tilde{F}_X(x) = \frac{\sum_{i=1}^n I_{x \leq x_i} \tilde{E}[w(x_i, \tilde{Y})]^{-1}}{\sum_{i=1}^n \tilde{E}[w(x_i, \tilde{Y})]^{-1}}. \quad (10)$$

This estimate can be used in turn to estimate $\tilde{F}_Y$, suggesting an iterative procedure as described in Algorithm 4. For the important case $w(x, y) = 1_{\{x < y\}}$, the algorithm reduces to the standard product-limit (PL) estimator for left and right truncated data, implemented, for example, in the DTDA package of R (Moreira et al. 2010). For more details and an extension to more than two variables, see Supplementary Materials, Section E.

Algorithm 4 Estimation of Marginal Distributions Under Quasi-independence

Input: $\mathcal{D}$ - sample, $w(x, y)$ - bias function, $d(F_1, F_2)$ - distance function.
Parameters: $\epsilon$ - convergence criterion.

1: Generate initial estimates $F_{X}^{\text{new}}, F_{Y}^{\text{new}}$ and set $F_{X}^{\text{fold}} = F_{Y}^{\text{fold}} \equiv 0$.
2: while $d(F_{X}^{\text{fold}}, F_{X}^{\text{new}}) + d(F_{Y}^{\text{fold}}, F_{Y}^{\text{new}}) > \epsilon$ do
3: Set $F_{X}^{\text{fold}} = F_{X}^{\text{new}}$ and $F_{Y}^{\text{fold}} = F_{Y}^{\text{new}}$.
4: Calculate $\tilde{E}[w(x, \tilde{Y})]$, and update $F_{X}^{\text{new}}$ using (10).
5: Calculate $\tilde{E}[w(x, y)]$, and update $F_{Y}^{\text{new}}$ using the equivalent for $\tilde{F}_Y$ of (10).
6: Output $F_{X}^{\text{new}}$ and $F_{Y}^{\text{new}}$.

While Algorithm 4 provides a general procedure to estimate a distribution under independence, for testing purposes it may result in low power. Consider the truncation model $w(x, y) = 1_{\{x < y\}}$. The PL estimators are consistent under the null hypothesis, but using them in our test leads to low power, even in seemingly very extreme situations of a strong dependence. For a test to perform reasonably well for moderate sample sizes, $F_X$ and $F_Y$ should be estimated well not only under the null, but also under the alternative hypothesis (see Section 6 and supplementary materials, Section B). The next example demonstrates this claim.

Example: Difficulties in Detecting Quasi-Independence

Consider the case where data is generated from a uniform bivariate distribution over $0 < x, y < 1 : |x - y| < 0.3$, and let $w(x, y) = 1_{\{x < y\}}$ be the standard truncation model.

We drew $n = 500$ samples from this model and estimated the univariate marginals CDFs using the PL estimators. The left panel of Figure 1 shows the sampled data points. The green and red curves in the right panel are the resulting PL estimates of $\tilde{F}_X$ and $\tilde{F}_Y$, respectively. Because the unbiased variables $\tilde{X}$ and $\tilde{Y}$ are exchangeable, they share the same underlying marginal distribution, depicted by the blue line. The product-limit estimates differ considerably from the true marginal distribution. When such CDFs generate the truncated data, the probability of a selection $(\tilde{X} < \tilde{Y})$ is small, and when it happens, the values of $\tilde{X}$ and $\tilde{Y}$ tend to be close, yielding a scatterplot somewhat similar to the observed data. Indeed, the middle panel of Figure 1 shows pairs obtained by generating independent variables from the estimated product-limit curves $\tilde{F}_X$, $\tilde{F}_Y$ and retaining only observations satisfying $\tilde{X} < \tilde{Y}$. This example shows that independent variables under selection bias can produce data similar to that obtained by strongly dependent variables. Applying the bootstrap test using estimates of the marginal that are consistent only under the null independence assumption may result in a test with low power.

A possible solution is to find estimators for the marginal CDFs that are consistent also under the alternative hypothesis of quasi-dependence. However, as the model is not identifiable under the alternative (Cheng, Hall, and Yang 2007), such estimators can be calculated only under additional assumptions, either on $w(x, y)$ or on the underlying joint distribution (or both). We next demonstrate this through two different settings.

![Figure 1](image-url) (left): a scatterplot of samples drawn from the true underlying distribution. (middle): a scatterplot of samples drawn from the truncated independence distribution, using the PL estimates. (right): the PL estimates of the univariate marginal CDFs $\tilde{F}_X$ and $\tilde{F}_Y$ are biased and do not resemble the true underlying CDF $F_X = F_Y$. 
4.2.2. Case 2: Strictly Positive \( w(x, y) > 0 \)

The problem of estimating nonparametrically a general multivariate distribution \( F \) using weighted data is well known (e.g., Vardi 1985) and for \( w > 0 \) the nonparametric maximum likelihood estimator (NPMLE) is given by

\[
\hat{F}_{XY}^n(t, s) = \frac{\sum_{i=1}^n 1_{[X_i \leq t, Y_i \leq s]} w(X_i, Y_i)^{-1}}{\sum_{i=1}^n w(X_i, Y_i)^{-1}}.
\]

(11)

Estimators for \( F_X \) and \( F_Y \) can be then obtained by marginalization of Equation (11),

\[
\hat{F}_X^n(t) = \hat{F}_{XY}^n(x, \infty) = \frac{\sum_{i=1}^n 1_{[X_i \leq t]} w(X_i, Y_i)^{-1}}{\sum_{i=1}^n w(X_i, Y_i)^{-1}}.
\]

(12)

The estimator above can be used whenever \( w > 0 \) in the entire support of \( F_{XY} \). By the law of large numbers, \( n^{-1} \sum_{i=1}^n 1_{[X_i \leq t]} w(X_i, Y_i)^{-1} \rightarrow F_X(t) / \mathbb{E}_{f_{XY}}[w(X, Y)] \) a.s. and \( n^{-1} \sum_{i=1}^n w(X_i, Y_i)^{-1} \rightarrow 1 / \mathbb{E}_{f_{XY}}[w(X, Y)] \) a.s. so by the continuous mapping theorem \( \hat{F}_X^n(t) \rightarrow F_X(t) \) a.s. By similar arguments, \( \hat{F}_Y^n(s) \rightarrow F_Y(s) \) a.s.

4.2.3. Case 3: Left Truncation \( w(x, y) = 1_{[x < y]} \)

In contrast to the former case, when \( w(x, y) \) is a truncation function, estimating the marginals under quasi-independence is more challenging due to the actual loss of data and identifiability issues. Nevertheless, for certain types of truncation mechanisms, additional assumptions on the joint distribution may allow to reconstruct the marginals. In particular, the most familiar type of truncation in the statistical literature is left (or right) truncation, described by \( w(x, y) = 1_{[x < y]} \). The next proposition shows that under exchangeability, the empirical distribution function of the joint sample \( X_1, \ldots, X_n, Y_1, \ldots, Y_n \) is a consistent estimator for the marginals (the proof is in the supplementary material; methods, Section A).

Proposition 1. Let \( F_{XY}(x, y) \) be an exchangeable joint distribution having a density \( f_{XY}(x, y) \equiv f_{XY}(y, x) \) and let \( D \sim (F_{XY}^{(w)})^n \) be a sample with the truncation weight function \( w(x, y) = 1_{[x < y]} \). Let \( \hat{F}_{XY}^{(w), n} \) and \( \hat{F}_{XY}^{(w), n} \) be the empirical CDFs of \( (X_1, \ldots, X_n) \) and \( (Y_1, \ldots, Y_n) \), respectively. Define:

\[
\hat{F}_X^n(x) = \hat{F}_{XY}^{(w), n}(x) = \frac{f_{XY}^{(w), n}(x)}{2 \sum_{i=1}^n 1_{[x_i \leq x]} + \sum_{i=1}^n 1_{[y_i \leq x]}}.
\]

(13)

Then \( \hat{F}_X^n \), \( \hat{F}_Y^n \rightarrow F_X = F_Y \) a.s.

5. Test Statistics

5.1. The Adjusted Hoeffding Statistic

While the permutation and bootstrap approaches can be applied with any test statistic, our goal is to modify an existing omnibus test to weighted models in general and to truncation in particular. For the latter, most tests used to date are tailored to specific alternatives, such as monotone dependence. A recent new approach studied by Chiu et al. (2018) can test against a general alternative, but uses either significant computational resources or permutations with different sample sizes, so its significance level is not guaranteed. We are inspired by some popular nonparametric tests of independence such as Thas and Ottoy (2004), Heller, Heller, and Gorfinke (2012), and Heller et al. (2016), and for concreteness, we describe the approach of the latter, which generalizes a modified version of Hoeffding (1948).

Our problem requires two major modifications. First, biased sampling should be taken into account when computing the null distribution of the test statistic using a bootstrap or permutations resampling approach; this was addressed in the previous sections. Second, the test statistic compares observed counts with their expectations under the null, and the computation of these expectations needs to be modified to accommodate biased sampling.

The test belongs to a family of tests which compare the observed counts \( a_A \) to the expected counts \( e_A \) for different sets \( A \in \mathbb{R}^2 \). As in Heller et al. (2016), our test statistic is based on Pearson’s Chi-squared statistics, and we consider all partitions defined by the data \( D \). Specifically, each data point \( (x_i, y_i) \in D \) defines a partition of \( \mathbb{R}^2 \) into four quadrants:

\[
Q_{jk}^i \equiv \{ (x', y') \in \mathbb{R}^2 : 1_{[x' > x_i]} = j, 1_{[y' > y_i]} = k \} \quad j, k \in \{0, 1\}.
\]

(14)

For example, \( Q_{00}^i = (-\infty, x_i] \times (-\infty, y_i] \) and \( P((X, Y) \in Q_{00}^i) = F_{XY}^{(w)}(x_i, y_i) \).

Let \( (x_i, y_i) \) be a point in the sample \( D \). For a quadrant \( Q_{jk}^i \), we denote the observed number of points by \( a_{Q_{jk}^i} \equiv a_{Q_{jk}} \) and the expected number of points under the null by \( e_{Q_{jk}^i} \equiv e_{Q_{jk}} \). We then compute, for each quadrant, the scaled squared difference between the observed and expected number of points under \( H_0 \).

Finally, we sum over all the sample points to get our test statistic:

\[
T = \sum_{i=1}^n \sum_{j,k \in \{0,1\}} \frac{(a_{Q_{jk}^i} - e_{Q_{jk}})^2}{e_{Q_{jk}}}. \]

(15)

Estimating the expected values \( e_{Q_{jk}} \) requires the estimation of the null distribution, which may become highly non-trivial in the biased sampling setting. First, \( F_X \) and \( F_Y \) may be non-identifiable, and therefore using plug-in estimators of \( F_X \) or \( F_Y \) may give a poor approximation of the distribution of the test statistic under the null. Second, evaluation of expectations or probabilities under the null may require computationally costly integration of the null distribution, \( [f_{XY}^{(w)}]^{(w)} \), as opposed to a simple multiplication of the empirical marginals in the standard setting. If the expectations \( e_j \) are not estimated correctly, Our weighted permutations test is still valid according to Corollary 1 and the bootstrap approach can be also applied, but power can be severely reduced, as is shown in the supplementary materials, Section C.

5.1.1. Computing Expectations Under Biased Sampling

A natural approach is to estimate the marginals \( F_X \) and \( F_Y \) under the null independence model and use them to calculate the expected count in a certain cell. However, as discussed in Section 4, this approach works well only for special models. We therefore suggest here an alternative method that directly estimates the expected counts.
Let $P_{ij} \equiv P_{W}(\pi(i) = j) = \sum_{\pi \in S_{n}} P_{W}(\pi) 1_{\{\pi(i) = j\}}$, and define the Bernoulli random variables $\xi_{ij} \equiv 1_{\{\pi(i) = j\}}$, so $P_{ij} = \mathbb{E}(\xi_{ij})$. Let $A \subset \mathbb{R}^{2}$ be an arbitrary set. Given a sample $D$, for any permutation $\pi$ of the data denote the number of points in $A$ under $\pi$ (i.e., after permuting the dataset $D$) by $\alpha_{A}(\pi) = \sum_{i=1}^{n} 1_{\{(x_{i},y_{i})\in A\}}$, and let $e_{A} \equiv \mathbb{E}_{P_{W}}[\alpha_{A}(\pi)]$ be the expected number of data points in $A$, under the permutations distribution $P_{W}$.

The $P_{ij}$ values determine the expected number $e_{A}$ for any set $A$ via the following claim:

Claim 2. For any $A \subset \mathbb{R}^{2}$, the expected number of points $e_{A}$ under the permutations distribution $P_{W}$ is given by:

$$e_{A} = \sum_{i,j=1}^{n} 1_{\{(x_{i},y_{i})\in A\}} \mathbb{E}_{P_{W}}(\xi_{ij}^{n}) = \sum_{i,j=1}^{n} 1_{\{(x_{i},y_{i})\in A\}} P_{ij}. \quad (16)$$

Proof. By definition, we have:

$$e_{A} = \mathbb{E}_{P_{W}}\left[\sum_{i=1}^{n} 1_{\{(x_{i},y_{i})\in A\}}\right] = \sum_{i,j=1}^{n} 1_{\{(x_{i},y_{i})\in A\}} P_{ij} = \sum_{i,j=1}^{n} 1_{\{(x_{i},y_{i})\in A\}} P_{ij}. \quad (17)$$

The probabilities $P_{ij}$ can be easily estimated using the MCMC scheme described in Algorithm 2: let $\pi_{0}$ be the identity permutation and $\pi_{1}, \ldots, \pi_{B}$ be the sampled permutations, and define the following estimator:

$$\hat{P}_{ij} \equiv \frac{1}{B+1} \sum_{b=0}^{B} 1_{\{\pi_{b}(i) = j\}}. \quad (18)$$

When we sample permutations using an importance distribution $P_{IS}$ (see Section 3.4), the above estimator for $P_{ij}$ is replaced by:

$$\hat{P}_{ij}^{(IS)} = \frac{\sum_{b=0}^{B} P_{W}(\pi_{b}) 1_{\{\pi_{b}(i) = j\}}}{\sum_{b=0}^{B} P_{W}(\pi_{b})}. \quad (19)$$

Plugging Equation (18) (or Equation (19)) into Equation (16) gives an estimator of $e_{A}$,

$$\hat{e}_{A} = \sum_{i,j=1}^{n} 1_{\{(x_{i},y_{i})\in A\}} \hat{P}_{ij}, \quad (20)$$

which can be used in the chi-squared statistic.

For the bootstrap approach, the estimate of the null distribution is used in a straightforward manner. Consider, for example, the bottom-left quadrant with respect to a point $(x_{i},y_{i})$, $Q_{i}^{0}$, given estimators $\hat{F}_{X}, \hat{F}_{Y}$ of the univariate CDFs, a natural estimator for the mass (up to a normalizing constant) that the null puts on $Q_{i}^{0}$ is given by

$$\hat{c}_{i}^{0} = n[\hat{F}_{X}\hat{F}_{Y}]^{(w)}(x_{i},y_{i}). \quad (21)$$

5.2. An Inverse Weighting Statistic for Strictly Positive $w$

When $w$ is strictly positive, a test can utilize an inverse weighting approach. For a set $A \subset \mathbb{R}^{2}$, define the inverse weighted observed and expected counts $o^{(w)}_{A}$ and $e^{(w)}_{A}$, respectively:

$$o_{A}^{(w)} = \sum_{\ell=1}^{n} 1_{\{(x_{\ell},y_{\ell})\in A\}} w(x_{\ell}, y_{\ell})^{-1},$$

$$e_{A}^{(w)} = nE_{[f_{X},f_{Y}]=w}(1_{A} w(X, Y)^{-1}). \quad (22)$$

For the quadrants $Q_{i}^{k}$, we can compute estimates for the expected weighted counts $e_{i}^{(w)} = e_{i}^{(w)}$ by multiplying the corresponding marginal inverse weighted counts. For example, for $Q_{1}^{0}$:

$$e_{1}^{(w)} = \left(\sum_{\ell=1}^{n} w(x_{\ell}, y_{\ell})^{-1} 1_{\{y_{\ell}\leq 0\}}\right) / \sum_{\ell=1}^{n} w(x_{\ell}, y_{\ell})^{-1} 1_{\{y_{\ell}\leq 0\}}. \quad (23)$$

Similarly to Equation (15), the weighted statistic is given by

$$T^{(w)} = \sum_{i=1}^{n} \sum_{j,k \in \{0,1\}} \left(\frac{c_{i}^{(w)} - o_{i}^{(w)}}{e_{i}^{(w)}}\right)^{2}. \quad (24)$$

5.3. Unknown $w$: Left Truncated Right-Censored Data

In some applications, the biased sampling function $w$ is unknown. However, our methodology is still applicable when $w$ can be estimated consistently. In particular, we can tackle the important case of left truncation with censoring. Consider the standard left-truncation right-censoring model where $(X_{i}, Y_{i}, C_{i})$ are independent triplets, the joint density of $(X_{i}, Y_{i})$ is proportional to $f_{XY}(x_{i}, y_{i}) 1_{\{y_{i}\leq x_{i}\}}$ for some density $f_{XY}$, and $(X_{i}, Y_{i})$ are independent of the censoring variables $C_{i}$. We observe triplets $(X_{i}, \min(Y_{i}, X_{i} + C_{i}), \Delta_{i})$, where $\Delta_{i} = 1_{\{Y_{i}\leq X_{i} + C_{i}\}}$ are the censoring indicators with $\Delta_{i} = 0$ and $1$ for censored and uncensored observations, respectively. We suggest testing independence based on the uncensored observations, where censored observations are used only for estimation of the weight function. Specifically, the conditional density of an uncensored observation is simply

$$S(y-x) f_{XY}(x,y) 1_{\{y\leq x\}} / \mathbb{E}_{f_{XY}}[S(Y-X) 1_{\{Y\leq X\}}], \quad (25)$$

where $S(t) = P(C > t)$ is the survival function of $C$. Thus, the density of uncensored observations has exactly the form of Equation (1), with $w(x,y) = 1_{\{y\leq x\}} S(y-x)$, a function involving a continuous and a truncated part. The methods developed in previous sections are flexible enough to accommodate such functions. The survival function $S$ can be estimated by the standard Kaplan-Meier estimator applied to the data $(\min(C_{i}, Y_{i} - X_{i}), 1 - \Delta_{i}), i = 1, \ldots, n$.
6. Simulation Studies

We investigated, using simulation, the performances of the weighted permutation (WP) and bootstrap tests, and compared them to that of Tsai’s (1990) and the minimum P-value (minP2) test of Chiu et al. (2018). The latter are applicable only to truncated data of the form \( w(x, y) = \mathbb{I}_{[x < y]} \).

We implemented the simulations in R, with time consuming parts implemented in C++ using the Rcpp package (Eddelbuettel et al. 2011). Scripts reproducing all figures and tables are available online at https://github.com/YanivTenzer/TIBS. P-values for the minP2 test were calculated using the package permDep in R (Chiu 2018), version 1.0.3 (Aug, 14th, 2019) from https://github.com/stc04003/permDep.

We calculated the rejection rate (power) at a significance level of \( \alpha = 0.05 \) by averaging results of 500 replications. As the tests are computationally demanding, we used small sample sizes of \( n = 100 \) and \( n = 200 \) observations for uncensored and censored settings, respectively, in order to perform extensive simulations under different settings. We used \( B = 1000 \) permuted or bootstrap null datasets for all tests except minP2 for which we used only \( B = 100 \) null datasets and 100 replications, as it was much slower. The average running times of the tests on a standard laptop with an i7 2.8Ghz dual core Intel processor were \( \approx 0.02 \) seconds for Tsai's test, \( \approx 0.11 \) seconds for the new weighted permutation test, \( \approx 2.86 \) seconds for the bootstrap test and \( \approx 93.45 \) seconds for the minP2 test.

6.1. Truncation, \( w(x, y) = \mathbb{I}_{[x < y]} \)

We study the performances of the tests under truncation for various dependence models with and without censoring. The censoring variable, \( C \), was sampled from Gamma distributions, with the shape and scale parameters set such that roughly 25% – 30% of the observations were censored for each model.

We first simulated data under monotone dependence models with an exchangeable joint distribution, where consistent estimators of the marginals exist (as shown in Section 4) and we expect the bootstrap procedure to perform well. We generated \( X \) and \( Y \) from a standard bivariate Gaussian distribution with different correlations \( \rho \) (Norm(\( \rho \))). In addition, we generated \( X \) and \( Y \) with standard Gaussian marginal distributions under two copula models: (i) The Gumbel copula (GC) with dependence parameter \( \theta = 1.6 \) (Kendall’s \( \tau = 0.375 \)), and (ii) The Clayton copula (CC) with dependence parameter \( \theta = 0.5 \) (\( \tau = 0.2 \)). Although both the Gumbel and Clayton copulas produce monotone dependence structures, the two are different in nature—while the former provides upper tail dependence structure the latter produces lower tail dependence (Nelsen 2007). Figure B.1 in the supplementary materials presents scatterplots of simulated pairs from the three models.

The results are summarized in Table 1, with the test having the highest power shown in boldface. As expected, in the Gaussian settings, under the null distribution (i.e., \( \rho = 0 \)), all tests achieve the correct \( \alpha = 0.05 \) error rate. Under the alternative, the bootstrap procedure demonstrates favorable performance in all three settings. In the Gaussian settings, for \( \rho < 0 \) the WP test consistently outperforms minP2. The minP2 has the lowest power in this setting. For \( \rho > 0 \), the WP test has poorer performance, probably due to the difficulty of sampling permutations consistent with the truncation, while Tsai’s test shows the second highest power, after the bootstrap.

Following Chiu et al. (2018), we next simulated data from a lifetime distribution (LD) where the joint distribution is non-exchangeable having marginal distributions \( X \sim \text{Exp}(5) \) and \( Y \sim \text{Weibull}(3, 8.5) \). We specified the dependence of \((X, Y)\) through a normal copula, where the strength of dependence is determined by the correlation parameter \( \rho \). We simulated data under independent \((\rho = 0)\) and dependent \((\rho = 0.4)\) \(F_{XY}\)

### Table 1. Power at a significance level of \( \alpha = 0.05 \) for left-truncated data \( w(x, y) = \mathbb{I}_{[x < y]} \) for uncensored (left 4 columns) and censored (right 3 columns) models.

| Setting                      | Model          | Uncensored (n = 100) | Censored (n = 200) |
|------------------------------|----------------|----------------------|---------------------|
|                              | Tsai | minP2 | WP | Bootstrap | Tsai | MinP2 | Tsai |
| Monotone Exchangeable        |      |       |    |           |      |       |      |
| Norm(−0.9)                   | 1    | 1     | 1  | 1         | 1    | 1     | 1    |
| Norm(−0.7)                   | 0.998| 0.960 | 0.998| 1         | 1    | 0.828 | 1    |
| Norm(−0.5)                   | 0.764| 0.510 | 0.742| 0.998     | 0.813| 0.284 | 0.895|
| Norm(−0.3)                   | 0.284| 0.130 | 0.278| 0.828     | 0.273| 0.096 | 0.366|
| Norm(0.0)                    | 0.056| 0.050 | 0.064| 0.058     | 0.052| 0.042 | 0.046|
| Norm(0.3)                    | 0.178| 0.110 | 0.118| 0.780     | 0.097| 0.083 | 0.202|
| Norm(0.5)                    | 0.352| 0.140 | 0.194| 1         | 0.211| 0.103 | 0.404|
| Norm(0.7)                    | 0.498| 0.290 | 0.262| 1         | 0.389| 0.097 | 0.729|
| Norm(0.9)                    | 0.658| 0.260 | 0.236| 1         | 0.495| 0.076 | 0.907|
| GC (\( \theta = 1.6 \))      | 0.196| 0.130 | 0.104| 1         | 0.079| 0.080 | 0.192|
| CC (\( \theta = 0.5 \))      | 0.110| 0.130 | 0.074| 0.782     | 0.126| 0.127 | 0.181|
| LD (\( \rho = 0.0 \))        | 0.042| 0.010 | 0.046| 0.996     | 0.005| 0.048 | 0.051|
| LD (\( \rho = 0.4 \))        | 0.634| 0.330 | 0.578| 0.704     | 0.060| 0.056 | 0.250|
| Nonmonotone Exchangeable     |      |       |    |           |      |       |      |
| CLmix(0.5)                   | 0.278| 0.140 | 0.412| 0.338     | 0.398| 0.120 | 0.308|
| CNorm(−0.9)                  | 0.992| 1     | 1   | 1         | 1    | 0.987 | 0.897|
| CNorm(−0.7)                  | 0.844| 0.950 | 0.992| 0.998     | 0.972| 0.609 | 0.635|
| CNorm(−0.5)                  | 0.514| 0.600 | 0.794| 0.988     | 0.626| 0.262 | 0.321|
| CNorm(−0.3)                  | 0.176| 0.290 | 0.272| 0.982     | 0.212| 0.138 | 0.157|
| CNorm(0.0)                   | 0.042| 0.020 | 0.046| 0.986     | 0.047| 0.053 | 0.055|
| CNorm(0.3)                   | 0.132| 0.170 | 0.216| 0.998     | 0.228| 0.035 | 0.089|
| CNorm(0.5)                   | 0.266| 0.370 | 0.654| 1         | 0.753| 0.037 | 0.163|
| CNorm(0.7)                   | 0.490| 0.790 | 0.992| 1         | 0.999| 0.080 | 0.268|
| CNorm(0.9)                   | 0.698| 1     | 1   | 1         | 1    | 0.298 | 0.367|
before truncation. Figure B.2 in the Supp. Materials displays scatterplots of simulated pairs from both models.

The results are shown in the second part of Table 1. Although X and Y are not exchangeable, we applied the bootstrap procedure (shown in light gray) as well, using marginal estimates according to Equation (13), in order to investigate the impact of model misspecification on its performance. The devastating impact of model misspecification under the null distribution for the bootstrap procedure is now apparent: the test does not retain the desired rejection rate under the null hypothesis (ρ = 0). For ρ = 0.4, Tsai’s test has the highest power, followed by the WP test and then minP2.

The third simulation study evaluates the performance of the tests under nonmonotone dependence. Starting with a non-monotone exchangeable model, we simulate data from a mixture of two Clayton copulas with dependence parameters θ = 0.5 (τ = 0.2) and θ = −0.5 (τ = −0.333), respectively, and equal population proportions. Figure B.3 in the supplementary materials presents scatterplots of simulated pairs from the model. The third part of Table 1 presents the results. Our bootstrap approach has the highest power, followed by the WP test, Tsai’s test and lastly minP2. It is somewhat surprising that Tsai’s test outperforms here minP2, as the former is tailored to monotone alternatives.

Finally, we considered a nonmonotone and non-exchangeable model. We used a normal copula, CNorm(ρ), with varied correlation coefficient ρ to specify the joint distribution of X and Y, where X ∼ Weibull(0.5, 4) and Y ∼ U[0, 16] are set such that E(X) = E(Y) = 8, and retained only pairs satisfying Y ≥ X. Figure B.4 in the supplementary materials displays scatterplots of simulated pairs from the model. The last part of Table 1 presents the results. As expected, the bootstrap procedure (light gray) does not retain the desired rejection rate under the null hypothesis. Under the alternative, both the WP and minP2 tests outperform Tsai’s procedure across the entire range of the dependence parameter ρ. This behavior is expected because both tests were designed to detect non-monotone dependence. Our method has the highest power for all values of ρ. minP2 is more powerful than Tsai’s test for strong (absolute) correlations and less powerful for weaker correlations.

### 6.2. Strictly Positive Bias Functions

Our new tests can be used to detect dependency in a general weighted model. To study the performance of our approach, we consider two cases of positive biased sampling. In the first case, we took w(x, y) = x + y, where X and Y were sampled from the log-normal bivariate distribution with zero mean, unit variance and correlation ρ. Figure B.5 top panels in the supplementary materials present scatterplots of simulated pairs from the model, for different values of ρ. Recall that for a strictly positive w, the inverse weighted Hoeffding statistic from Section 5.2 can be used as an alternative to the adjusted Hoeffding statistic. We also compared here the bootstrap and MCMC approaches to the importance sampling approach from Section 3.4, with four different importance sampling distributions, described in the Supp. Methods, Section D. We applied six different sampling methods for p-value calculation for each of the two test statistics, resulting in twelve different tests. When applying the Bootstrap, we estimated the univariate marginals using the weighted estimators in Equation (12). Table 2 shows the rejection rates of the various tests at a significance level α = 0.05, for sample size n = 100 and for ρ = 0 (independence) and ρ = 0.2. All tests except the bootstrap seem to maintain the significance level at approximately α = 0.05 under the null. Under the alternative, the importance sampling approach with a uniform and “grid” importance distribution is most powerful, but the MCMC approach is not far behind.

In the second example, we considered a bivariate Gaussian distribution Norm(ρ) for X and Y as in Section 6.1, and with w(x, y) ∝ Norm(−ρ). Figure B.5 bottom panels in the supplementary materials present scatterplots of simulated pairs from the model, for different values of ρ. The biased sampling function here masks the dependence so the observed pairs, (X_i, Y_i), are independent Gaussian random variables.

| Model          | IS-Dist. | WPIS Kou-McCullagh | WPIS Uniform | WPIS Monotone | WPIS Grid | WP | Bootstrap |
|----------------|----------|---------------------|--------------|---------------|-----------|---------|-----------|
| LogNormal(ρ = 0) |          | 0.048               | 0.050        | 0.016         | 0.050     | 0.056   | 0.072     |
|                |          | 0.050               | 0.038        | 0.016         | 0.036     | 0.060   | 0.080     |
| LogNormal(ρ = 0.2) |        | 0.606               | 0.676        | 0.004         | 0.676     | 0.602   | 0.632     |
|                |          | 0.384               | 0.432        | 0.000         | 0.428     | 0.382   | 0.414     |
| Norm(0.0)      |          | 0.046               | 0.048        | 0.032         | 0.040     | 0.046   | 0.004     |
|                |          | 0.056               | 0.058        | 0.032         | 0.054     | 0.050   | 0.004     |
| Norm(0.1)      |          | 0.084               | 0.084        | 0.052         | 0.080     | 0.082   | 0.006     |
|                |          | 0.076               | 0.076        | 0.032         | 0.076     | 0.076   | 0.008     |
| Norm(0.3)      |          | 0.296               | 0.322        | 0.134         | 0.350     | 0.302   | 0.052     |
|                |          | 0.282               | 0.296        | 0.136         | 0.318     | 0.298   | 0.024     |
| Norm(0.5)      |          | 0.582               | 0.640        | 0.298         | 0.630     | 0.582   | 0.166     |
|                |          | 0.502               | 0.526        | 0.228         | 0.532     | 0.492   | 0.038     |
| Norm(0.7)      |          | 0.856               | 0.860        | 0.516         | 0.850     | 0.852   | 0.334     |
|                |          | 0.730               | 0.706        | 0.456         | 0.684     | 0.726   | 0.078     |
| Norm(0.9)      |          | 0.964               | 0.914        | 0.712         | 0.918     | 0.964   | 0.672     |
|                |          | 0.876               | 0.800        | 0.662         | 0.786     | 0.878   | 0.114     |
This example shows that biased sampling can not only create spurious dependencies, but can also mask true dependencies. Nevertheless, knowing $w$ we can apply our tests and detect the dependence, as is shown in Table 2. As expected, for all tests, as $\rho$ increases (we used only $\rho > 0$ due to symmetry) the power increases quite rapidly. The importance sampling with a uniform distribution is usually the most powerful, with the MCMC approach very close or superior for strong correlation. The bootstrap approach shows poor performance for this case. As in the previous example, here too the inverse weighting statistic is inferior to the adjusted Hoeffding statistic. The relative success of the importance sampling schemes for both examples can be explained by the small sample size. As shown in the supplementary materials, Section D, when the sample size increases the importance sampling distributions become unrepresentative of the distribution $P_{X,Y}$, resulting in poor performance, whereas the MCMC approach is much more robust to changes in sample size.

**Figure 2.** Scatterplots of four real datasets. Red points indicate the true data. Green ‘+’ signs represent points $(x_i, y_{\pi(i)})$ for randomly drawn permutations $\pi$ under the null quasi-independence model. The biased sampling functions are shown as background in grayscale, with $w(x, y) = 1_{x<y}$ for the AIDS dataset, $w(x, y) = 1_{x<y}S(y-x)$ for the Channing housing dataset, $w(x, y) = x + y$ for the Infection dataset and $w(x, y) = \min(65 - x - y, 18)1_{x+y<65}$ for the Huji dataset.
7. Real-Life Datasets

We applied the various tests to four datasets, shown in Figure 2. The results are shown in Table 3. P-values for the WP and bootstrap tests are based on \( B = 10^5 \) samples, and for the minP2 test on \( B = 10^4 \), due to computational restrictions.

1. Time from Infection to AIDS in HIV Carriers—A classical example of truncated data occur in AIDS retrospective studies, where the time from HIV infection to AIDS (Y) is restricted to be smaller than the time from HIV to sampling (X) (Lagakos, Barraj, and de Gruttola 1988). Here, we analyze data on 295 AIDS cases, available in the DTDATA package of R (Moreira et al. 2010). By design, the sample comprised only patients satisfying \( 0 \leq X \leq Y \). The new WP, Tsai’s and minP2 tests all yielded significant P-values at the standard \( \alpha = 0.05 \) level, suggesting that dependence exists between the two time variables. To examine the effect of ignoring the truncation mechanism when testing for independence, we also performed a WP test with a constant \( w \); the P-value remained significant and in fact was reduced to \( 10^{-5} \).

2. Survival in the Channing House Community—Hyde (1977) analyzed survival data for residents of the Channing House retirement community in Palo Alto, CA; the full dataset is available in the boot R package (Canty and Ripley 2017). The survival time, \( Y \), of a resident is left truncated by the entering age to the community, \( X \), and is right censored by the age at the end of followup. After removing five observations that were not consistent with the criterion \( X < Y \), the data consisted of \( n = 457 \) individuals, 282 (61.7%) of which were censored. Quasi independence of survival time and entering time was tested applying the approach described in Section 5.3. Tsai’s and minP2 tests were calculated for comparison. The WP, Tsai’s and minP2 tests were all non-significant, showing no evidence for dependence between entering age and survival. As expected, when we ignored the truncation mechanism, a naive WP test of independence considering only the censoring and using \( w(x, y) = S(y-x) \) (see Equation (25)) yielded a spurious signal of dependence, with a P-value of \( 10^{-5} \), demonstrating the need to account for biased sampling when testing.

3. Time before and after infection in Intensive-Care-Units—Data on 137 patients that were hospitalized in Intensive-Care-Units (ICUs) on a random day were collected in five Israeli hospitals as part of a national cross-sectional study (Mandel 2010). Infection data were collected from admission to the ICU until discharge or 30 days, whatever comes first. An important question was whether the time of infection is associated with the remaining time in the ICU. To test this independence hypothesis, we use the sub-sample of patients who admitted to the ICU without delay and who acquired infections during their first 30 days of hospitalization in the ICU. Thus, we test independence of \( X \), the time from admission to the ICU to infection and \( Y \), the time from infection to discharge from the ICU. Due to the sampling mechanism, the data are length biased according to the total length of stay in the ICU, yielding the weight function \( w(x, y) = x + y \). The new WP and bootstrap tests were both significant, indicating that the time of acquiring infection shows a significant effect on prolonging the remaining time in the ICU. When we used a WP test while ignoring the biased sampling function \( w \), the signal for dependence disappeared, and we got a P-value of 0.605. Thus, for this dataset the biased sampling masks the true dependence between \( X \) and \( Y \), and our test that takes \( w \) into account was able to reveal it.

4. Time to Promotion to the Rank of Full Professor—The data consist of cross-sectional records on all faculty members of the Hebrew University of Jerusalem who were employed in 1998. We tested whether the age at promotion to the associate professor rank depends on the service time in that rank. Let \( A_{AP} \) and \( A_{FP} \) denote the age at promotion to the ranks associate professor (AP) and full professor (FP), respectively; we test independence between \( X = A_{AP} \) and \( Y = A_{FP} - A_{AP} \) for associate professors promoted to full professor before the age of 65, back then the retirement age in Israel. We used the sub-sample of 306 faculty members who were promoted to the FP rank after 1980 and were younger than 65 at sampling time (1998). As in Mandel and Rinott (2012), we assume that professors will stay in the university until age 65. Assuming a stable entrance process to the AP rank, the cross-sectional study design leads to length biased sampling according to the length of service at the FP rank. The restriction of the data to professors who promoted after 1980 resulted in the weight function \( w(x, y) = \min(65 - x - y, 18) \mathbf{1}_{\{x+y<65\}} \). Thus, the weight is neither a truncation function nor strictly positive, and the only test applicable is the permutation test of Section 3.

The WP test with \( B = 10^5 \) permutations yielded a very small P-value of 0.00002, meaning that the age of promotion to associate professor rank does depend on the service time in that rank. Ignoring the biased sampling function \( w \) still yielded a significant P-value of 0.00060.

8. Discussion

In this article we addressed the problem of testing quasi-independence in the presence of a general bias function and in possibly nonmonotone settings. As demonstrated using real-life datasets, testing independence naïvely, while ignoring the bias function, can either create spurious dependence or mask true dependence.

We introduced two general machineries to simulate samples under the biased null distribution, namely, the permutations
and bootstrap approaches, and examined the challenges possessed by both. Concretely, the former requires drawing permutations from a (general) nonuniform distribution, while the latter requires consistent estimation of the univariate marginals under the alternative hypothesis. We tackle the first challenge by utilising an MCMC scheme and an importance sampling methodology. Our simulation study indicates that for large sample size, the latter approach may suffer some degradation in performance, due to the difficulty in finding a proposal distribution suited for a general bias function.

On the bootstrap front, we identified two settings in which consistent estimators can be derived, both under the null and the alternative. We also introduced a new algorithm for estimating the marginal CDFs under the null, for a general bias function. This is of independent interest in cases where the null hypothesis is not rejected.

Importantly, both the permutations and bootstrap approaches can be combined with different statistics, thus result in a different test. As shown in simulations and real-life datasets, the choice of statistic can affect the power of the resulting test.

Our proposed tests extend the applicability of testing independence under biased sampling beyond existing methods in the literature; a high-level overview is shown in Table 4. The table presents the family of weight functions, alternatives that can be detected by the tests and algorithms for calculating the $P$-values. While truncation problems have been studied quite intensively in the literature, testing independence under a general weight function is, for the best of our knowledge, new. Tests based on importance sampling (see Section 3.4) can in principle be applied to any bias function $w$ and can detect any dependence for which they were designed, but may be impractical as shown in the supplementary information, Section D. Our new bootstrap test is applicable for positive weight functions and a special case of a truncation function with exchangeable distributions (see Section 4). Our weighted permutations test (WP) is the first test applicable for general biased sampling mechanisms and general alternatives.

An appealing feature of the methodologies developed here is that they can be easily adapted to cases where the bias function is not known, but can be estimated from the data. An important instance of such cases is that of censoring. In particular, in observational studies, left truncation right censoring settings are frequently encountered. Our tests can accommodate such settings and it would be interesting to explore this approach further to more general cases of censoring and biased sampling.

We demonstrated the merit of our proposed tests, using simulated and real-life datasets, and showed that even for truncated data they attain similar and often higher power in most settings considered here, compared to the recent minP2 test.

Finally, some theoretical aspects of the proposed algorithms are yet to be explored and left for future work. In particular we conjecture that, under the assumption of quasi-dependence, both the WP and bootstrap test are consistent.

### Supplementary Material

A. Proofs, B. Simulated Data, C. Computing the Expectations for the Statistical Under the Null, D. Algorithms for Importance Sampling, E. An Iterative Algorithm for Marginal Estimation, and F. Implementation Details.

### Acknowledgments

The authors thank Tair Heller, Tair Goldberg, Malka Gorfine, Sy Han Chiou, and Rebecca Betensky for fruitful discussions. They also thank three anonymous referees for helpful comments and suggestions.

### Funding

This research was supported by the NIH (grant no. R01NS094610) and by The Israel Science Foundation (grant No. 519/14).

### References

Bickel, P. J., and Ritov, J. (1991), "Large Sample Theory of Estimation in Biased Sampling Regression Models. I." The Annals of Statistics, 797–816. [2194,2199]

Canty, A., and Ripley, B.D. (2017), *boot: Bootstrap R (S-Plus) Functions*, R package version 1.3-20. [2204]

Chen, C.-H., Tsai, W.-Y., and Chao, W.-H. (1996), "The Product-moment Correlation Coefficient and Linear Regression for Truncated Data," Journal of the American Statistical Association, 91, 1181–1186. [2195]

Chen, Y., and Liu, J. S. (2007), "Sequential Monte Carlo Methods for Permutation Tests on Truncated Data," Statistica Sinica, 857–872. [2195,2197]

Cheng, M.-Y., Hall, P., and Yang, Y.-I. (2007), "Nonparametric Inference Under Dependent Truncation," Acta Scientiarum Mathematicarum, 73, 397–422. [2195,2198]

Chiou, S.H. (2018), *permDep: Permutation Tests for General Dependent Truncation*. https://CRAN.R-project.org/package=permDep. R package version 1.0.2. [2201]

Chiou, S.H., Qian, J., Mormino, E., Betensky, R. A. (2018), "Lifestyle Flagship Study of Aging, Harvard Aging Brain Study, Alzheimer's Disease Neuroimaging Initiative, et al. Permutation tests for General Dependent Truncation," *Computational Statistics & Data Analysis*, 126, 308–324. [2194,2195,2199,2201]

de Uña-Alvarez, J. (2012), "On the Markov Three-state Progressive Model," in *Recent Advances in System Reliability*, eds. A. Lisiannisi and I. Frenkel, pp. 269–281. London: Springer. [2195]

Diaconis, P., Graham, R., and Holmes, S. P. (2001), "Statistical Problems Involving Permutations With Restricted Positions," *Journal Notes Monograph Series*, pp. 195–222. [2197]

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011), "Rcpp: Seamless r and c++ Integration," *Journal of Statistical Software*, 40, 1–18. [2201]
Efron, B., and Petrosian, V. (1999), “Nonparametric Methods for Doubly Truncated Data,” *Journal of the American Statistical Association*, 94, 24–834. [2195,2197]

Elderton, E.M., Barrington, A., Jones, H. G., Lamotte, E., Laski, H., and Pearson, K. (1913), “On the Correlation of Fertility With Social Value. A Cooperative Study,” *The Eugenics Review*, 5, 369. [2194]

Emura, T., and Wang, W. (2010), “Testing Quasi-independence for Truncation Data,” *Journal of Multivariate Analysis*, 101, 223–239. [2195]

Goodman, L. A. (1968), “The Analysis of Cross-classified Data: Independence, Quasi-independence, and Interactions in Contingency Tables With or Without Missing Entries: Ra Fisher Memorial Lecture,” *Journal of the American Statistical Association*, 63, 1091–1131. [2194]

Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., and Smola, A. J. (2008), “A Kernel Statistical Test of Independence,” In *Advances in Neural Information Processing Systems*, pp. 585–592. [2194]

Harrison, M. T. (2012), “Conservative Hypothesis Tests and Confidence Intervals Using Importance Sampling.” *Biometrika*, 99, 57–69. [2197]

Hastings, W. K. (1970), *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Oxford: Oxford University Press. [2197]

Heller, R., Heller, Y., and Gorfine, M. (2012), “A Consistent Multivariate Test of Association Based on Ranks of Distances,” *Biometrika*, 100, 503–510. [2194,2199]

Heller, R., Heller, Y., Kaufman, S., Brill, B., and Gorfine, M. (2016), “Consistent Distribution-free k-sample and Independence Tests for Univariate Random Variables,” *Journal of Machine Learning Research*, 17, 1–54. [2194,2195,2199]

Kou, S. C., and McCullagh, P. (2009), “Approximating the α-permanent,” *Biometrika*, 96, 635–644. [2197]

Lagakos, S. W., Barraj, L. M., and de Gruttola, V. (1988), “Nonparametric Analysis of Truncated Survival Data, With Application to AIDS,” *Biometrika*, 75, 515–523. [2204]

Mandel, M. (2010), “The Competing Risks Illness–death Model Under Cross-sectional Sampling,” *Biostatistics*, 11, 290–303. [2204]

Mandel, M., and Rinott, Y. (2012), *Cross-Sectional Sampling, Bias, Dependence, and Composite Likelihood*, Center for the Study of Rationality. [2204]

Martin, E. C., and Betensky, R. A. (2005), “Testing Quasi-independence of Failure and Truncation Times Via Conditional Kendall’s Tau,” *Journal of the American Statistical Association*, 100, 484–492. [2195]

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), “Equation of State Calculations by Fast Computing Machines,” *The Journal of Chemical Physics*, 21, 1087–1092. [2197]

Nelsen, R. B. (2007), *An Introduction to Copulas*, Springer Science & Business Media, New York: Springer. [2201]

Rodriguez-Girondo, M., and de Uña-Álvarez, J. (2012), “A Nonparametric Test for Markovianity in the Illness-Death Model,” *Statistics in Medicine*, 31, 4416–4427. [2194,2195]

Székely, G. J., Rizzo, M.L. (2009), “Brownian Distance Covariance,” *The Annals of Applied Statistics*, 3, 1236–1265. [2194]

Tsai, W-Y. (1990), “Testing the Assumption of Independence of Truncation Time and Failure Time,” *Biometrika*, 169–177. [2194,2195,2201]

Vardi, Y. (1985), “Empirical Distributions in Selection Bias Models,” *The Annals of Statistics*, 13, 178–203. [2198,2199]