Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Anytime-Valid Tests of Conditional Independence Under Model-X

Peter Grünwald, Alexander Henzi & Tyron Lardy

To cite this article: Peter Grünwald, Alexander Henzi & Tyron Lardy (2024) Anytime-Valid Tests of Conditional Independence Under Model-X, Journal of the American Statistical Association, 119:546, 1554-1565, DOI: 10.1080/01621459.2023.2205607

To link to this article: https://doi.org/10.1080/01621459.2023.2205607

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 546, 1554–1565: Theory and Methods
https://doi.org/10.1080/01621459.2023.2205607

Anytime-Valid Tests of Conditional Independence Under Model-X

Peter Grünwald [a,b], Alexander Henzi[c], and Tyron Lardy[a,b]

[a]Department of Machine Learning, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands; [b]Department of Statistics, Leiden University, Leiden, The Netherlands; [c]Seminar for Statistics, ETH Zürich, Switzerland

ABSTRACT

We propose a sequential, anytime-valid method to test the conditional independence of a response Y and a predictor X given a random vector Z. The proposed test is based on e-statistics and test martingales, which generalize likelihood ratios and allow valid inference at arbitrary stopping times. In accordance with the recently introduced model-X setting, our test depends on the availability of the conditional distribution of X given Z, or at least a sufficiently sharp approximation thereof. Within this setting, we derive a general method for constructing e-statistics for testing conditional independence, show that it leads to growth-rate optimal e-statistics for simple alternatives, and prove that our method yields tests with asymptotic power one in the special case of a logistic regression model. A simulation study is done to demonstrate that the approach is competitive in terms of power when compared to established sequential and nonsequential testing methods, and robust with respect to violations of the model-X assumption. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2022
Accepted April 2023

KEYWORDS
E-value; Logistic regression; Optional stopping; Sequential test

1. Introduction

A fundamental task in many areas of research, such as economics, genetics, and pharmacology, is to find out whether there is an association between a response Y and an explanatory variable X, given a vector of covariates Z. Mathematically, absence of such an association is defined as conditional independence (CI) between X and Y given Z, denoted by $X \perp\!\!\!\perp Y \mid Z$ (Dawid 1979). A standard way to tackle these problems is to assume a (semi)parametric model on Y given X and Z, encoding the dependence between Y and X in a model parameter. For example, in the logistic model the probability that a binary random variable Y equals one is regressed on (X, Z), and the regression coefficient corresponding to X is zero if and only if $X \perp\!\!\!\perp Y \mid Z$. Within these parametric models, there are well-established methods to test conditional independence, such as the likelihood ratio test for generalized linear models (McCullagh and Nelder 1989). However, all of these tests have in common that they fail to uphold a Type-I error guarantee when the model assumptions are not satisfied. If Z is continuously distributed such error inflation is in fact unavoidable: unless further assumptions on the distribution of (X, Y, Z) are imposed, there exist no nontrivial tests of CI which maintain Type-I error guarantees (Shah and Peters 2020).

However, Candès et al. (2018) show that given additional knowledge, that is, the distribution of X conditional on Z, nontrivial tests of conditional independence can be designed without further assumptions on the distribution of (X, Y, Z). This has been dubbed the Model-X (MX) setting, and the condition that the distribution of X conditional on Z is known is called the Model-X (MX) assumption, though there are settings where the MX "assumption" is *known* to be true. Perhaps the most prominent example is a randomized clinical trial, where the distribution of treatment/control is imposed by the researchers. Another example is conjoint analysis, which is a survey-based experiment where respondents are asked to express a preference between multiple hypothetical products with different attributes. These attributes are randomized according to a distribution chosen by the researchers, with the aim to find out whether one or more of the attributes have an influence on consumer preference (see e.g., Ham, Imai, and Janson 2022). Furthermore, Candès et al. (2018) show that there are ample scenarios where at least an accurate estimate of the conditional distribution of X given Z is known, while also acknowledging that the MX assumption might not be appropriate if this is not the case.

Under the MX assumption, Candès et al. derive the conditional randomization test (CRT), which has nontrivial power to detect CI. Recently, much effort has gone into relaxing the MX assumption and improving the robustness of the CRT under misspecification of the conditional distribution of $X \mid Z$ (Katsevich and Ramdas 2022; Li and Liu 2022; Niu et al. 2022). The CRT and most of its extensions have in common that they are based on p-values computed on batches of data, and therefore designed for fixed sample size experiments. In this article, we focus on anytime-valid tests for CI under MX. Anytime-valid tests allow for more flexibility when testing compared to the nonsequential CRT and also allow covariate adaptive designs, where the distribution of X does not only depend on Z, but also on past data, such as in response-adaptive sampling schemes. Previous work on anytime-valid tests of CI under MX has been

CONTACT Peter Grünwald *pdg@cwi.nl* Department of Machine Learning, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

done by Duan, Ramdas, and Wasserman (2022), who propose tests for the case that X is binary. Their approach is based on pragmatic game-theoretic principles, whereas in this article, we aim to describe and theoretically analyze anytime-valid tests for CI for general X. Shaer, Maman, and Romano (2022) have worked on an extension of the work by Duan, Ramdas, and Wasserman (2022) concurrently, and we discuss their work and the connections to this article in Section 6.

Our hypothesis tests are based on the concept of e-statistics[1] (Grünwald, de Heide, and Koolen 2019; Vovk and Wang 2021; Ramdas et al. 2022). E-statistics have been introduced as an alternative to p-values that is inherently more suitable for testing under optional stopping and continuation (Grünwald, de Heide, and Koolen 2019; Ramdas et al. 2020; Vovk and Wang 2021). While they have their roots in the work on anytime-valid testing by H. Robbins and students (e.g., Darling and Robbins 1967), interest in them has exploded in recent years; see, for example, also Shafer (2021) and Grünwald (2022). E-statistics can be associated with a natural notion of optimality, called GRO (growth-rate optimality) by Grünwald, de Heide, and Koolen (2019), which may be viewed as an analogue of statistical power in an optional stopping context. We derive a general method for constructing e-statistics for conditional independence testing under model-X and show that the method that we propose is optimal in this GRO sense for testing conditional independence against point alternatives under MX. This result should be seen as an anytime-valid analogue to the result by Katsevich and Ramdas (2022), who use the Neyman-Pearson lemma to derive test statistics for which the conditional randomization test is the most powerful conditionally valid MX CI test. Furthermore, we show that under misspecification of the distribution of X given Z, our method retains Type-I error sequentially just as well as the CRT does for blocks of data (Berrett et al. 2020). Finally, we discuss in detail an application to the setting where Y is binary, where we use logistic regression to construct an anytime-valid test of conditional independence. Under the MX assumption, this test is valid even if the logistic model assumptions are not satisfied, and when they are, it is guaranteed to have asymptotic power one.

The rest of this article is structured as follows. In Section 2 we discuss the background needed for this article. This includes an introduction to the conditional randomization test in Section 2.1 and to e-statistics in Section 2.2. In Section 3 we discuss in detail our anytime-valid method of testing conditional independence under MX. This includes an analysis of optimality in Section 3.1, a bound on the worst-case rejection rate in Section 3.2, and an application to binary response variable in Section 3.3. We compare our method to established tests of conditional independence in a simulation study in Section 4. This includes a comparison in terms of Type-I error (Sections 4.1 and supplementary Section S1) and in terms of power (Section 4.2). Our method is further highlighted by an application to a real-world dataset in Section 5. This article is concluded with a discussion of our results in Section 6. All proofs are deferred to the supplementary material.

[1] E-statistics are commonly known as e-variables; we use the former to stress data dependence.

2. Background

In this section, we give a brief introduction to the conditional randomization test, as well as to e-statistics. Throughout this section, as well as the rest of the article, we assume that the data consists of iid tuples $D_i = (X_i, Y_i, Z_i) \in \mathcal{D} := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. In the section on the CRT, we assume that data comes in as a single block $D^n = (D_1, \ldots, D_n)$ of $n \in \mathbb{N}$ data points, so that i above ranges from 1 to n. In contrast, in the section on e-statistics, we assume that the data comes in sequentially as a stream $(D_n)_{n\in\mathbb{N}}$. We use the notation $f_{Y|X,Z}(y \mid x, z)$ to denote the conditional density of Y given X and Z evaluated at (x, y, z) when the joint density of (X, Y, Z) is f, and analogous notation for other conditional densities, for example, $f_{X|Z}(x \mid z)$ for the density of X given Z.

2.1. The Conditional Randomization Test

The conditional randomization test by Candès et al. (2018) works under the MX assumption, that is, that the distribution of $X \mid Z$ is known. This holds, for example, when X corresponds to a randomized treatment in a clinical trial, since the researchers specify the randomization mechanism themselves. Candès et al. (2018) give other examples where the this distribution is at least known to a much higher precision than the joint distribution of (X, Y, Z) because much more samples of pairs (X, Z) are available than data including Y. We let Q_z denote the distribution of X given that $Z = z$. We are interested in testing $X \overset{P}{\perp\!\!\!\perp} Y \mid Z$, which in the MX setting corresponds to the null hypothesis

$$\mathcal{H}_0 = \left\{ P \in \mathcal{P}(\mathcal{D}) : X \overset{P}{\perp\!\!\!\perp} Y \mid Z,\ P_{X|Z} = Q_Z \right\}, \tag{1}$$

where $\mathcal{P}(\mathcal{D})$ denotes the set of all probability distributions on $\mathcal{D}$, and $P_{X|Z}$ is a shorthand for the conditional law of X given Z. The starting point of the CRT is to choose any test statistic $T : \mathcal{D}^n \mapsto \mathbb{R}$ that measures the dependence of X and Y given Z, so that larger values of T indicate stronger dependence. Since the conditional distribution of X given Z is known, one can simulate independent new realizations $\tilde{X}_{j,i} \sim Q_{Z_i}$, $j = 1, \ldots, M$, $i = 1, \ldots, n$, so that under the null hypothesis $\tilde{X}^n_j = (X_{j,1}, \ldots, X_{j,n})$ and X^n have the same distribution conditional on Y^n and Z^n. Hence, the triplets $(\tilde{X}^n_j, Y^n, Z^n)$, $j = 1, \ldots, M$, and (X^n, Y^n, Z^n) are exchangeable, and

$$p_M(X^n, Y^n, Z^n) = \frac{1 + \sum_{j=1}^{M} 1\{T(\tilde{X}^n_j, Y^n, Z^n) \geq T(X^n, Y^n, Z^n)\}}{1 + M} \tag{2}$$

is a "Monte Carlo" p-value for the null hypothesis (1). More precisely, for any distribution $P \in \mathcal{H}_0$,

$$P(p_M(X^n, Y^n, Z^n) \leq \alpha) \leq \alpha.$$

In the case that the distributions Q_z are not known exactly, but only some estimate $\hat{Q}_z$, an obvious question is how robust a test based on $\hat{Q}_z$ may be. Berrett et al. (2020, Theorem 4) show that

$$P(p_M(X^n, Y^n, Z^n) \leq \alpha \mid Y^n, Z^n) \leq \alpha + d_{\mathrm{TV}}(\hat{Q}^n_{Z^n}, Q^n_{Z^n}), \tag{3}$$

where $Q^n_{Z^n}$ is the product of the distributions Q_{Z_i}, $i = 1, \ldots, n$, and d_{TV} the total variation distance. Precise estimation of conditional distributions in total variation distance is admittedly

a challenging problem, but also not completely unrealistic, as discussed among others by Berrett et al. (2020, sec. 5.1). We furthermore give a brief discussion of available literature on estimation in terms of KL divergence at the end of Section 3.1, and a bound on KL gives a bound on total variation.

There remains the question how to choose the statistic T. Katsevich and Ramdas (2022) show that for a point alternative with density f for (X, Y, Z), using the conditional density $f_{Y|X,Z}$ of Y given (X, Z) as statistic T leads to the most powerful conditionally valid test against $\mathcal{H}_0$. This result suggests that a reasonable method is to use an estimator $\hat{f}_{Y|X,Z}$ of the true conditional density as statistic. Importantly, any estimation error does not lead to a loss of Type-I error guarantee, but only to a decrease in power, as the p-value in (2) is valid for *any* statistic. Alternatively, one could use any measure of conditional independence, for example, the absolute value of the fitted coefficient in a lasso regression model, as originally proposed by Candès et al. (2018). A downside to this method is that it is computationally expensive, as it requires a lasso model to be fit for the original data as well as for all the simulated data points. Liu et al. (2021) propose a "leave-one-covariate-out" variant of these lasso statistics, which is significantly less computationally demanding, while leading to similar power.

2.2. E-statistics, Test Martingales, and Anytime-Valid Tests

An e-statistic is any function of the data $S_n : \mathcal{D}^n \to [0, \infty)$ such that $\mathbb{E}_P[S_n(D^n)] \leq 1$ for all $P \in \mathcal{H}_0$. An e-statistic evaluated on a realization of the data will be referred to as an e-value. Previously, e-statistics and e-values have appeared in the literature as likelihood ratios (although the concept is vastly more general than such ratios), particular Bayes factors and betting scores (Shafer 2021). Large e-values constitute evidence against the null hypothesis, since $P(S_n(D^n) \geq 1/\alpha) \leq \alpha$ by Markov's inequality, so that the Type-I error of the test $1\{S_n(D^n) \geq 1/\alpha\}$ is bounded by α. However, such a test is defined for a block of data D^n. In a sequential setting, one instead observes a stream of data $(D_n)_{n\in\mathbb{N}}$. We therefore define, more generally, a sequence of conditional e-statistics as a sequence of statistics $(E_n(D^n))_{n\in\mathbb{N}}$, such that $\mathbb{E}_P[E_n(D^n) \mid D^{n-1}] \leq 1$ for all $n \in \mathbb{N}$ and $P \in \mathcal{H}_0$. For $n = 1$, the expectation is supposed to be read unconditionally. Intuitively, the conditional e-statistic at time n measures the evidence in round n against $\mathcal{H}_0$ conditional on the past data, and the cumulative product of these conditional e-statistics $S_n(D^n) = \prod_{i=1}^n E_i(D^i)$ is a measure of the total accumulated evidence against the null hypothesis. Formally, the sequence $(S_n(D^n))_{n\in\mathbb{N}}$ of cumulative products forms a nonnegative supermartingale with starting value bounded by 1, a so-called test martingale, that is, $\mathbb{E}_P[S_{n+1}(D^{n+1}) \mid D^n] \leq S_n(D^n)$ for all $P \in \mathcal{H}_0$ and $n \in \mathbb{N}$, and $\mathbb{E}_P[S_1(D_1)] \leq 1$. By Ville's inequality (see, e.g., Shafer 2021), such test martingales satisfy for any $\alpha > 0$

$$P(\exists\, n \in \mathbb{N}\colon S_n(D^n) \geq 1/\alpha) \leq \alpha. \tag{4}$$

A sequential test can thus be defined by monitoring $S_n(D^n)$ and rejecting if it exceeds $1/\alpha$. The inequality (4) ensures that this test retains Type-I error control, no matter how we choose the moments to peek at $S_n(D^n)$. In fact, (4) is easily derived from a more basic property: $(S_n(D^n))_{n\in\mathbb{N}}$ satisfies $\mathbb{E}_P[S_\tau(D^\tau)] \leq 1$ for any stopping time τ, that is, the stopped process $S_\tau(D^\tau)$ is again an e-statistic, both for data dependent and externally imposed stopping rules (Grünwald, de Heide, and Koolen 2019, Proposition 2). Tests with this property will be referred to as anytime-valid tests.

Perhaps the most prominent example of a test martingale is the likelihood ratio process $L_n = \prod_{i=1}^n p_1(Y_i)/p_0(Y_i)$ for testing the null hypothesis that independent $(Y_n)_{n\in\mathbb{N}}$ stem from a distribution with density p_0 against the alternative density p_1. Tests based on likelihood ratio processes $(L_n)_{n\in\mathbb{N}}$ are called sequential probability ratio test (SPRT) and have first been studied by Wald (1947), though Wald, like most of the subsequent sequential analysis literature, uses them with reject/accept rules different from ours (e.g., one rejects if the value exceeds $(1-\beta)/\alpha$ instead of $1/\alpha$) that preclude optional stopping and require specific stopping times/rules. We refer to Grünwald, de Heide, and Koolen (2019, sec. 7) for further discussion of related literature on sequential testing.

Under violations of the null hypothesis, one would hope that an e-statistic or test martingale attains high values, which gives evidence to reject the null hypothesis. This requires a suitable analogue of power, or notion of optimality, for e-statistics. We follow Grünwald, de Heide, and Koolen (2019) and Shafer (2021) and try to find e-statistics that maximize the logarithmic expected value under the alternative. That is, suppose the alternative hypothesis is given by a single distribution $\mathcal{H}_1 = \{P^*\}$, then the growth-rate optimal (GRO) e-statistic is defined as the e-statistic that maximizes $S_n \mapsto \mathbb{E}_{P^*}[\log S_n(D^n)]$ over all e-statistics. At first glance, this notion of optimality seems counterintuitive for sequential tests, since it is defined for individual e-statistics and not test martingales. Really, one would hope to find a test martingale $(S_n)_{n\in\mathbb{N}}$ that maximizes $\mathbb{E}_{P^*}[\log S_\tau(D^\tau)]$ simultaneously for all stopping times τ. Remarkably, in the special setting of this article, the sequence of GRO statistics that we will consider for a point alternative does have exactly this property, as follows from Theorem 12 of Koolen and Grünwald (2022), providing additional justification for our focus on GRO. This is discussed briefly in the supplementary material (Section S3).

3. Conditional Independence Testing With *E*-statistics

The conditional randomization test and its permutation version by Berrett et al. (2020) are defined for batches of data D^n. We show here how to create e-statistics for sequential testing with a stream of data $(D_n)_{n\in\mathbb{N}}$ under the MX assumption. Our method can be seen as a broad generalization of an e-statistic introduced in the proof of the main theorem of Turner, Ly, and Grünwald (2021), who essentially handle the case that Y and X are both Bernoulli.

Theorem 1. Let $h_n : \mathcal{D} \to [0, \infty)$, $n \in \mathbb{N}$, be nonnegative measurable functions such that h_n is determined after seeing data D^{n-1}. Then the sequence $(E^{\mathrm{CI}}_{h_n}(D_n))_{n\in\mathbb{N}}$ defined by

$$E^{\mathrm{CI}}_{h_n}(D_n) = \frac{h_n(X_n, Y_n, Z_n)}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n)\, \mathrm{d}Q_{Z_n}(x)} \tag{5}$$

is a sequence of conditional e-statistics for the null hypothesis (1). Consequently, $S^{\mathrm{CI}}_{h^n}(D^n) = \prod_{i=1}^n E^{\mathrm{CI}}_{h_i}(D_i)$ is a test martingale.

To be explicit, the general workflow of our method is as follows. At each time $n = 1, 2, \ldots$, a test function h_n is chosen, usually depending on the past data D^{n-1}. After seeing the data point D_n, the conditional e-statistic (5) is computed and the cumulative product is updated according to $S^{\mathrm{CI}}_{h^n}(D^n) = S^{\mathrm{CI}}_{h^{n-1}}(D^{n-1}) \cdot E^{\mathrm{CI}}_{h_n}(D_n)$. For a test at level α, one could stop and reject the null hypothesis as soon as $S^{\mathrm{CI}}_{h^n}(D^n) \geq 1/\alpha$ for the first time.

Remark. At this point, it should be noted that the MX assumption is stronger than needed within our sequential setup: the MX assumption requires that the data points (X_n, Y_n, Z_n) are iid and that the distribution of X_n given Z_n is given by Q_{Z_n}. However, in our sequential setting it would actually be allowed for the distribution of (X_n, Y_n, Z_n) to depend on the data D^{n-1}. At each time n, we would use in the denominator a distribution $Q_{Z_n, D^{n-1}}$. It should be clear that the resulting sequence of random variables still defines a test martingale, but with respect to the altered null hypothesis that $Y_n \perp\!\!\!\perp X_n \mid Z_n, D^{n-1}$ and $X_n \mid Z_n \sim Q_{Z_n, D^{n-1}}$ for all $n \in \mathbb{N}$. Clearly, the assumption that $Q_{Z_n, D^{n-1}}$ is known (or can be estimated with high precision) is not realistic in many settings with temporal dependence. One example where this extension would be useful are clinical trials with so-called covariate- or response-adaptive designs, where the allocation of patients to a treatment depends either on the imbalance of treatment/control in certain covariate groups or on previously observed responses from patients (Robbins 1952; Pocock and Simon 1975; Zhang et al. 2007). To avoid cluttering notation, we will not explicitly denote this potential past data dependence, but it is good to keep in mind.

In most cases in practice, one has to resort to simulations to approximate the integral in the denominator of (5). That is, one simulates M independent $\tilde{X}_1, \ldots, \tilde{X}_M$ according to Q_{Z_n} and replaces the integral by the empirical average $\sum_{i=1}^M h_n(\tilde{X}_i, Y_n, Z_n)/M$. The following proposition shows that a slight modification of this procedure is guaranteed to yield an e-statistic.

Proposition 2. Let $\tilde{X}_1, \ldots, \tilde{X}_M$ be independent with distribution Q_{Z_n}. Then

$$\check{E}^{\mathrm{CI}}_{h_n}(D_n) := \frac{h_n(X_n, Y_n, Z_n)}{(h_n(X_n, Y_n, Z_n) + \sum_{i=1}^M h_n(\tilde{X}_i, Y_n, Z_n))/(M+1)}$$

satisfies $\mathbb{E}_P[\check{E}^{\mathrm{CI}}_{h_n}(D_n) \mid D^{n-1}] = 1$, $P \in \mathcal{H}_0$, where h_n is as in Theorem 1 and the expectation is taken both over the data and $\tilde{X}_1, \ldots, \tilde{X}_M$.

In fact, the proof of Proposition 2 only requires that $X_n, \tilde{X}_1, \ldots, \tilde{X}_M$ are exchangeable, so it will also be applicable in many situations where data is randomly permuted. Furthermore, the naive approach without including $h_n(X_n, Y_n, Z_n)$ in the denominator is anti-conservative and does not define a sequence of conditional e-statistics. Indeed, taking expectation over $X_n, \tilde{X}_1, \ldots, \tilde{X}_M$,

$$\mathbb{E}_P\left[\frac{h_n(X_n, Y_n, Z_n)}{\sum_{i=1}^M h_n(\tilde{X}_i, Y_n, Z_n)/M} \middle| Y_n, Z_n\right] \geq \frac{\int h_n(x, Y_n, Z_n)\, dQ_{Z_n}(x)}{\sum_{i=1}^M \int h_n(x, Y_n, Z_n)\, dQ_{Z_n}(x)/M} = 1. \quad (6)$$

Here we use that $X_n, \tilde{X}_1, \ldots, \tilde{X}_M$ are independent with distribution Q_{Z_n}, and invoke Jensen's inequality with the strictly convex function $s \mapsto 1/s$. Equality in (6) holds if any only if h_n is constant in x, and in that trivial case the e-statistic is equal to the constant 1. In all our applications, we use the variant proposed in Proposition 2 to compute the e-statistics.

3.1. Optimality

In order to accumulate evidence against the null hypothesis, the functions h_i in (5) should measure the conditional (in)dependence between X and Y. This will ensure that the test martingale $(S^{\mathrm{CI}}_{h^n}(D^n))_{n\in\mathbb{N}}$ will grow if the null hypothesis is violated. Ideally, we would be able to choose a measure of conditional independence that requires little to no assumptions on the distribution of (X, Y, Z). Many such measures have been proposed in the literature, see, for example, Fukumizu et al. (2007), Shah and Peters (2020), and Azadkia and Chatterjee (2021). However, as far as we can tell, none of these measures allow for a sequential decomposition. That is, they are defined for fixed sample size n as a function $T_n : \mathcal{D}^n \to [0, \infty)$, but cannot be decomposed in a nontrivial way as a product $T_n(D^n) = \prod_{i=1}^n T_i(D_i)$. Our only option to use them in our test martingale is therefore to set $h_i(D_i) = T_i(D^{i-1}, D_i)$ in (5). This would have two major drawbacks: first, it would be computationally involved, because T_i needs to be recalculated entirely *within* the integral in the denominators in (5). Second, it would generally be ineffective, because $T_i(D^{i-1}, D_i)$ will depend very little on D_i for large i, so $h_i(D_i)$ will generally not be sensitive to changing X_i to $x' \sim Q_{Z_i}$. As a result, the fraction in (5) will be close to 1, preventing us from accumulating much evidence against the null.

However, if we are willing to assume that under the alternative, the density of (X, Y, Z) is given by f, then the conditional density $f_{Y|X,Z}$ itself is a measure of conditional independence. Moreover, evaluating the density in n data points is equivalent to taking the product of the density evaluated at all single data points $i = 1, \ldots, n$, because the data stream is iid. This gives a sequential decomposition as desired above. Furthermore, Katsevich and Ramdas (2022) showed that an optimal conditionally valid p-value based test can be achieved by running the CRT with the conditional density $f_{Y|X,Z}$ as test function. It turns out that this choice also yields the GRO e-statistic among all e-statistics defined on n observations, and the expectation of the logarithm of this e-statistic is the conditional mutual information (Cover and Thomas 1991), an established conditional dependence measure which has also been applied for conditional independence testing (Runge 2018). Note that the expectation is taken over the entirety of the data, that is, $D^n = (X^n, Y^n, Z^n)$, as opposed to the result by Katsevich and Ramdas (2022) which only holds conditionally on (Y^n, Z^n); see their article for a more thorough discussion.

Theorem 3. The GRO e-statistic for testing $\mathcal{H}_0$ as in (1) against the alternative distribution with density f is given by

$$S^{\mathrm{CI}}_{f_{Y|X,Z}}(D^n) = \prod_{i=1}^{n} E^{\mathrm{CI}}_{f_{Y|X,Z}}(D_i) = \prod_{i=1}^{n} \frac{f_{Y|X,Z}(Y_i \mid X_i, Z_i)}{f_{Y|Z}(Y_i \mid Z_i)} \tag{7}$$

and achieves growth rate $\mathbb{E}_f[\log S^{\mathrm{CI}}_{f_{Y|X,Z}}(D^n)] = nI_f(X;Y \mid Z)$, where $I_f(X;Y \mid Z)$ denotes the conditional mutual information if (X,Y,Z) follows the distribution with density f.

Remark. A simple application of Bayes theorem allows one to rewrite (7) to $S^{\mathrm{CI}}_{f_{Y|X,Z}}(D^n) = \prod_{i=1}^{n} f_{X|Y,Z}(X_i \mid Y_i, Z_i)/f_{X|Z}(X_i \mid Z_i)$, which shows that the resulting test martingale is in fact a likelihood ratio process. That is, it is the ratio between the true density of X given (Y,Z) under the alternative and that under the null (whereas the density in the denominator of (7) does not have to correspond to the data generating distribution). The latter follows, because under the null $f_{X|Y,Z}$ is equal to $f_{X|Z}$, which we assume to be well-specified and equal the density of Q_{Z_i}. Hence, the resulting test for a simple alternative hypothesis f is in fact a generalization of the SPRT, where the distribution under the null hypothesis changes with the variable Z_i. It depends on the application at hand which formulation, that is, (7) or conditional densities of X, is more suitable for constructing a test. For example, we show in Section 3.3 that for binary $Y \in \{0,1\}$ and $(X,Z) \in \mathbb{R}^p \times \mathbb{R}^q$, one can construct a test based on logistic regression, which is often simpler than directly trying to find a suitable conditional density of X given Y and Z, especially when $p > 1$.

The information inequality (Cover and Thomas 1991) implies that $I_f(X;Y \mid Z) \geq 0$, with equality if and only if $Y \perp\!\!\!\perp X \mid Z$, which shows that $S^{\mathrm{CI}}_{f_{Y|X,Z}}$ has nontrivial power to detect deviations from conditional independence if f is the true density of (X,Y,Z). Assuming a simple (point) alternative f is, of course, an unrealistically strong assumption. We now proceed to develop a method that also gets uniform growth rates for potentially large classes of alternative densities $\mathcal{F}$ by building on the simple alternative case. The following result states that if we do not know the density $f_{Y|X,Z}$, but instead use a different density $g_{Y|X,Z}$, the loss in expected growth rate is directly related to a measure of distance between $f_{Y|X,Z}$ and $g_{Y|X,Z}$.

Proposition 4. For any conditional density $g_{Y|X,Z}$, the following holds:

$$\mathbb{E}_f\left[\log E^{\mathrm{CI}}_{g_{Y|X,Z}}(D)\right] \geq I_f(X;Y \mid Z) - \mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z} \| g_{Y|X,Z})]. \tag{8}$$

Since we are in a sequential setting, this proposition implies that if we do not know the density f, we could try to estimate it, using estimates that improve as sample size increases. That is, let $\hat{f}_n$ be an estimator of $f_{Y|X,Z}$ based on data D^n, and $\hat{f}_0$ an initial guess. The test martingale we use is then given by $\prod_{i=1}^{n} E^{\mathrm{CI}}_{\hat{f}_{n-1}}(D^i)$. It follows from the combination of Theorem 3 and Proposition 4 that if the estimator is consistent in a KL sense, then the expected growth per outcome converges to that of the GRO e-statistic.

Corollary 5. (i) Assume that $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z} \| \hat{f}_{i-1}) \mid D^{i-1}] \xrightarrow[n\to\infty]{\text{a.s.}} 0$, then

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_f\left[\log E^{\mathrm{CI}}_{\hat{f}_{i-1}}(D_i) \middle| D^{i-1}\right] \xrightarrow[n\to\infty]{\text{a.s.}} I_f(Y;X \mid Z).$$

(ii) Assume that for some function $b(n) : \mathbb{N} \to \mathbb{R}_0^+$ with $b(n) = o(n)$, we have

$$\mathbb{E}_f\left[\sum_{i=1}^{n} \mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z} \| \hat{f}_{i-1}) \mid D^{i-1}]\right] \leq b(n). \tag{9}$$

Then

$$\frac{1}{n}\mathbb{E}_f\left[\sum_{i=1}^{n} \log E^{\mathrm{CI}}_{\hat{f}_{i-1}}(D_i)\right] \geq I_f(X;Y \mid Z) - \frac{b(n)}{n}. \tag{10}$$

Consequently, to achieve an asymptotic optimal growth rate, we need to use a conditional density estimator $\hat{f}_i$ that converges in KL divergence to f, where we may assume $f \in \mathcal{F}$ for some given set of densities $\mathcal{F}$, that is, our statistical model. Barron (1998) showed that, for a wide variety of parametric and nonparametric models $\mathcal{F}$, we have *convergence in information* (his terminology for (9) holding for all n) if we set $\hat{f}_i$ to be the Bayes predictive density, under no further conditions, uniformly for all $f \in \mathcal{F}$, as long as a suitable prior is used. This means that, for some fixed function b (9) holds uniformly for all $f \in \mathcal{F}$, so that also (10) holds uniformly for all $f \in \mathcal{F}$: we could put a $\inf_{f\in\mathcal{F}}$ to the left of (10) and the result would still hold. Thus, we get the optimal growth rate (which itself can only be achieved by an oracle that knows the "true" $f \in \mathcal{F}$) up to an additive term of $b(n)/n$ uniformly for all $f \in \mathcal{F}$. The rate $b(n)/n$ is then usually, up to log factors, equal to the minimax rate in squared Hellinger distance. The same rates (potentially up to further log factors) are available for the Bayesian posterior mean under a weak additional condition on the model introduced by Grünwald and Mehta (2020) under the name *witness-of-badness*; it generalizes a well-known earlier condition of Wong and Shen (1995); see also (Bilodeau, Foster, and Roy 2021) for related results. In Section 3.3, we demonstrate our approach with $\mathcal{F}$ set to the logistic regression model with $X \in \mathbb{R}^p$ and $Z \in \mathbb{R}^q$, for which a result by Foster et al. (2018) in combination with (Barron 1998) implies that, if we use the Bayes predictive distribution as above, then $b(n)$ can be chosen as $O((p+q)\log n)$ as long as the first four moments of all components of X and Z exist, implying a parametric rate. However, in our experiments in Section 4, rather than the Bayes predictive distribution, we use a (slightly regularized) MLE, since it can be computed much more efficiently. Although we suspect that the MLE converges at the same rates as Bayesian methods under the same weak conditions, the methods of the aforementioned papers cannot be used to prove this, and instead in Proposition 7 we show almost sure convergence of the MLE, without rates, under a stronger sub-Gaussianity assumption on (X,Z).

3.2. Worst-Case Bounds on Rejection Rate

Up to now, we have discussed the construction and properties of e-statistics when the conditional distributions Q_z are known

exactly. In this section, we prove a result analogous to Theorem 4 of Berrett et al. (2020) (see (3)) on the error rate of our sequential test under the null hypothesis when the distributions Q_z are only approximations. The approximation of Q_z will be denoted by $\hat{Q}_z$, and the (approximate) e-statistic at time n is given by

$$\tilde{E}^{\mathrm{CI}}_{h_n}(D^n) = \frac{h_n(X_n, Y_n, Z_n \mid D^{n-1})}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n \mid D^{n-1})\, \mathrm{d}\hat{Q}_{Z_n}(x)}.$$

Here the nonnegative function h_n depends on D^{n-1} since it can be constructed sequentially, for example by estimating the conditional density of Y_n given X_n and Z_n with all past data (X_i, Y_i, Z_i), $i = 1, \ldots, n-1$. Recall that $Q^n_{Z^n}$ denotes the product distribution of Q_{Z_i}, $i = 1, \ldots, n$, that is, for measurable $A \subseteq \mathcal{X}^n$

$$Q^n_{Z^n}(A) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} 1\{x^n \in A\}\, \mathrm{d}Q_{Z_1}(x_1) \cdots \mathrm{d}Q_{Z_n}(x_n)$$

In particular, $Q^n_{Z^n}(A) = P(X^n \in A \mid Z^n) = P(X^n \in A \mid Y^n, Z^n)$ for $P \in \mathcal{H}_0$, due to conditional independence of Y_i and X_i given Z_i, $i = 1, \ldots, n$. The distribution $\hat{Q}^n_{Z^n}$ is defined in the same way as $Q^n_{Z^n}$ but with $\hat{Q}_{Z_i}$ instead of Q_{Z_i}.

Theorem 6. Assume that $h_1, \ldots, h_N > 0$ are measurable. For any $N \in \mathbb{N}$, $\alpha \in (0, 1)$ and $P \in \mathcal{H}_0$,

$$P\left(\exists\, n \le N\colon \prod_{i=1}^{n} \tilde{E}^{\mathrm{CI}}_{h_i}(D^i) \ge \frac{1}{\alpha} \,\middle|\, Y^N, Z^N\right) \le \alpha + d_{\mathrm{TV}}(Q^N_{Z^N}, \hat{Q}^N_{Z^N}). \tag{11}$$

Theorem 6 gives the same worst case bound on the rejection rate as Theorem 4 in Berrett et al. (2020) for a sample size N, but optional stopping at any sample size $n \le N$ is allowed. Berrett et al. (2020, sec. 5.1) discuss conditions under which the total variation distance between $Q^N_{Z^N}$ and $\hat{Q}^N_{Z^N}$, which bounds the excess rejection rate, is small. For example, they obtain an upper bound of the form $\mathcal{O}_p(\sqrt{Nk/m})$ when $(X, Z) \in \mathbb{R}^k$ follow a multivariate Gaussian distribution and the conditional law of $X \mid Z$ is estimated with an unlabeled sample of size m. Hence, when m remains constant but N diverges to infinity, the bound on the distance between $Q^N_{Z^N}$ and $\hat{Q}^N_{Z^N}$ becomes trivial. Furthermore, the discussion at the end of Section 3.1 on estimation in terms of KL divergence (which bounds the total variation distance) can also be applied to the estimation of Q_z.

3.3. Application to Logistic Regression

The general construction strategy for e-statistics and all results so far assume no specific model for the outcome Y or covariates (X, Z). In this section, we consider the special but important case of a binary outcome $Y \in \{0, 1\}$ which follows a logistic regression model under the alternative, that is, $(X, Z) \in \mathbb{R}^{p+q}$, and Y equals $y = 0, 1$ with probability

$$p_\theta(y \mid X, Z) = \frac{\exp(y(\beta^\top X + \gamma^\top Z))}{1 + \exp(\beta^\top X + \gamma^\top Z)}, \tag{12}$$

with an unknown $(p+q)$-dimensional coefficient vector $\theta = (\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)$. Conditional independence of Y and X given Z holds if and only if $\beta_1 = \cdots = \beta_p = 0$. It turns out that in this setting, one can construct e-statistics which not only have power on average, as in Corollary 5, but which even reject the null hypothesis with probability one if it is violated and the sample size grows to infinity. At this point, it is important to recall that the validity of the e-statistics does not require the logistic model to be correctly specified: the validity only depends on the null hypothesis, which is still the set of *all* distributions under which conditional independence holds in the sense of (1), including many distributions that violate the logistic model assumption. The result rather shows that *if* the logistic model is suitable (in the sense that, if the alternative is true, then data are sampled from a distribution in this model), then the e-statistic has guaranteed power to detect violations of conditional independence.

Following our general strategy, an e-statistic for testing CI is given by

$$S^{CI}_n(D^n) = \prod_{i=1}^{n} \frac{p_{\hat{\theta}_{i-1}}(Y_i \mid X_i, Z_i)}{\int p_{\hat{\theta}_{i-1}}(Y_i \mid x, Z_i)\, \mathrm{d}Q_{Z_i}(x)}, \tag{13}$$

where $\hat{\theta}_k$ may be any estimator for θ based on the first k samples, (X_i, Y_i, Z_i), $i = 1, \ldots, k$. When the observations (X_i, Y_i, Z_i), $i \in \mathbb{N}$, are independent and identically distributed, the growth rate optimal e-statistic is obtained if $\hat{\theta}_k = \theta$ for all $k \in \mathbb{N}$. Nevertheless, the following proposition shows that tests based on S^{CI}_n have asymptotic power one when $\hat{\theta}_k$ is the maximum likelihood estimator. From now on, $\|v\| = (v^\top v)^{1/2}$ denotes the Euclidean norm of a vector v, and we denote by (X, Z) the stacked vector $(X_1, \ldots, X_p, Z_1, \ldots, Z_q)$. We will take (X, Y, Z) as a generic observation which has the same distribution as (X_i, Y_i, Z_i), $i \in \mathbb{N}$, for writing probability statements about elements of this sequence.

Proposition 7. Let (X_i, Y_i, Z_i), $i \in \mathbb{N}$, be independent and identically distributed such that (12) holds with $(\beta_1, \ldots, \beta_p) \neq 0$. Assume furthermore that

(i) (a) (X, Z) satisfies $P(u^\top (X, Z) \neq 0) > 0$ for all $u \in \mathbb{R}^{p+q} \setminus \{0\}$, and (b) it is sub-Gaussian with variance parameter σ^2, that is

$$\mathbb{E}[\exp(u^\top((X, Z) - \mathbb{E}[(X, Z)]))] \le \exp(\|u\|^2 \sigma^2/2), \quad \forall\, u \in \mathbb{R}^{p+q},$$

(ii) $\hat{\theta}_n$ in (13) is the logistic MLE based on data (X_i, Y_i, Z_i), $i = 1, \ldots, n$, for all $n \in \mathbb{N}$, with $\hat{\theta}_n$ arbitrarily defined but finite if the MLE does not exist.

Then S^{CI}_n satisfies $\liminf_{n\to\infty} \log(S^{CI}_n)/n \ge I(X; Y|Z) > 0$ almost surely.

Assumption (i)(a) ensures that the MLE converges almost surely at a fast rate, as shown by Qian and Field (2002). Instead of sub-Gaussianity ((i)(b)) their result requires only moment assumptions (which are implied by sub-Gaussianity), but sub-Gaussianity is indeed required in our setting: see the proof of Proposition 7, given in the supplementary material.

4. Simulations

To investigate the robustness and power of tests of conditional independence based on *e*-statistics, we compare the *e*-statistic (13) for binary Y to other methods applicable in this setting. The covariate X is univariate, $X \in \mathbb{R}$, while $Z = (1, Z_1, \ldots, Z_{q-1})$ is a q-dimensional vector containing an intercept term. The distribution of $(X, Z_1, \ldots, Z_{q-1})$ is the q-dimensional normal distribution with zero mean and a Toeplitz covariance matrix, $\Sigma_{i,j} = 1/(1+|i-j|)$ for $i, j = 1, \ldots, q$. Then Q_Z is the Gaussian distribution with mean

$$\mu_Z = \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}(Z_1, \ldots, Z_{q-1})^\top \tag{14}$$

where $\Sigma_{-1,-1}$ is the submatrix $(\Sigma_{i,j})_{i,j=2,\ldots,q}$ and $\Sigma_{1,-1}$ is the row vector $(\Sigma_{1,2}, \ldots, \Sigma_{1,q})$. The conditional variance equals $\sigma_Z^2 = 1 - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{1,-1}^\top$. The binary response Y has probabilities given by $p_\theta(y \mid X, Z)$, $y \in \{0, 1\}$, as in (12), where the intercept and the coefficients of Z, and $\gamma_1, \ldots, \gamma_q$, are drawn independently and uniformly distributed on the interval $[-1, 1]$. The coefficient of X, that is, $\beta = \beta_1$, is chosen in $[0, 1]$, with 0 corresponding to conditional independence of Y and X given Z. Below are implementation details for all methods considered in the simulations. In the following sections, these methods are compared in terms of Type-I error and power. A further study on the robustness of the MX-based methods with respect to violations of the MX assumption is given in the supplementary material.

Conditional randomization e-statistic (E-CRT). The parameter vector θ in (13) is re-estimated after each new observation with the maximum likelihood method, starting from a minimal sample size of $5q + 1$, so that $5q$ observations are available for the first parameter estimate. In addition, the probabilities $p_{\hat{\theta}_k}(y \mid X, Z)$ are truncated to $[\varepsilon, 1-\varepsilon]$ for some small $\varepsilon > 0$. This is to account for the fact that at small sample sizes the MLE sometimes yields predicted probabilities in $\{0, 1\}$. We also include an oracle version of this *e*-statistic (E-CRT-O), which uses the true θ starting from the first observation, instead of the maximum likelihood estimator. For both variants, the integral in the *e*-statistic is approximated by an average over 500 Monte Carlo samples.

Conditional randomization test (CRT). The CRT is applied nonsequentially with the likelihood of the logistic regression model as test statistic and 500 samples for randomization. That is, X is sampled 500 times from the conditional distribution given Z, the logistic regression model is re-estimated with this simulated covariate, and the likelihood achieved with these models with simulated X is compared to the likelihood achieved with the actual values of X.

The following methods are for testing whether the coefficient β in the logistic regression model equals zero. Unlike the E-CRT and CRT, they are not based on the MX assumption, but their Type-I error guarantee does require that the true probabilities of Y are given by the logistic model (12). A comparison is of interest since these methods are, to the best of our knowledge, currently the only ones that allow sequential testing in a logistic model.

Running maximum likelihood (R-MLE). We apply the running MLE method by Wasserman, Ramdas, and Balakrishnan (2020, sec. 7), an instance of the generic method introduced in that paper which they call *universal inference*. Parameter estimation is also started with a minimum $5q$ observations, and we additionally investigate an L_1 penalized version for estimation under the alternative hypothesis, abbreviated as R-MLE-P, which, like for the E-CRT, is to prevent predicted probabilities close to 0 or 1 due to divergence of the MLE. In this second variant, the penalization parameter is chosen by 10-fold cross-validation on the available data at the given time, with likelihood as optimization criterion. The penalization parameter is only updated every 10 observations, since the cross-validation is computationally expensive.

Likelihood ratio test (LRT). The classical asymptotic likelihood ratio test for the null hypothesis that $\beta = 0$ is applied with fixed sample size, and group sequential versions of it with K equally sized groups and the methods by Pocock (1977) (LRT-PK) and O'Brien and Fleming (1979) (LRT-OF).

The results shown in this section are for dimension $q = 4$ of the covariate vector (X, Z). A maximum sample size of $n = 2000$ is considered, after which the evaluation is terminated independently of whether the null hypothesis is rejected. The sequential methods apply the most aggressive stopping rule, namely, reject the null hypothesis as soon as the test statistic exceeds $1/\alpha$ once; more discussion on this is provided in Section 4.2. All results, that is, rejection rates and average sample sizes, are computed over 800 simulations of the data-generating process. The same simulation but with higher dimension ($q = 8$) or with negative correlations between the covariates ($\Sigma_{i,j} = (-1)^{i-j}/(1+|i-j|)$) yields similar results. For the running MLE method, we additionally tested whether not penalizing the coefficient of interest, β, may achieve higher power, which was not the case.

Simulations are performed in R 4.2 (R Core Team 2022), with the `glm` function for maximum likelihood estimation in logistic regression, the package `glmnet` (Simon et al. 2011) for L_1 penalized estimation, and the package `ldbounds` (Casper, Cook, and on FORTRAN program ld98 2022) for computing critical values for the Pocock and O'Brien-Fleming group sequential tests. Replication material for the simulations and the case study, as well as additional figures, are available on *https://github.com/AlexanderHenzi/eindependence*.

4.1. Sequential Tests Under the Null

Table 1 shows the rejection rates of the different sequential methods with a maximum sample size of $n = 2000$. The methods based on *e*-statistics and the running MLE yield rejection rates below the nominal level. For the *e*-statistics, the chosen truncation level is $\varepsilon = 0.05$, but the rejection rate is also below α for $\varepsilon = 0, 0.01, 0.1$. The group sequential methods with $K = 20$ equally sized groups, each of size 100, are slightly anti-conservative, and similar rejection rates are obtained for $K = 5, 10, 40$.

4.2. Simulations Under the Alternative

We proceed to compare the different methods under violations of the null. This is commonly done by comparing the achieved power at given sample sizes n for different effect sizes β, or

Table 1. Rejection frequencies (and standard errors) of the different methods, with implementation details as described in Section 4.1.

	E-CRT	R-MLE	R-MLE-P	LRT-PK	LRT-OF
$\alpha = 0.01$	0.0075 (0.0031)	0.0038 (0.0022)	0.0038 (0.0022)	0.0138 (0.0013)	0.0127 (0.0040)
$\alpha = 0.05$	0.0438 (0.0072)	0.0038 (0.0022)	0.0038 (0.0022)	0.0700 (0.0090)	0.0599 (0.0084)

NOTE: Frequencies are given in [0, 1], not in percentages.

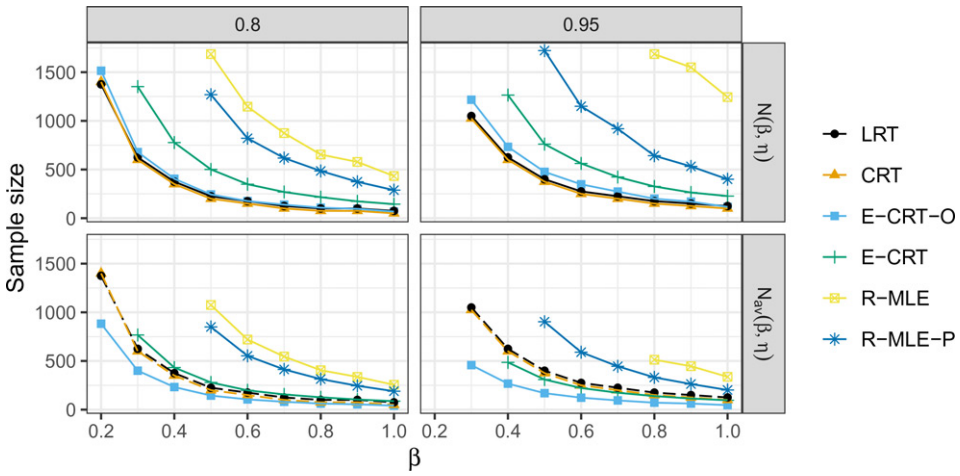

Figure 1. Sample sizes to achieve power $1 - \eta = 0.8, 0.95$ with Type-I error 0.05. For the nonsequential methods (LRT, CRT), dashed lines show $N(\beta, \eta)$ also in the lower panels for better comparison. Simulations were conducted up to $n = 2000$, and no results are shown if $N(\beta, \eta) > 2000$ for a given method, β and η, that is, if more than 2000 observations would be required to achieve a power of $1 - \eta$.

the inverse of that function, that is, the minimum sample size required to achieve power $1 - \eta$,

$$N(\beta, \eta) = \min\{n \in \mathbb{N}\colon P_\beta(\phi_n = 1) \geq 1 - \eta\}.$$

For a fixed Type-I error probability α, the test decision ϕ_n is given by $\phi_n = 1\{\max_{m \leq n} S_m \geq 1/\alpha\}$ for anytime-valid tests based on $(S_n)_{n \in \mathbb{N}}$, or $\phi_n = 1\{p_n \leq \alpha\}$ for a method based on a fixed sample size p-value p_n. For an anytime-valid test, $N(\beta, \eta)$ can be regarded as the worst case sample size a researcher has to plan for in order to achieve power $1 - \eta$; the actual sample size at rejection may be smaller thanks to optional stopping. Therefore, we additionally consider the average sample size of the anytime-valid tests when evaluation is terminated at the latest at $N(\beta, \eta)$,

$$N_{\text{av}}(\beta, \eta) = \mathbb{E}_{P_\beta}[\min(N(\beta, \eta), \inf\{n \in \mathbb{N}\colon S_n \geq 1/\alpha\})].$$

The rationale is that—even though we have seen that anytime-valid methods retain Type-I error validity under arbitrary stopping times—in practice, a natural way to proceed with an anytime-valid test is to run the experiment until either a rejection or a given upper bound on the number of samples is reached. The obvious choice for this upper bound is $N(\beta, \eta)$, as it ensures that the test will have a power of $1 - \eta$. Then $N_{\text{av}}(\beta, \eta)$ gives the average sample size of anytime-valid tests designed to have power $1 - \eta$.

A comparison of the different methods in terms of $N(\beta, \eta)$ and $N_{\text{av}}(\beta, \eta)$ is given in Figure 1. The group sequential methods are excluded from this figure and are analyzed in more detail at the end of the section. The upper two panels of Figure 1 depict $N(\beta, \eta)$ for $1 - \eta = 0.8, 0.95$ as a function of the parameter β (for clarity note that, as before, β stands for the parameter vector in (13); we never use $1 - \beta$ for power). It can be seen that $N(\beta, \eta)$ is higher for the anytime-valid methods than for fixed sample size tests. This is to be expected: the sample size $N(\beta, \eta)$ ensures power $1 - \eta$, but the actual sample size of an anytime-valid test is random and often smaller thanks to early stopping. The lower two panels of Figure 1 similarly show $N_{\text{av}}(\beta, \eta)$ as a function of β. It can be seen that the average sample size is not more and sometimes even less than the sample size of the nonsequential tests. These results suggest that the average sample size to reject the null hypothesis with the E-CRT is not higher than the fixed sample size one would have to plan for with the CRT or LRT. Similar observations have already been made for e-statistics in other settings, such as comparisons with the t-test (Grünwald, de Heide, and Koolen 2019), Fisher's Exact Test (Turner, Ly, and Grünwald 2021), or the logrank test (Ter Schure et al. 2020).

Furthermore, the running MLE method requires much more data to achieve a rejection than the other methods, even with the superior penalized estimation under the null hypothesis. The large sample sizes required for the R-MLE to have a power of 0.95 are mainly due to predicted probabilities close or equal to zero or one at early stages, a problem which is remedied by using penalization, as already proposed by Wasserman, Ramdas, and Balakrishnan (2020); even then, more data are required though. We again emphasize that the running MLE methods are based on different assumptions than the randomization based tests: they do not require the MX assumption, but are only valid for a correctly specified logistic model. However, even when compared to the classical likelihood ratio test, which requires almost the same sample sizes as the CRT, one would have to plan for substantially higher sample sizes with the running MLE methods.

For the conditional randomization e-statistics, we tested different levels of truncation $[\varepsilon, 1 - \varepsilon]$ for the predicted probabilities. Truncating at a small level $\varepsilon = 0.01, 0.05$ is superior to no truncation, $\varepsilon = 0$, since it prevents e-values close to

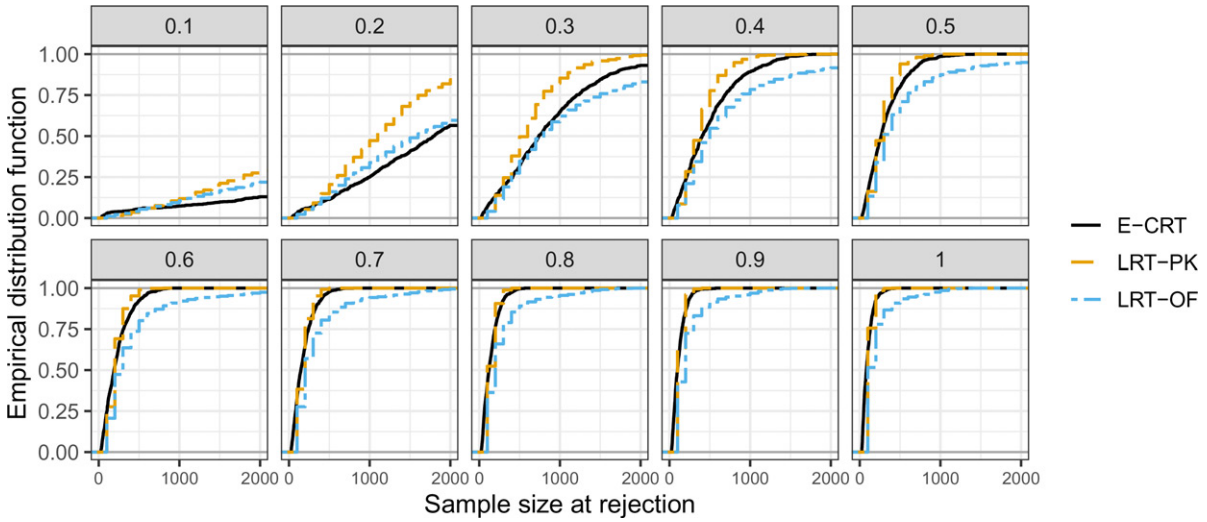

Figure 2. Empirical distribution of the sample size at rejection, with level $\alpha = 0.05$, for the randomization based *e*-statistics and group sequential methods, and $\beta \in \{0.1, 0.2, \ldots, 1\}$.

or equal to zero, but if the truncation level becomes too high, $\varepsilon = 0.1$, it limits the power of the test for observations with predicted probability close to zero or one. See Figure S1 in the supplementary material for the case $\varepsilon = 0$. The results shown in this and the next section are for $\varepsilon = 0.05$. In principle, one could also apply a penalized estimator to remedy convergence problems of the MLE, but truncation is computationally less demanding as it does not require the selection of a penalization parameter.

Finally, in Figure 2, the conditional randomization *e*-statistic is compared to group sequential methods with the Pocock and the O'Brien and Fleming method with $K = 20$ groups. We see that for small parameter $\beta \in \{0.1, 0.2\}$ these methods achieve a higher power than the *e*-statistics, but as already shown in Table 1, the group sequential methods do not control the rejection rate below the nominal level when $\beta = 0$. Also, the E-CRT yields slightly more rejections at small sample sizes than the group sequential methods. As β increases, the conditional randomization *e*-statistics tend to outperform the O'Brien and Fleming method, and achieve a rejection rate very close to Pocock's method. Even for large β, the O'Brien and Fleming requires higher sample sizes due to the fact that the method is designed to yield fewer rejections with small samples. Different numbers of groups for the group sequential methods give similar results, except for the fact that rejecting at small sample sizes becomes impossible if the number of groups is small and the group size large. The performance of the *e*-statistics compared to the group sequential methods is in line with the results of Ter Schure et al. (2020) for survival analysis. Also in their study, group sequential methods and alpha-spending approaches, which have to stop at a certain maximum sample size, tend to achieve a higher power than open-ended tests based on *e*-statistics, which do not require a finite upper bound on the sample size.

5. Data Application

Berrett et al. (2020) analyze conditional independence relations in the Capital Bikeshare dataset. Code and data for their study are available on *https://rinafb.github.io/research/*, *https://ride.capitalbikeshare.com/system-data*. The dataset collects information on bike rides with the Capital bikeshare System in Washington DC. One question in the analysis by Berrett et al. (2020) is whether there is dependence of the duration of ride and the binary variable member type, distinguishing casual users from people who purchased a long-term membership, conditional on the start location, destination, and the time of the day during which the bike ride took place.

Let X denote the logarithm of the ride duration and denote by Z the three dimensional vector of starting point, destination of ride, and of the starting time. We model the distribution of $X \mid Z$ as Gaussian, $\mathcal{N}(\mu_Z, \sigma_Z^2)$ in exactly the same way as Berrett et al. (2020). Separately for each combination of starting point and destination, the conditional mean μ_Z and variance σ_Z^2 are estimated by a kernel regression of X with the time of day as covariate; details are given in Appendix B of Berrett et al. (2020). We also apply the same preselection criteria, namely, exclude data from weekends and holidays, and values of Z where the estimation might be imprecise due to scarce data. The member type is encoded as $Y \in \{0, 1\}$, with $Y = 0$ referring to casual members.

The data analyzed is from the months September to November in 2011. Unlike Berrett et al. (2020), who took the months September and November as training data for estimating the distribution of $X \mid Z$ and October as test data, we perform estimation on data of September and October and perform tests on the November data, which would be the natural order for real-time sequential analysis. The test dataset, after the application of selection criteria, contains 7173 observations. The training data consists of 158,741 observations. Due to the temporal structure of the data there might be some short lag autocorrelation between the observations, but we did not find any distorting influence of this on the results presented below.

Berrett et al. (2020) apply the conditional randomization test with the test statistic $|\mathrm{cor}(Y, X - \mathbb{E}_{Q_Z}[X])|$. Applied to the November data, this yields a *p*-value of essentially zero: the observed value of the test statistic was greater than any value obtained with simulated X, over 10,000 simulations. For the

Figure 3. *E*-value for the bike sharing dataset. Lines at the bottom indicate observation times.

e-statistics, we model the probability that $Y = 1$ with logistic regression, taking X and μ_Z as covariates and starting with a minimal sample size of 200 observations in the test data. A full model including Z instead of only μ_Z would be problematic due to the high number of combinations of starting points and destinations relative to the size of the test data. We do not have this limitation in the estimation of the distribution of $X \mid Z$ due to the much larger size of the training dataset. However, for a valid test it is not necessary to include Z itself in the model. The probability predictions from the logistic model are then truncated to the interval $[0.01, 0.99]$. For the sequential analysis, the rides are arranged in the order of the start date and time of ride. Figure 3 shows how evidence accumulates over time. At the end of the period, an *e*-value of more than 10^6 is attained, giving decisive evidence against the null hypothesis of conditional independence. An *e*-value of 10^4 is already reached at the 4380th observation, on November 15th, and hence after about half of the total observation time.

6. Discussion, Related and Future Work

We have proposed and analyzed anytime-valid tests of conditional independence in the model-X setting. Our method gives a general procedure to transform statistics that measure conditional independence to *e*-statistics. We have shown that for a simple alternative, using the conditional density as statistic leads to the growth rate optimal *e*-statistic, and derived a bound on the inflation of Type-I error under violations of the MX assumption.

Duan, Ramdas, and Wasserman (2022) and Shaer, Maman, and Romano (2022) have also proposed methods to test CI under MX, but they address the problem in fundamentally different ways than we do. In the fully sequential setting, Duan, Ramdas, and Wasserman (2022, Appendix E.4) construct a test martingale which grows if a researcher can better than randomly predict a binary X_n given past information and (Y_n, Z_n); notice, however, that this is just a small part of their work, which, for example, also includes tests for a batch setting. Shaer, Maman, and Romano (2022) construct a test martingale which measures how much better a forecaster can predict Y_n based on past data and general (X_n, Z_n), relative to a forecaster only having access to a randomized $\tilde{X}_n \sim Q_{Z_n}$. A derandomization of this procedure is obtained by taking the expectation over $\tilde{X}_n$, which yields similar test martingales as in Theorem 1. Both of these methods are very flexible in the choice and tuning of the prediction method, and may therefore behave quite differently from ours. However, ours are justified in terms of the strong GRO optimality criterion by Proposition 4 and Corollary 5, whenever a fast-converging estimator is used. While this makes our approach in some sense the optimal one, it requires specifying a reasonable (potentially nonparametric) set of densities $\mathcal{F}$ to define the estimator $\hat{f}$. We can basically use any set of densities we like, as long as the estimators converge in information in the sense below Corollary 5. However, it may not always be easy to find $\mathcal{F}$ for which estimation is computationally feasible and practically successful, especially if Z is high-dimensional. In that case the approaches of Duan, Ramdas, and Wasserman (2022) and Shaer, Maman, and Romano (2022) may have the advantage of being more flexible—whether or not this is the case may be domain-dependent, and is an interesting avenue for future research. Finally, the issue of robustness with respect to misspecification of the distribution of $X|Z$ is only assessed by Shaer, Maman, and Romano (2022) through simulations, whereas our Theorem 6 yields the same worst-case bound for our approach as in the batch setting.

For a comparison of our method to the CRT, the following aspects are worth highlighting. Our simulations suggest that our anytime-valid tests with optional stopping do not need more samples, on average, to achieve the same power as the CRT. To be precise, for a given desired power, researchers have to plan for a higher maximum sample size with our anytime-valid tests. But thanks to early stopping, the average sample size of experiments is not more or even less than for the classical CRT. This confirms the findings by Grünwald, de Heide, and Koolen (2019), Ter Schure et al. (2020), and Turner, Ly, and Grünwald (2021) on other anytime-valid tests and their nonsequential counterparts. In terms of computational complexity, our method is not less efficient than the classical CRT if the functions h_n in the test martingale can be updated in constant time; in our application on logistic regression we did not make use of recursive updating, though. A slight advantage of our method compared to the fixed sample size CRT is that the latter requires at least $\lceil 1/\alpha - 1 \rceil$ resamples in order to be able to obtain a *p*-value below α, whereas the test martingale $(S^{\mathrm{CI}}_{h^n})_{n\in\mathbb{N}}$ can exceed any level independently of the number M of resamples. If the strategy

of Proposition 2 is applied, then each factor in $S_{h^n}^{\mathrm{CI}}$ is bounded by $M + 1$, but not their cumulative product.

There are various avenues for future research. With respect to robustness, our Theorem 6 shows that for a fixed upper bound N on the sample size, the sequential test achieves the same worst case inflation of rejection rate as the nonsequential CRT with a sample size N. It would be of interest to investigate the robustness of the test when this upper bound grows and the approximation $\hat{Q}_Z$ of the conditional distributions of $X \mid Z$ are sequentially updated with new samples. Furthermore, our simulations indicate that our e-statistics have competitive power compared to existing methods with relatively small dimension of Z, but as stated above, further research is necessary to investigate suitable e-statistics and their power when Z is of higher dimension.

Supplementary Materials

The supplementary material includes all the code needed to reproduce the results in Sections 4 and 5, as well as a pdf file with three sections: Section S1 provides additional simulations, Section S2 gives the proofs of the main results in the paper, and Section S3 connects the GRO E-statistic discussed in the main text to anytime-valid E-statistics/E-processes.

Acknowledgments

We are grateful to Aaditya Ramdas, Yaniv Romano, and three anonymous referees for their helpful feedback on this article. Part of the computations have been performed on UBELIX (*https://ubelix.unibe.ch/*), the HPC cluster of the Universiy of Bern.

Funding

Alexander Henzi gratefully acknowledges support from the Swiss National Science Foundation (grant number 192269).

ORCID

Peter Grünwald https://orcid.org/0000-0001-9832-9936

References

Azadkia, M., Chatterjee, S. (2021), "A Simple Measure of Conditional Dependence," *The Annals of Statistics*, 49, 3070–3102. [1557]

Barron, A. (1998), "Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems," in *Bayesian Statistics* (Vol. 6), eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M., pp. 27–52, Oxford: Oxford University Press. [1558]

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020), "The Conditional Permutation Test for Independence While Controlling for Confounders," *Journal of the Royal Statistical Society*, Series B, 82, 175–197. [1555,1556,1559,1562]

Bilodeau, B., Foster, D. J., and Roy, D. M. (2021), "Minimax Rates for Conditional Density Estimation via Empirical Entropy," arXiv preprint arXiv:2109.10461. [1558]

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [1554,1555,1556]

Casper, C., Cook, T., and on FORTRAN program ld98., O. A. P. B. (2022), *ldbounds: Lan-DeMets Method for Group Sequential Boundaries*, R package version 2.0.0. [1560]

Cover, T. M., and Thomas, J. A. (1991), *Elements of Information Theory*, Wiley Series in Telecommunications, New York: Wiley. [1557,1558]

Darling, D., and Robbins, H. (1967), "Confidence Sequences for Mean, Variance, and Median," *Proceedings of the National Academy of Sciences*, 58, 66–68. [1555]

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, Series B, 41, 1–15. [1554]

Duan, B., Ramdas, A., and Wasserman, L. (2022), "Interactive Rank Testing by Betting," in *Proceedings of the First Conference on Causal Learning and Reasoning*, Vol. 177 of Proceedings of Machine Learning Research, pp. 201–235. [1555,1563]

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. (2018), "Logistic Regression: The Importance of Being Improper," in *Proceedings COLT*, pp. 167–208. [1558]

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007), "Kernel Measures of Conditional Dependence," in *Advances in Neural Information Processing Systems* (Vol. 20). [1557]

Grünwald, P. (2022), "Beyond Neyman-Pearson," arXiv preprint arXiv: 2205.00901. [1555]

Grünwald, P. D., and Mehta, N. A. (2020), "Fast Rates for General Unbounded Loss Functions," *Journal of Machine Learning Research*, 21, 1–80. [1558]

Grünwald, P., de Heide, R., and Koolen, W. (2019), "Safe Testing," arXiv preprint arXiv:1906.07801. [1555,1556,1561,1563]

Ham, D. W., Imai, K., and Janson, L. (2022), "Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis," arXiv preprint arXiv:2201.08343. [1554]

Katsevich, E., and Ramdas, A. (2022), "On the Power of Conditional Independence Testing Under Model-X," *Electronic Journal of Statistics*, 16, 6348–6394. [1554,1555,1556,1557]

Koolen, W. M., and Grünwald, P. (2022), "Log-Optimal Anytime-Valid E-Values," *International Journal of Approximate Reasoning*, 141, 69–82. [1556]

Li, S., and Liu, M. (2022), "Maxway CRT: Improving the Robustness of Model-X Inference," arXiv preprint arXiv:2203.06496. [1554]

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2021), "Fast and Powerful Conditional Randomization Testing via Distillation," *Biometrika*, 109, 277–293. [1556]

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.), CRC Monographs on Statistics and Applied Probability Series, New York: Chapman & Hall. [1554]

Niu, Z., Chakraborty, A., Dukes, O., Katsevich, E., (2022), "Reconciling Model-X and Doubly Robust Approaches to Conditional Independence Testing," arXiv preprint arXiv:2211.14698. [1554]

O'Brien, P. C., and Fleming, T. R. (1979), "A Multiple Testing Procedure for Clinical Trials," *Biometrics*, 35, 549–556. [1560]

Pocock, S. J. (1977), "Group Sequential Methods in the Design and Analysis of Clinical Trials," *Biometrika*, 64, 191–199. [1560]

Pocock, S. J., and Simon, R. (1975), "Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial," *Biometrics*, 31, 103–115. [1557]

Qian, G., and Field, C. (2002), "Law of Iterated Logarithm and Consistent Model Selection Criterion in Logistic Regression," *Statistics & Probability Letters*, 56, 101–112. [1559]

R Core Team (2022), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [1560]

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020), "Admissible Anytime-Valid Sequential Inference Must Rely on Nonnegative Martingales," arXiv preprint arXiv:2009.03167. [1555]

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022), "Testing Exchangeability: Fork-Convexity, Supermartingales and E-Processes," *International Journal of Approximate Reasoning*, 141, 83–109. [1555]

Robbins, H. (1952), "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, 58, 527–535. [1557]

Runge, J. (2018), "Conditional Independence Testing Based on a Nearest-Neighbor Estimator of Conditional Mutual Information," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (Vol. 84) of Proceedings of Machine Learning Research, pp. 938–947. [1557]

Shaer, S., Maman, G., and Romano, Y. (2022), "Model-Free Sequential Testing for Conditional Independence via Testing by Betting," arXiv preprint arXiv:2210.00354. [1555,1563]

Shafer, G. (2021), "Testing by Betting: A Strategy for Statistical and Scientific Communication," *Journal of the Royal Statistical Society*, Series A, 184, 407–431. [1555,1556]

Shah, R. D., and Peters, J. (2020), "The Hardness of Conditional Independence Testing and the Generalized Covariance Measure," *The Annals of Statistics*, 48, 1514–1538. [1554,1557]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011), "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent," *Journal of Statistical Software*, 39, 1–13. [1560]

Ter Schure, J., Perez-Ortiz, M. F., Ly, A., and Grunwald, P. (2020), "The Safe Logrank Test: Error Control under Continuous Monitoring with Unlimited Horizon," arXiv preprint arXiv:2011.06931. [1561,1562,1563]

Turner, R., Ly, A., and Grünwald, P. (2021), "Generic E-Variables for Exact Sequential k-Sample Tests That Allow for Optional Stopping," arXiv preprint arXiv:2106.02693. [1556,1561,1563]

Vovk, V., and Wang, R. (2021), "E-values: Calibration, Combination and Applications," *The Annals of Statistics*, 49, 1736–1754. [1555]

Wald, A. (1947), *Sequential Analysis*, New York: Wiley. [1556]

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020), "Universal Inference," *Proceedings of the National Academy of Sciences*, 117, 16880–16890. [1560,1561]

Wong, W. H., and Shen, X. S. (1995), "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES," *The Annals of Statistics*, 23, 339–362. [1558]

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007), "Asymptotic Properties of Covariate-Adjusted Response-Adaptive Designs," *The Annals of Statistics*, 35, 1166–1182. [1557]