Nonparametric conditional local independence testing

Christgau, Alexander Mangulad; Petersen, Lasse; Hansen, Niels Richard

Published in:
Annals of Statistics

DOI:
10.1214/23-AOS2323

Publication date:
2023

Document version
Publisher’s PDF, also known as Version of record

Document license:
Unspecified

Citation for published version (APA):
Christgau, A. M., Petersen, L., & Hansen, N. R. (2023). Nonparametric conditional local independence testing. Annals of Statistics, 51(5), 2116-2144. https://doi.org/10.1214/23-AOS2323
Conditional local independence is an asymmetric independence relation among continuous time stochastic processes. It describes whether the evolution of one process is directly influenced by another process given the histories of additional processes, and it is important for the description and learning of causal relations among processes. We develop a model-free framework for testing the hypothesis that a counting process is conditionally locally independent of another process. To this end, we introduce a new functional parameter called the Local Covariance Measure (LCM), which quantifies deviations from the hypothesis. Following the principles of double machine learning, we propose an estimator of the LCM and a test of the hypothesis using nonparametric estimators and sample splitting or cross-fitting. We call this test the (cross-fitted) Local Covariance Test ((X)-LCT), and we show that its level and power can be controlled uniformly, provided that the nonparametric estimators are consistent with modest rates. We illustrate the theory by an example based on a marginalized Cox model with time-dependent covariates, and we show in simulations that when double machine learning is used in combination with cross-fitting, then the test works well without restrictive parametric assumptions.

1. Introduction. Notions of how one variable influences a target variable are central to both predictive and causal modeling. Depending on the objective, the relevant notion of influence can be variable importance in a predictive model of the target, but it can also be the causal effect of the variable on the target. In either case, we can investigate influence conditionally on a third variable—to quantify the added predictive value, the direct causal effect or the causal effect adjusted for a confounder. Our interests are in an asymmetric notion of direct influence among stochastic processes, which is not adequately captured by classical (symmetric) notions of conditional dependence. The objective of this paper is therefore to quantify this notion of asymmetric influence and specifically to develop a new nonparametric test of the hypothesis that one stochastic process does not directly influence another.

The hypothesis we consider is that of local independence—a concept introduced by Schweder (1970) for Markov processes as a continuous time formalization of the phenomenon that the past of one stochastic process does not directly influence the evolution of another stochastic process. Generalizations to other continuous time processes were given by Aalen (1987) and studied by Commenges and Gégout-Petit (2009), who systematically used the term conditional local independence for the general concept. We will in this paper follow that convention whenever we want to emphasize the conditional nature of the local independence. We note that (conditional) local independence is a continuous time version of the discrete time concept of Granger noncausality (Aalen (1987), Granger (1969)).

To illustrate the concept of conditional local independence, we will in this Introduction consider an example involving three processes: $X$, $Z$ and $N$; see Figure 1. The process $N$ is
FIG. 1. Local independence graph illustrating a dependence structure among the three processes X, Z and N. Here, N is the indicator of death for an individual, X is their cumulative pension savings and Z is a covariate process. All nodes in this graph have implicit self-loops. There is no edge from X to N, which indicates that death is not directly influenced by pension savings. This can be formalized as N being conditionally locally independent of X, which is the hypothesis we aim to test.

the indicator of death, \( N_t = 1(T \leq t) \), for an individual with survival time \( T \), and \( X_t \) denotes the total pension savings of the individual at time \( t \). The process \( Z \) is a covariate process, for example, health variables or employment status, that may directly affect both the pension savings and the survival time. This is indicated in Figure 1 by edges pointing from \( Z \) to \( X \) and \( N \). Edges pointing from \( N \) to \( X \) and \( Z \) indicate that a death event directly affects both \( X \) and \( Z \) (which take the values \( X_T \) and \( Z_T \), respectively, after time \( T \), see Section 2.2).

To define conditional local independence, let \( \mathcal{F}_t^{N,Z} = \sigma(N_s, Z_s; s \leq t) \) denote the filtration generated by the \( N \)- and \( Z \)-processes. The \( \sigma \)-algebra \( \mathcal{F}_t^{N,Z} \) represents the information contained in the \( N \)- and the \( Z \)-processes before time \( t \). Informally, the process \( N_t \) is conditionally locally independent of the process \( X_t \) given \( \mathcal{F}_t^{N,Z} \) if \( (X_s)_{s \leq t} \) does not add predictable information to \( \mathcal{F}_t^{N,Z} \) about the infinitesimal evolution of \( N_t \). For this particular example, this means that the conditional hazard function of \( T \) does not depend on \( (X_s)_{s \leq t} \) given \( \mathcal{F}_t^{N,Z} \). In Figure 1, the hypothesis of interest, that \( N_t \) is conditionally locally independent of \( X_t \) given \( \mathcal{F}_t^{N,Z} \), is represented by the lack of an edge from \( X \) to \( N \).

A systematic investigation of algebraic properties of conditional local independence was initiated by Didelez (2007, 2008, 2015). She also introduced local independence graphs, such as the directed graph in Figure 1, to graphically represent all conditional local independencies among several processes, and she studied the semantics of these graphs. This work was extended further by Mogensen and Hansen (2020) to graphical representations of partially observed systems. While we will not formally discuss local independence graphs, the problem of learning such graphs from data was an important motivation for us to develop a non-parametric test of conditional local independence. A constraint-based learning algorithm of local independence graphs was given by Mogensen, Malinsky and Hansen (2018) in terms of a conditional local independence oracle, but a practical algorithm requires that the oracle is replaced by conditional local independence tests.

Another important motivation for considering conditional local independence arises from causal models. With a structural assumption about the stochastic process specification, a conditional local independence has a causal interpretation (Aalen (1987), Aalen et al. (2012), Commenges and Gégout-Petit (2009)), and if the causal stochastic system is completely observed, a test of conditional local independence is a test of no direct causal effect. See also Røysland et al. (2022), who use local independence graphs to formulate criteria for identification of causal effects in continuous-time survival models. If the causal stochastic system is only partially observed, a conditional local dependency need not correspond to a direct causal effect due to unobserved confounding, but the projected local independence graph, as introduced by Mogensen and Hansen (2020), retains a causal interpretation, and its Markov equivalence class can be learned by conditional local independence testing. In addition, within the
framework of structural nested models, testing the hypothesis of no total causal effect can also be cast as a test of conditional local independence (Lok (2008), Theorem 9.2).

To appreciate what conditional local independence means—and, in particular, what it does not mean—it is useful to compare with ordinary conditional independence. In our example, $N_t$ is conditionally locally independent of $X_t$ given $\mathcal{F}_{t-}^{N,Z}$, but this implies neither that $N \perp X \mid Z$ (as processes), nor that $N_t \perp X_t \mid \mathcal{F}_{t-}^Z$. In fact, these conditional independencies cannot hold in this example where $X_t = X_T$ for $t \geq T$—except in special cases such as $T$ being a deterministic function of $Z$. Theorem 2 in Didelez (2008) gives a sufficient condition for $N_t \perp X_t \mid \mathcal{F}_{t-}Z$ to hold in terms of the local independence graph, but this condition is also not fulfilled by the graph in Figure 1 due to the edge from $N$ to $X$. Moreover, conditional local independence is in general also different from the baseline conditional independence $T \perp X_0 \mid Z_0$ unless both $X$ and $Z$ are time-independent; see Section 3.2. In Section E in the Supplementary Material (Christgau, Petersen and Hansen (2023)), we elaborate further upon the connection to semiparametric survival models. Didelez (2008) argues that $N_t$ being conditionally locally independent of $X_t$ given $\mathcal{F}_{t-}^{N,Z}$ heuristically means that $N_t \perp \mathcal{F}_{t-}X \mid \mathcal{F}_{t-}^{N,Z}$, but this is technically problematic in continuous time. If $T$ has a continuous distribution, then for any fixed $t$, $N_t = N_{t-}$ almost surely, whence $N_t$ is almost surely $\mathcal{F}_{t-}^{N,Z}$-measurable and conditionally independent of anything given $\mathcal{F}_{t-}^{N,Z}$. This heuristic can thus not be used to formally define conditional local independence in continuous time. See instead the formal Definition 2 by Didelez (2008) or our Definition 2.1.

Several examples from health sciences given by Didelez (2008) demonstrate the usefulness of conditional local independence for multivariate event systems, and more recent attention to event systems in the machine learning community (Achab et al. (2017), Bacry et al. (2017), Cai et al. (2022), Xu, Farajtabar and Zha (2016), Zhou, Zha and Song (2013)) testifies to the relevance of conditional local independence. This line of research relies primarily on the linear Hawkes process model, which is effectively used to infer local independence graphs—sometimes even interpreted causally. The Hawkes model is attractive because conditional local independencies can be inferred from corresponding kernel functions being zero—and statistical tests can readily be based on parametric or nonparametric estimation of kernels.

A less attractive property of the Hawkes model is that it is not closed under marginalization. As with any model based statistical test, the validity of the test is jeopardized by model misspecification, hence even within a subsystem of a linear Hawkes process, a test of conditional local independence based on a Hawkes model may be invalid.

The challenges resulting from model misspecification and marginalization is investigated further in Sections 2.2 and 6 based on an extension of our introductory example and Cox’s survival model. Both the Hawkes model and the Cox model illustrate that conditional local independence might be expressed and tested within a (semi)parametric model, but model misspecification makes us question the validity of any such model based test. Thus, there is a need for a nonparametric test of the hypothesis of conditional local independence. Moreover, since we cannot translate the hypothesis into an equivalent hypothesis about classical conditional independence, we cannot directly use existing nonparametric tests, such as the GCM (Shah and Peters (2020)) or the GHCM (Lundborg, Shah and Peters (2022)), of conditional independence.

We propose a new nonparametric test when the target process $N$ is a counting process and $X$ is a real valued process, and where the hypothesis is that $N$ is conditionally locally independent of $X$ given a filtration $\mathcal{F}_t$. In the context of the introductory example, $\mathcal{F}_t = \mathcal{F}_{t-}^{N,Z}$. We consider a counting process target primarily because the theory of conditional local independence is most complete in this case, but generalizations are possible—we refer to the discussion in Section 7. Within our framework, we base our test on an infinite-dimensional parameter, which we call the Local Covariance Measure (LCM). It is a function of time,
which is constantly equal to zero under the hypothesis. Our main result is that the LCM can be estimated by using the ideas of double machine learning (Chernozhukov et al. (2018)) in such a way that the estimator converges uniformly at a \( \sqrt{n} \)-rate to a mean zero Gaussian martingale under the hypothesis of conditional local independence. We use the LCM to develop the (cross-fitted) Local Covariance Test ((X)-LCT), for which we derive uniform level and power results.

1.1. Organization of the paper. In Section 2, we introduce the general framework for formulating the hypothesis of conditional local independence. This includes the introduction in Section 2.1 of an abstract residual process, which is used to define the LCM as a functional target parameter indexed by time. The LCM equals the zero-function under the hypothesis of conditional local independence, and to test this hypothesis we introduce an estimator of the LCM in Section 2.3. The estimator is a stochastic process, and we describe how sample splitting is to be used for its computation via the estimation of two unknown components.

In Section 3, we give interpretations of the LCM and its estimator. We show that the LCM estimator is a Neyman orthogonalized score statistic in Section 3.1, and in Section 3.2 we relate LCM to the partial copula when \( X \) is time-independent.

In Section 4, we state the main results of the paper. We establish in Section 4.1 that the LCM estimator generally approximates the LCM with an error of order \( n^{-1/2} \). Under the hypothesis of conditional local independence, we show that the (scaled) LCM estimator converges weakly to a mean zero Gaussian martingale. The estimator requires a model of the target process \( N \) as well as the process \( X \) conditionally on \( \mathcal{F}_t \) to achieve the orthogonalization at the core of double machine learning. The model of \( X \) is in this paper expressed indirectly in terms of the residual process, and we show that if we can learn the residual process at rate \( g(n) \) and the model of \( N \) at rate \( h(n) \) such that \( g(n), h(n) \to 0 \) and \( \sqrt{n} g(n) h(n) \to 0 \) for \( n \to \infty \) then we achieve a \( \sqrt{n} \)-rate convergence of the LCM estimator. We also show that the variance function of the Gaussian martingale can be estimated consistently, and we give a general result on the asymptotic distribution of univariate test statistics based on the LCM estimator. All asymptotic results are presented in the framework of uniform stochastic convergence.

Section 5 gives explicit examples of univariate test statistics, including the local covariance test based on the normalized supremum of the LCM estimator. Its asymptotic distribution is derived and we present results on uniform asymptotic level and power. In Section 5.2, we present the generalization from the sample split estimator to the cross-fit estimator. Though this estimator and the corresponding cross-fit Local Covariance Test (X-LCT) are a bit more involved to compute and analyze, X-LCT is more powerful, and thus our recommended test for practical usage.

The survival example from the Introduction is used and elaborated upon throughout the paper. We introduce a Cox model in terms of the time-varying covariate processes, and we report in Section 6 the results from a simulation study based on this model.

The paper is concluded by a discussion in Section 7.

The Supplementary Material (Christgau, Petersen and Hansen (2023)), henceforth referred to as the supplement, consists of Sections A through G and contains: proofs of results in this paper (A); definitions and results on uniform asymptotics (B); a uniform version of Rebolledo’s martingale CLT (C); an overview of achievable rate results for estimation of nuisance parameters that enter into the LCM estimator (D); a comparison with semiparametric survival models (E); details on Neyman orthogonality (F) and additional results from the simulation study (G).
2. The local covariance measure. In this section, we present the general framework of the paper, we define conditional local independence and we introduce the local covariance measure as a means to quantify deviations from conditional local independence. In Section 2.3, we outline how the local covariance measure can be estimated using double machine learning and sample splitting. We illustrate the central concepts and methods by an example based on Cox’s survival model with time-varying covariates.

We consider a counting process \( N = (N_t) \) and another real value process \( X = (X_t) \), both defined on the probability space \( (\Omega, \mathcal{F}, P) \). All processes are assumed to be defined on a common compact time interval. We assume, without loss of generality, that the time interval is \([0, 1]\). We will assume that \( N \) is adapted w.r.t. a right continuous and complete filtration \( \mathcal{F}_t \), and we denote by \( \mathcal{G}_t \) the right continuous and complete filtration generated by \( \mathcal{F}_t \) and \( X_t \). We assume throughout that \( X \) is càglàd (i.e., has sample paths that are continuous from the left and with limits from the right), which will ensure bounded sample paths and that the process is \( \mathcal{G}_t \)-predictable.

In the survival example of the Introduction, \( N_t = 1(T \leq t) \) is the indicator of whether death has happened by time \( t \), and there can only be one event per individual observed. Furthermore, \( \mathcal{F}_t = \mathcal{F}^{N,Z}_t \) and \( \mathcal{G}_t = \mathcal{F}^{N,X,Z}_t \). Our general setup works for any counting process, thus it allows for recurrent events and censoring, and the filtration \( \mathcal{F}_t \) can contain the histories of any number of processes in addition to the history of \( N \) itself.

2.1. The hypothesis of conditional local independence. The counting process \( N \) is assumed to have an \( \mathcal{F}_t \)-intensity \( \lambda_t \), that is, \( \lambda_t \) is \( \mathcal{F}_t \)-predictable and with

\[
\Lambda_t = \int_0^t \lambda_s \, ds
\]

being the compensator of \( N \),

(1) \[ M_t = N_t - \Lambda_t \]

is a local \( \mathcal{F}_t \)-martingale. Within this framework, we can define the hypothesis of conditional local independence precisely.

**Definition 2.1 (Conditional local independence).** We say that \( N_t \) is conditionally locally independent of \( X_t \) given \( \mathcal{F}_t \) if the local \( \mathcal{F}_t \)-martingale \( M_t \) defined by (1) is also a local \( \mathcal{G}_t \)-martingale.

For simplicity, we may also refer to this hypothesis as local independence and write

(2) \[ H_0 : M_t = N_t - \Lambda_t \text{ is a local } \mathcal{G}_t \text{-martingale.} \]

As argued in the Introduction, the hypothesis of local independence is the hypothesis that observing \( X \) on \([0, t]\) does not add any information to \( \mathcal{F}_{t-} \) about whether an \( N \)-event will happen in an infinitesimal time interval \([t, t + dt]\). Definition 2.1 captures this interpretation by requiring that the \( \mathcal{F}_t \)-compensator, \( \Lambda_t \), of \( N \) is also the \( \mathcal{G}_t \)-compensator. Thus, \( \lambda \) is also the \( \mathcal{G}_t \)-intensity under \( H_0 \).

If \( N \) has \( \mathcal{G}_t \)-intensity \( \lambda \), the innovation theorem, Theorem II.T14 in Brémaud (1981), gives that the predictable projection \( \lambda_t = \mathbb{E}(\lambda_t | \mathcal{F}_{t-}) \) is the (predictable) \( \mathcal{F}_t \)-intensity. Local independence follows if \( \lambda \) is \( \mathcal{F}_t \)-predictable. Intensities are, however, only unique almost surely, and we can have local independence even if \( \lambda \) is not a priori \( \mathcal{F}_t \)-predictable but have an \( \mathcal{F}_t \)-predictable version. When \( N \) has \( \mathcal{G}_t \)-intensity \( \lambda \), \( H_0 \) is thus equivalent to \( \lambda \) having an \( \mathcal{F}_t \)-predictable version. We find Definition 2.1 preferable because it directly gives an operational criterion for determining whether \( N \) has an \( \mathcal{F}_t \)-predictable version of a \( \mathcal{G}_t \)-intensity.
Remark 2.2 (Censoring). Suppose that the data is censored such that \((N_t, X_t, \mathcal{F}_t) = (N^*_t \land C, X^*_t \land C, \mathcal{F}^*_t \land C)\), where \((N^*, X^*, \mathcal{F}^*)\) are uncensored data and where \(C\) is the censoring time. The hypothesis regarding the uncensored data,

\[
H^*_0 \colon N^*_t \text{ is locally independent of } X^*_t \text{ given } \mathcal{F}^*_t,
\]

might then be the hypothesis of interest. If \(\mathbb{1}(C \geq t)\) happens to be \(\mathcal{F}^*_t\)-predictable, it is straightforward to show that \(H^*_0\) implies \(H_0\), and consequently a test of \(H_0\) is also a test of \(H^*_0\). However, \(\mathcal{F}^*_t\) may not \textit{a priori} contain information about the censoring process. Suppose instead that the common condition of independent censoring (Andersen et al. (1993)) holds, which is equivalent to \(N^*_t\) being locally independent of \(C_t := \mathbb{1}(C \leq t)\) given \(\mathcal{G}^*_t\) (Roysland et al. (2022)). Then \(H^*_0\) implies that \(N_t\) is locally independent of \(X_t\) given \(\mathcal{F}_t \vee \mathcal{F}^*_t\). Thus, in order to test \(H^*_0\), we replace \(\mathcal{F}_t\) by the enlarged filtration \(\mathcal{F}_t \vee \mathcal{F}^*_t\) and proceed \textit{mutatis mutandis} with testing \(H_0\) using the observed data.

Since \(X\) is assumed càglàd, and thus especially \(\mathcal{G}_t\)-predictable, the stochastic integral

\[
\int_0^t X_s \, dM_s,
\]

is under \(H_0\) a local \(\mathcal{G}_t\)-martingale. A test could be based on detecting whether (3) is, indeed, a local martingale. We will take a slightly different approach where we replace the integrand \(X\) by a residual process as defined below. We do so for two reasons. First, to achieve a \(\sqrt{n}\)-rate via double machine learning we need the integrand to fulfill (4) below. Second, other choices of integrands than \(X\) could potentially lead to more powerful tests.

Definition 2.3 (Residual process). A residual process \(G = (G_t)_{t \in [0,1]}\) of \(X_t\) given \(\mathcal{F}_t\) is a càglàd stochastic process that is \(\mathcal{G}_t\)-adapted and satisfies

\[
\mathbb{E}(G_t | \mathcal{F}_{t-}) = 0, \quad t \in [0,1].
\]

The geometric interpretation is that the residual process evolves such that \(G_t\) is orthogonal to \(L_2(\mathcal{F}_{t-})\) within \(L_2(\mathcal{G}_{t-})\) at each time \(t\). One obvious residual process is the \textit{additive residual process} given by

\[
G_t = X_t - \Pi_t = X_t - \mathbb{E}(X_t | \mathcal{F}_{t-}),
\]

where \(\Pi_t = E(X_t | \mathcal{F}_{t-})\) denotes the predictable projection of the càglàd process \(X_t\); see Theorem VI.19.2 in (Rogers and Williams (2000)). The additive residual projects \(X_t\) onto the orthogonal complement of \(L_2(\mathcal{F}_{t-})\), but this may not necessarily remove all \(\mathcal{F}_t\)-predictable information from \(X_t\). An alternative choice that does so under sufficient regularity conditions is the \textit{quantile residual process} given by

\[
G_t = F_t(X_t) - \frac{1}{2},
\]

where \(F_t\) is the conditional distribution function given by \(F_t(x) = \mathbb{P}(X_t \leq x | \mathcal{F}_{t-})\). The quantile residual process satisfies (4) provided that \((t, x) \mapsto F_t(x)\) is continuous. In Section 3.1, we discuss additional transformations of \(X\) that can also be applied before any residualization procedure.

We will formulate the general results in terms of an abstract residual process, but we focus on the additive residual process in the examples. Any nondegenerate residual process will contain a predictive model of (aspects of) \(X_t\) given \(\mathcal{F}_{t-}\) in order to satisfy (4). We use \(\hat{G}_t\) to denote the residual obtained by plugging in an estimate of that predictive model. For the additive residual process, the predictive model is \(\Pi_t\) and \(\hat{G}_t = X_t - \hat{\Pi}_t\). For the quantile residual process, the predictive model is \(F_t\) and \(\hat{G}_t = F_t(X_t) - \frac{1}{2}\).

We can now define our functional target parameter of interest, which we call the \textit{local covariance measure}.
DEFINITION 2.4 (Local covariance measure). With $G_t$ a residual process, define for $t \in [0, 1]$,
\[
\gamma_t = \mathbb{E}(I_t), \quad \text{where } I_t = \int_0^t G_s \, dM_s,
\]
whenever the expectation is well-defined. We call the function $t \mapsto \gamma_t$ the Local Covariance Measure (LCM).

The following propositions illuminate how $\gamma$ relates to the null hypothesis of $N_t$ being conditionally locally independent of $X_t$ given $\mathcal{F}_t$.

**PROPOSITION 2.5.** Under $H_0$, the process $I = (I_t)$ is a local $\mathcal{G}_t$-martingale with $I_0 = 0$. If $I$ is a martingale, then $\gamma_t = 0$ for $t \in [0, 1]$.

To interpret $\gamma$ in the alternative, we assume that $N$ has $\mathcal{G}_t$-intensity $\lambda$.

**PROPOSITION 2.6.** If $\int_0^1 \mathbb{E}(|G_s| (\lambda_s + \lambda_s)) \, ds < \infty$, then for every $t \in [0, 1]$,
\[
\gamma_t = \int_0^t \text{cov}(G_s, \lambda_s - \lambda_s) \, ds.
\]
In particular, $\gamma$ is the zero-function if and only if $\text{cov}(G_s, \lambda_s - \lambda_s) = 0$ for almost all $s \in [0, 1]$.

We note that under $H_0$, the condition $\int_0^1 \mathbb{E}(|G_s| \lambda_s) \, ds < \infty$ is sufficient to ensure that $I$ is a martingale and $\gamma_t = 0$ for all $t \in [0, 1]$. By Proposition 2.6, the LCM quantifies deviations from $H_0$ in terms of the covariance between the residual process and the difference of the $\mathcal{F}_t$- and $\mathcal{G}_t$-intensities. To this end, note that if $X$ happens to be $\mathcal{F}_t$-adapted, then $\mathcal{G}_t = \mathcal{F}_t$ and $N$ is trivially locally independent of $X$. The hypothesis of local independence is only of interest when $\mathcal{G}_t$ is a strictly larger filtration than $\mathcal{F}_t$, that is, when $X$ provides information not already in $\mathcal{F}_t$.

For the additive residual process, where $G_t = X_t - \Pi_t$,
\[
\gamma_t = \mathbb{E}\left(\int_0^t G_s \, dM_s\right) = \mathbb{E}\left(\int_0^t X_s \, dM_s\right) - \mathbb{E}\left(\int_0^t \Pi_s \, dM_s\right)
\]
provided that the expectations are well-defined. Since the predictable projection $\Pi_t$ has a càglàd version and is $\mathcal{F}_t$-predictable, and since $M_t$ is a local $\mathcal{F}_t$-martingale, $\int_0^t \Pi_s \, dM_s$ is a local $\mathcal{F}_t$-martingale. If it is a martingale, it is a mean zero martingale, and
\[
\gamma_t = \mathbb{E}\left(\int_0^t X_s \, dM_s\right) = \mathbb{E}\left(\sum_{\tau \leq t: \Delta N_\tau = 1} X_\tau - \int_0^t X_s \lambda_s \, ds\right).
\]

The computation above shows that the additive residual process defines the same functional target parameter $\gamma_t$ as the stochastic integral (3) would. It is, however, the representation of $\gamma_t$ as the expectation of the residualized stochastic integral that will allow us to achieve a $\sqrt{n}$-rate of convergence of the estimator of $\gamma_t$ in cases where the estimator of $\lambda_t$ converges at a slower rate.
2.2. A Cox model with a partially observed covariate process. To further illustrate the hypothesis of conditional local independence and the local covariance measure we consider an example based on Cox’s survival model with time dependent covariates. This is an extension of the example from the Introduction with \( T \) being the time to death of an individual, and with \( X \) and \( Z \) being time-varying processes. There is, moreover, one additional time-varying process \( Y \) in the full model.

An interpretation of the processes is as follows:

\[
X = \text{Pension savings}, \\
Y = \text{Blood pressure}, \\
Z = \text{BMI}.
\]

Periods of overweight or obesity may influence blood pressure in the long term, and due to, for example, job market discrimination, high BMI could influence pension savings negatively. Death risk is influenced directly by BMI and blood pressure but not the size of your pension savings. Figure 2 illustrates two possible dependence structures among the three processes and the death time as local independence graphs, and we will use these two graphs to discuss the concept of conditional local independence of pension savings on time to death.

We assume that \( T \in [0, 1] \) and that \( X, Y \) and \( Z \) have continuous sample paths. Recall also that \( N_t = 1(T \leq t) \) is the death indicator process. To maintain some form of realism, all processes are stopped at time of death, that is, \( X_t = X_T, Y_t = Y_T \) and \( Z_t = Z_T \) for \( t \geq T \). This feedback from the death event to the other processes is reflected in Figure 2 by the edges pointing out of \( N \).

Recall also that \( F_{N,X,Y,Z}^t = \sigma(N_s, Z_y, Y_s; s \leq t) \) is the filtration generated by the \( N \)- and \( Z \)-processes. We use a similar notation for other processes and combinations of processes. For example, \( F_{N,Y,Z}^t \) is the filtration generated by \( N \) and all three \( X \)-, \( Y \)- and \( Z \)-processes. With \( \lambda_{i}^{\text{full}} \) denoting the \( F_{N,Y,Z}^t \)-intensity of time of death based on the history of all processes, we assume in this example a Cox model given by

\[
\lambda_{i}^{\text{full}} = 1(T \geq t)\lambda_{i}^{0}e^{Y_t+\beta Z_t}
\]

with \( \lambda_{i}^{0} \) a deterministic baseline intensity. It is not important that \( \lambda_{i}^{\text{full}} \) is a Cox model for our general theory, but it allows for certain theoretical computations in this example.

The fact that \( \lambda_{i}^{\text{full}} \) does not depend upon \( X_t \) implies that \( \lambda_{i}^{\text{full}} \) is also the \( F_{N,Y,Z}^t \)-intensity, and according to Definition 2.1, \( N_t \) is conditionally locally independent of \( X_t \) given \( F_{N,Y,Z}^t \).
This is in agreement with the local independence graphs in Figure 2 where there is no edge in either of them from $X$ to $N$.

We will take an interest in the case where $Y$ is unobserved and test the hypothesis:

$$H_0: N_t \text{ is conditionally locally independent of } X_t \text{ given } F_{N,Z,t}.$$  

That is, with $Y$ unobserved we want to test if the intensity of time to death given the history of $N, X$ and $Z$ depends on $X$. To simplify notation, let $F_t = F_{N,Z,t}^{N,Z}$ and $G_t = F_{N,X,Z,t}^{N,X,Z}$, in accordance with the general notation. The $G_t$-intensity is by the innovation theorem given as

$$\lambda_t = \mathbb{E}(\lambda_{t}^{\text{full}} \mid G_{t}) = 1(T \geq t)\lambda_0 e^{\beta Z_t} \mathbb{E}(e^{Y_t} \mid G_{t}),$$

while the $F_t$-intensity is

$$\lambda_t = \mathbb{E}(\lambda_{t}^{\text{full}} \mid F_{t}) = 1(T \geq t)\lambda_0 e^{\beta Z_t} \mathbb{E}(e^{Y_t} \mid F_{t}),$$

and $H_0$ is equivalent to $\lambda_t = \lambda_t$ almost surely. Comparing (8) and (9), we see that $H_0$ holds in this example if

$$\mathbb{E}(e^{Y_t} \mid G_{t}) = \mathbb{E}(e^{Y_t} \mid F_{t}),$$

and a sufficient condition for this is

$$F_{X,t} \perp \perp F_{Y,t} \mid F_t.$$  

The condition (10) is in concordance with the left graph in Figure 2 (see Theorem 2 in Didelez (2008)), but not the right, and it implies $H_0$. We will in Section 6.1 elaborate on condition (10) and give explicit examples.

We recall that $H_0$ can be reformulated as $\lambda_t$ not depending on $X$, and we could investigate the hypothesis via a marginal Cox model

$$\lambda^{\text{cox}}_t = 1(T \geq t)\lambda_0 e^{\alpha_1 X_t + \alpha_2 Z_t}$$

and test if $\alpha_1 = 0$. The Cox model is, however, not closed under marginalization and the semiparametric model (11) is quite likely misspecified. Consequently, a test of $\alpha_1 = 0$ is not equivalent to a test of $H_0$.

Our proposed nonparametric test of $H_0$ does not rely on a specific (semi)parametric model of $\lambda_t$. To test $H_0$, we consider the LCM using the additive residual process. Then (6) implies that

$$\gamma_t = \mathbb{E}(X_T N_t - \int_0^T X_s \lambda_s ds).$$

By Proposition 2.5, $\gamma_t = 0$ for $t \in [0, 1]$ under $H_0$, whence conditional local independence implies $\gamma_t = 0$, and we test $H_0$ by estimating $\gamma_t$ and testing if it is constantly equal to 0.

Before introducing a general estimator of the LCM in Section 2.3, we outline how to estimate the end point parameter $\gamma_1$ in this example. Due to $T \leq 1$ and the appearance of the indicator $1(T \geq t)$ in (9),

$$\gamma_1 = \mathbb{E}(X_T - \int_0^T X_s \lambda_s ds).$$

With i.i.d. observations $(T_1, X_1, Z_1), \ldots, (T_n, X_n, Z_n)$ and (nonparametric) estimates, $\hat{\lambda}_{j,t}$, based on $(T_1, Z_1), \ldots, (T_n, Z_n)$, we could compute the plug-in estimate

$$\hat{\gamma}_{1,\text{plug-in}}^{(n)} = \frac{1}{n} \sum_{j=1}^n \left( X_{j,T_j} - \int_0^{T_j} X_{j,s} \hat{\lambda}_{j,s} ds \right).$$

However, we cannot expect the plug-in estimator to have a $\sqrt{n}$-rate unless $\hat{\lambda}$ has $\sqrt{n}$-rate, which effectively requires parametric model assumptions on the intensity. Using the definition of $\gamma_1$ in terms of the additive residual process $G_t = X_t - \Pi_t$, we also have that

$$\gamma_1 = \mathbb{E}(X_T - \Pi_T - \int_0^T (X_s - \Pi_s) \lambda_s ds).$$
FIG. 3. Histograms of the distributions of three different estimators of $\gamma_1$. Each histogram contains 1000 estimates fitted to samples of size $n = 500$. The samples were sampled from a model that satisfies the hypothesis of conditional local independence and hence the ground truth is $\gamma_1 = 0$. See Section 6.2 for further details of the data generating process.

A double machine learning estimator based on the ideas by Chernozhukov et al. (2018) is therefore obtained by plugging in two nonparametric estimators:

$$\hat{\gamma}_{1,\text{double}}^{(n)} = \frac{1}{n} \sum_{j=1}^{n} \left( X_{j,T_j} - \hat{N}_{j,T_j} - \int_0^{T_j} (X_{j,s} - \hat{N}_{j,s}) \hat{\lambda}_{j,s} \, ds \right).$$

To achieve a small bias and a $\sqrt{n}$-rate of convergence, we use sample splitting. The nonparametric estimates $\hat{N}_{j}$ and $\hat{\lambda}_{j}$ are based on one part of the sample only, and are thus independent of the other part of the sample used for testing; see Section 2.3. To obtain a fully efficient estimator, multiple sample splits can be combined, for example, via cross-fitting; see Section 5.2.

Figure 3 shows the distributions of $\hat{\gamma}_{1,\text{plug-in}}^{(500)}$ and $\hat{\gamma}_{1,\text{double}}^{(500)}$ for the Cox example with $\gamma_1 = 0$; see Section 6.2 for details on the full model specification. The latter estimator was computed using cross-fitting but also without using any form of sample splitting. The figure illustrates the bias of $\hat{\gamma}_{1,\text{plug-in}}^{(500)}$, which is somewhat diminished by double machine learning without sample splitting and mostly eliminated by double machine learning in combination with cross-fitting.

2.3. Estimating the local covariance measure. To estimate the LCM, we assume that we have observed $n$ i.i.d. replications of the processes, $(N_1, X_1, F_1), \ldots, (N_n, X_n, F_n)$, where observing $F_j = (F_{j,t})$ signifies that anything adapted to the $j$th filtration is computable from observations. The process $N_j$ is adapted to $F_j$, while $X_j$ is not, and $G_j$ denotes the smallest right continuous and complete filtration generated by $X_j$ and $F_j$.

For each $n$, we consider a sample split corresponding to a partition $J_n \cup J_n^c = \{1, \ldots, n\}$ of the indices into two disjoint sets. We let $\hat{\lambda}^{(n)}$ and $\hat{G}^{(n)}$ be estimates of the intensity and the residualization map, respectively, fitted on data indexed by $J_n^c$ only. By an estimate, $\hat{\lambda}^{(n)}$, of $\lambda$, we mean a (stochastic) function that can be evaluated on the basis of $F_{j,t}$ for $j \in J_n$, and its value, denoted by $\hat{\lambda}_{j,t}^{(n)}$, is interpreted as a prediction of $\lambda_{j,t}$. The stochasticity in $\hat{\lambda}^{(n)}$ arises from its dependence on data indexed by $J_n^c$, from which its functional form is completely determined. Similarly, $\hat{G}^{(n)}$ is a function that can be evaluated on the basis of $G_{j,t}$ for $j \in J_n$. 
Algorithm 1: Sample split estimator of LCM

1 **input**: processes \((N_j, X_j, Z_j)\)\(j = 1, \ldots, n\), partition \(J_n \cup J_n^c\) of indices;
2 **options**: historical regression methods for estimation of \(\hat{\lambda}\) and \(\hat{G}\) given \(N\) and \(Z\),
3 discrete time grid \(0 = t_0 < \cdots < t_k \leq 1\);
4 **begin**
5 historically regress \((X_j)_{j \in J_n^c}\) on \((N_j, Z_j)_{j \in J_n^c}\) to obtain a fitted model \(\hat{G}^{(n)}\);
6 historically regress \((N_j)_{j \in J_n^c}\) on \((N_j, Z_j)_{j \in J_n^c}\) to obtain a fitted model \(\hat{\lambda}^{(n)}\);
7 compute out of sample residuals \(\hat{G}^{(n)}_{j, t_i}\) and \(\hat{M}^{(n)}_{j, t_i}\) for \(j \in J_n\) and \(i = 0, \ldots, k\);
8 for each \(i = 1, \ldots, k\), compute
   \[
   \hat{\gamma}^{(n)}_{t_i} = \frac{1}{|J_n|} \sum_{j \in J_n} \sum_{1 \leq l \leq i} \hat{G}^{(n)}_{j, t_l} (\hat{M}^{(n)}_{j, t_l} - \hat{M}^{(n)}_{j, t_{l-1}})
   \]
9 **end**
10 **output**: Local covariance measure \(\hat{\gamma}^{(n)}\) numerically approximated on grid;

To give a prediction \(\hat{G}^{(n)}_{j, t}\) of \(G_{j, t}\). In Section 6.1, we illustrate through the Cox example how \(\hat{\lambda}^{(n)}\) and \(\hat{G}^{(n)}\) are to be computed in practice when we use sample splitting. In Section D in the supplement, we give more examples of such estimation procedures and discuss their statistical properties in greater detail.

To ease notation, we will throughout assume that \((N, X, F)\) denotes one additional process and filtration—indeed of and with the same distribution as the observed processes. Then the estimated intensity \(\hat{\lambda}^{(n)}\) and estimated residual process \(\hat{G}^{(n)}\) can be evaluated on \((N, X, F)\), and thus we may write \(\hat{\lambda}^{(n)}_{t}\) and \(\hat{G}^{(n)}_{t}\) to denote template copies of \(\hat{\lambda}^{(n)}_{j, t}\) and \(\hat{G}^{(n)}_{j, t}\) for \(j \in J_n\).

In terms of the estimates \(\hat{\lambda}^{(n)}\) and \(\hat{G}^{(n)}\), we estimate LCM by the stochastic process \(\hat{\gamma}^{(n)}\) given by

\[
\hat{\gamma}^{(n)}_{t} = \frac{1}{|J_n|} \sum_{j \in J_n} \int_0^t \hat{G}^{(n)}_{j, s} d\hat{M}^{(n)}_{j, s},
\]

where \(\hat{M}^{(n)}_{j, t} = N_{j, t} - \int_0^t \hat{\lambda}^{(n)}_{j, s} ds\). We can regard \(\hat{\gamma}^{(n)}_{t}\) as a double machine learning estimator of \(\gamma_t\), with the observations indexed by \(J_n^c\) used to learn models of \(\hat{\lambda}\) and \(\hat{G}\), and with observations indexed by \(J_n\) used to estimate \(\gamma_t\) based on these models. In Section 5.2, we define the more efficient estimator that uses cross-fitting, but it is instructive to study the simpler estimator based on sample splitting first.

In practical applications, we do not directly observe the filtration \(F_j\), but rather samples from the stochastic processes generating the filtration. In accordance with the introductory Cox example, consider \(F_j\) and \(G_j\) given by \(F_j, t = \sigma(Z_{j, s}, N_{j, s}; s \leq t)\) and \(G_j, t = \sigma(X_{j, s}, Z_{j, s}, N_{j, s}; s \leq t)\) for a third stochastic process \(Z_j\), with \(Z_j\) possibly being multivariate. Within this setup, a general procedure for numerically computing the LCM is described in Algorithm 1. Here, historical regression refers to any method, which regresses the outcome at a given time on the history of the regressors up to that time. For example, historical linear regression is discussed in Section 6 and various alternative methods are discussed in Section D in the supplement. The choice of sample split will be discussed further in Section 5.2 in the context of cross-fitting.

As in Section 2.2, we could suggest estimating the entire function \(t \mapsto \gamma_t\) by a simple plug-in estimator of \(\hat{\lambda}\) using the representation (6). Figure 4 illustrates the distribution of
FIG. 4. A time dependent extension of Figure 3 showing the distribution of the sample paths \( t \mapsto \hat{\gamma}_t^{(500)} \) and \( t \mapsto \hat{\gamma}_t^{(500)} \), the latter with and without using cross-fitting. The data were simulated under \( H_0 \) where \( t \mapsto \gamma_t \) is the zero function. See Section 6.2 for further details of the data generating process.

estimators of the entire time dependent LCM for this plug-in estimator together with the double machine learning estimator with and without using cross-fitting. The figure also shows the distribution of the endpoint being the same distribution shown in Figure 3. The simulation is under \( H_0 \), and we see that only the double machine learning estimator with cross-fitting results in estimated sample paths centered around 0.

3. Interpretations of the LCM estimator. In this section, we provide some additional perspectives on and interpretations of the LCM. First, we show that the LCM estimator can be seen as a Neyman orthogonalization of the score statistic for a particular one-parameter family. The abstract formulation of the residual process \( (G_t) \) permits that we transform \( X \) into another \( G_t \)-predictable processes. Using this perspective, we may optimize the choice of the process \( X \) in terms of power.

Next, we show that when \( X \) is independent of time, the test statistic reduces in a survival context to certain covariance measures between \( X \)-residuals and Cox–Snell residuals, which we can link to existing test statistics for ordinary conditional independence.

3.1. Neyman orthogonalization of a score statistic. Consider the one-parameter family of \( G_t \)-intensities

\[
\lambda_t^\beta = e^{\beta X_t} \lambda_t
\]

for \( \beta \in \mathbb{R} \). Within this one-parameter family, the hypothesis of conditional local independence is equivalent to \( H_0 : \beta = 0 \). The normalized log-likelihood with \( n \) i.i.d. observations in the interval \([0, t]\) is

\[
\ell_t(\beta) = \frac{1}{n} \sum_{j=1}^n \left( \int_0^t \log(\lambda_{j,s}^\beta) \, dN_{j,s} - \int_0^t \lambda_{j,s}^\beta \, ds \right)
\]

\[
= \frac{1}{n} \sum_{j=1}^n \left( \int_0^t \beta X_{j,s} + \log(\lambda_{j,s}) \, dN_{j,s} - \int_0^t e^{\beta X_{j,s}} \lambda_{j,s} \, ds \right).
\]

Straightforward computations show that

\[
\partial_\beta \ell_t(0) = \frac{1}{n} \sum_{j=1}^n \int_0^t X_{j,s} \, dM_{j,s} \quad \text{and} \quad -\partial_\beta^2 \ell_t(0) = \frac{1}{n} \sum_{j=1}^n \int_0^t X_{j,s}^2 \lambda_{j,s} \, ds.
\]
If $\lambda$ were known, the score statistic $\partial\beta \ell_t(0)$ satisfies $\mathbb{E}(\partial\beta \ell_t(0)) = \gamma_t$. Moreover, under $H_0 : \beta = 0$ we have that $-\partial^2\beta \ell_t(0) = (\partial\beta \ell_t(0))$ is a consistent estimate of the asymptotic variance of the mean zero martingale $\partial\beta \ell_t(0)$. The hypothesis of local independence—with $\lambda$ known—could thus be tested using the score test statistic $-\partial^2\beta \ell_t(0) / \partial\beta^2 \ell_t(0)$.

The nuisance parameter $\lambda$ is, however, unknown and we want to avoid restrictive parametric assumptions about $\lambda$. Replacing $X_{j,t}$ by the residual process $G_{j,t}$ in the score statistic $\partial\beta \ell_t(0)$ gives a Neyman orthogonalized score

$$\frac{1}{n} \sum_{j=1}^{n} \int_0^t G_{j,s} \, dM_{j,s}.$$ 

This score is linear in $\lambda$, which is used in supplementary Section F to show that it satisfies the Neyman orthogonality condition under $H_0$; cf. Definition 2.1 in Chernozhukov et al. (2018). In this section, it is also shown that the act of replacing $X_t$ with $G_t = X_t - \Pi_t$ can, in fact, be viewed as concentrating out the intensity of the score statistic in the sense of Newey (1994).

While Neyman orthogonality is never invoked explicitly, it is implicitly a central part of the asymptotic results for the LCM estimator (in particular Lemma A.7 in the supplement).

The perspective on the LCM from a Neyman orthogonalized score statistic suggests that a test based on the LCM has most power against alternatives in the one-parameter family $\lambda^\beta$. If it happens that the most important alternatives are of the form $\lambda^\beta_t = e^{\beta \overline{X}_t \lambda_t}$ for some $\mathcal{G}_t$-predictable process $\overline{X}_t$ different from $X_t$, then we should replace $X_t$ by $\overline{X}_t$ in our test statistic, that is, in the residualization procedure. Examples of processes $\overline{X}_t$ are

- transformations, $\overline{X}_t = f(X_t)$ for a function $f$.
- time-shifts, $\overline{X}_t = X_{t-s}$ for $s > 0$.
- linear filters, $\overline{X}_t = \int_0^t \kappa(t-s)X_s \, ds$ for a kernel $\kappa$.
- nonlinear filters, $\overline{X}_t = \phi(\int_0^t \kappa(t-s)f(X_s) \, ds)$ for a kernel $\kappa$ and functions $f$ and $\phi$.

Any finite number of such processes could, of course, also be combined into a vector process, and we could, indeed, generalize the LCM estimator (13) to a vector process. The generalization is straightforward.

### 3.2. Survival time with time-independent covariates.

A different perspective on the test statistic is obtained if $X$ is constant over time, if $N_t = 1(T \leq t)$ is the counting process of a survival time $T$, and if $F_t = \sigma(N_s, Z; s \leq t)$ where $Z$ is a vector of additional baseline variables. Then the $F_t$-intensity is

$$\lambda_t = 1(T \geq t)h(t, Z),$$

where $h(t, Z)$ is the hazard function for $T$ given the baseline $Z$. In this special case, the hypothesis of conditional local independence is equivalent to the ordinary conditional independence

$$(14) \quad X \perp\!\!\!\!\perp T | Z.$$ 

We also find that

$$\gamma_t = \mathbb{E}(X(1(T \leq t) - \Lambda_{t,T})), $$

and in particular $\gamma_1 = \mathbb{E}(X(1 - \Lambda_T))$ as $T \in [0, 1]$ by assumption. Since $\Lambda_T$ is exponentially distributed with mean 1, we may write

$$\gamma_1 = -\text{cov}(X, \Lambda_T).$$
Testing if $\gamma_1 \neq 0$ in this particular setup is effectively a test of the conditional independence (14). When (14) is true, it further holds that $\Pi_t = \mathbb{E}(X|\mathcal{F}_{t-}) = \mathbb{E}(X|Z) = \Pi_0$ is independent of $t$, and if we incorporate this into our model of $\Pi$, the LCM estimator of $\gamma_1$ equals

$$\hat{\gamma}_1^{(n)} = \frac{1}{|J_n|} \sum_{j \in J_n} (X_j - \hat{\Pi}_{j,0})(1 - \hat{\Lambda}_{T_j}).$$

This is a (nonnormalized) generalized covariance measure (GCM), see (Shah and Peters (2020)), which is simply the (negative) empirical covariance between the additive residuals $X_j - \hat{\Pi}_{j,0}$ and the Cox–Snell residuals $\hat{\Lambda}_{T_j}$.

Alternatively, consider the quantile residual process $G_{j,t} = F_t(X_j) - \frac{1}{2}$ where $F_t(x) = \mathbb{P}(X \leq x|\mathcal{F}_{t-})$. If (14) is true, it holds again that $F_t(x) = F_0(x) = \mathbb{P}(X \leq x|Z)$ is independent of $t$, and our LCM estimator becomes

$$\hat{\gamma}_1^{(n)} = \frac{1}{|J_n|} \sum_{j \in J_n} \hat{G}_{j,0}(1 - \hat{\Lambda}_{T_j}).$$

This is likewise an empirical covariance, but now between the generalized residuals and the Cox–Snell residuals. This is closely related to the partial copula between $X$ and $T$ given $Z$, which can be estimated as

$$\frac{1}{|J_n|} \sum_{j \in J_n} \hat{G}_{j,0} \left( \frac{1}{2} - \exp(-\hat{\Lambda}_{T_j}) \right).$$

See Petersen and Hansen (2021) for further details on the partial copula and how this statistic can be used to test conditional independence.

Under a combined rate condition on estimation of $G$ and $\Lambda$, the endpoint statistics above are known to be asymptotically Gaussian with mean zero when the hypothesis of conditional independence in (14) holds. Within this survival setting, the endpoint statistics (15) can furthermore be seen as a score test derivable from a semiparametric efficient score function. Section E in the supplement gives the details for two specific semiparametric survival models.

Whenever $\hat{G}_{j,t} = \hat{G}_{j,0}$ is independent of time, for example, if we incorporate (14) into the residual model, the $t$-indexed LCM estimator is

$$\hat{\gamma}_t^{(n)} = \frac{1}{|J_n|} \sum_{j \in J_n} \hat{G}_{j,0}(\mathbb{I}(T_j \leq t) - \hat{\Lambda}_{T_j \wedge t}),$$

which can be seen as a $t$-indexed extension of (16). For a general, time-dependent residual process, the full $t$-indexed LCM estimator is

$$\hat{\gamma}_t^{(n)} = \frac{1}{|J_n|} \sum_{j \in J_n} \mathbb{I}(T_j \leq t) \hat{G}_{j,T_j} - \int_0^{t \wedge T_j} \hat{G}_{j,s} \hat{\lambda}_s \, ds.$$

The general results of this paper show that the $t$-indexed LCM estimator is asymptotically distributed as a mean zero Gaussian martingale under $H_0$. This appears to be a novel result even when $X$ is constant over time. However, the main contributions of this paper is to the case where $X$ and $Z$ are stochastic processes varying with time—where the hypothesis of conditional local independence is also distinct from (14).
4. General asymptotic results. In this section, we derive uniform asymptotic results regarding the general LCM estimator as a stochastic process. In Section 5, we discuss how to construct tests of $H_0$ based on the asymptotic results.

We assume that $N$ has a $G_t$-intensity $\lambda_t$, we let $\Lambda_t = \int_0^t \lambda_s \, ds$ denote the $G_t$-compensator of $N$ and let $M_t = N_t - \Lambda_t$ be the compensated local $G_t$-martingale. We also recall that $\hat{\gamma}^{(n)}$ denotes the LCM estimator based on sample splitting as defined in Section 2.3. Within this framework, we consider the decomposition

$$\sqrt{|J_n|} \hat{\gamma}^{(n)} = U^{(n)} + R_1^{(n)} + R_2^{(n)} + R_3^{(n)} + D_1^{(n)} + D_2^{(n)},$$

where the processes $U^{(n)}$, $R_1^{(n)}$, $R_2^{(n)}$, $R_3^{(n)}$, $D_1^{(n)}$ and $D_2^{(n)}$ are given by

$$U_t^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t G_{j,s} \, dM_{j,s},$$

$$R_{1,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t G_{j,s} (\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}) \, ds,$$

$$R_{2,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t (\hat{\gamma}_{j,s}^{(n)} - G_{j,s}) (\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}) \, ds,$$

$$R_{3,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t (\hat{\gamma}_{j,s}^{(n)} - G_{j,s}) (\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}) \, ds,$$

$$D_{1,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t G_{j,s} (\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}) \, ds,$$

$$D_{2,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t (\hat{\gamma}_{j,s}^{(n)} - G_{j,s}) (\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}) \, ds.$$

Note that the processes $D_1$ and $D_2$ are (almost surely) the zero-process under $H_0$, since the null is equivalent to $\lambda_t$ being a version of $\lambda_t$. We proceed to show that

- the processes $U^{(n)}$ and $D_1^{(n)} - \sqrt{|J_n|} \gamma$ each converge in distribution,
- and the processes $R_1$, $R_2$, $R_3$, $D_2$ converge to the zero process.

For the analysis of each of $R_1$, $R_2$ and $D_2$, sample splitting is used to render the summands conditionally independent.

These asymptotic properties imply that $\sqrt{|J_n|} (\hat{\gamma}^{(n)} - \gamma)$ is stochastically bounded in general, so the LCM estimator will asymptotically detect if the LCM is nonzero. Moreover, it will follow that $U^{(n)}$ drives the asymptotic limit of the LCM estimator under $H_0$. Based on these general asymptotic results, we derive in Section 5 asymptotic error control for tests based on the LCM estimator.

4.1. Asymptotics of the LCM estimator. Our asymptotic results are formulated in terms of uniform stochastic convergence, which has also been discussed extensively in the recent literature on hypothesis testing (Lundborg et al. (2022), Lundborg, Shah and Peters (2022), Neykov, Balakrishnan and Wasserman (2021), Scheidegger, Hörmann and Bühlmann (2022), Shah and Peters (2020)). Uniform convergence allows us to establish uniform asymptotic level of our proposed test, as well as power under local alternatives. We have collected key definitions and results related to uniform convergence in Section B in the supplement.
To state uniform assumptions and asymptotic results, we need to indicate a range of possible sampling distributions for which the assumptions apply and the results hold. For this purpose, we extend our setup and allow all data to be parametrized by a fixed parameter set $\Theta$. The set $\Theta$ is not a priori assumed to have any structure, and $\theta \in \Theta$ simply indicates that $N^\theta$, $X^\theta$, $\lambda^\theta$, $G^\theta$, etc. have $\theta$-dependent distributions. We generally denote evaluation of processes or derived quantities for a specific $\theta$-value by a superscript, with the LCM, $\gamma^\theta$, in particular, depending on $\theta$. The LCM estimator is likewise written as $\hat{\gamma}^{(n),\theta} = (\hat{\gamma}^{(n),\theta}_t)$ for $\theta \in \Theta$ to denote its dependence on the sampling distribution. The superscript notation is, however, heavy and unnecessary in many cases and we will suppress the dependency on $\theta \in \Theta$ whenever it is not needed. Any result that does not explicitly involve $\Theta$ should be understood as a pointwise result for each $\theta \in \Theta$.

The parametrization allows us to express convergence in distribution and probability uniformly over $\Theta$, which are denoted by $\overset{D}{\longrightarrow}_{\Theta}$ and $\overset{P}{\longrightarrow}_{\Theta}$, respectively. These concepts are defined rigorously in Definition B.2 in the supplement. We note that uniform convergence reduces to classical (pointwise) convergence if $\Theta$ is a singleton, which corresponds to fixing the sampling distribution. We also introduce the parameter subset

\[(24) \quad \Theta_0 := \{ \theta \in \Theta | H_0 \text{ is valid} \},\]

consisting of all parameter values for which the hypothesis of conditional local independence holds. Correspondingly, we will use $\overset{D}{\longrightarrow}_{\Theta_0}$ and $\overset{P}{\longrightarrow}_{\Theta_0}$ to denote stochastic convergences uniformly over $\Theta_0$.

We are now ready to formulate the underlying assumptions on the data required for our asymptotic results. These assumptions may appear strong, but we argue in the discussion in Section 7 that they are not unreasonable from a practical viewpoint.

**Assumption 4.1.** There exist constants $C, C' > 0$, such that for any parameter value $\theta \in \Theta$:

(i) The $G^\theta_t$-intensity $\lambda_t^\theta$ of $N^\theta$ is càglàd with $\sup_{0 \leq t \leq 1} \lambda_t^\theta \leq C$ almost surely.

(ii) The residual process $G^\theta$ is càglàd with $\sup_{0 \leq t \leq 1} |G^\theta_t| \leq C'$ almost surely.

The estimator, $\hat{\lambda}^{(n)}_t$, of $\lambda_t$ and the estimator, $\hat{G}^{(n)}_t$, of the residual process are assumed to satisfy the same bounds as $\lambda_t$ and $G_t$. We note that Assumption 4.1(i) implies that $M_t$ is a true $G_t$-martingale, and by the innovation theorem, $\lambda_t = E(\lambda_t | F_{t-})$. As a consequence, the $F_t$-intensity $\lambda_t$ inherits the boundedness from the $G_t$-intensity $\lambda_t$, and $M_t$ is an $F_t$-martingale. More generally, we have the following proposition ensuring that stochastic integrals are true martingales, for example, that $I_t$ is a martingale under $H_0$.

**Proposition 4.1.** Under Assumption 4.1, it holds that each of the processes

\[\left( \int_0^t f(G_s) \, dM_s \right)_{t \in [0, 1]} \quad \text{and} \quad \left( \int_0^t f(\hat{G}^{(n)}_s) \, dM_s \right)_{t \in [0, 1]}\]

are mean zero, square integrable $G_t$-martingales for any $f \in C(\mathbb{R})$.

To express the asymptotic distribution of $U^{(n)}$, we need its variance function.

**Definition 4.2.** We define the variance function $\mathcal{V} : [0, 1] \to [0, \infty]$ as

\[(25) \quad \mathcal{V}(t) = E\left( \int_0^t G^2_s \, dN_s \right).\]
As everything else, the variance function, \( \mathcal{V} = \mathcal{V}^\theta \), is also indexed by the parameter \( \theta \), which we, for notational simplicity, suppress unless explicitly needed.

By taking \( f(x) = x^2 \) in Proposition 4.1, Assumption 4.1 implies that for each \( t \in [0, 1] \),

\[
\mathcal{V}(t) = \mathbb{E}\left( \int_0^t G_s^2 \lambda_s \, ds \right) < \infty.
\]

Moreover, \( \mathcal{V}(t) \) is the variance of \( \int_0^t G_s \, dM_s \), which under \( H_0 \) is the same as the variance of \( I_t = \int_0^t G_s \, dM_s \).

With the assumptions above, we can prove the following proposition about the uniform distributional limit of the process \( U(n) \) in the Skorokhod space \( D[0, 1] \), the space of càdlàg functions from \([0, 1]\) to \( \mathbb{R} \) endowed with the Skorokhod topology. A corresponding pointwise result is an application of Rebolledo’s classical martingale CLT. Our generalization to uniform convergence is based on a uniform extension of Rebolledo’s theorem; see Theorem C.4 in Section C in the supplement.

**Proposition 4.3.** Under Assumption 4.1, it holds that

\[
U(n, \theta) \overset{D/\Theta}{\rightarrow} U^\theta
\]

in \( D[0, 1] \) as \( n \rightarrow \infty \), where for each \( \theta \in \Theta \), \( U^\theta \) is a mean zero continuous Gaussian martingale on \([0, 1]\) with variance function \( \mathcal{V}^\theta \).

To control the remainder terms in (17), we will bound the estimation errors in terms of the 2-norm, \( \| \cdot \|_2 \), on \( L_2([0, 1] \times \Omega) \), that is,

\[
\| W \|_2^2 = \mathbb{E}\left( \int_0^1 W_s^2 \, ds \right)
\]

for any process \( W \in L_2([0, 1] \times \Omega) \). We will make the following consistency assumptions on \( \hat{\lambda}^{(n)} \) and \( \hat{G}^{(n)} \).

**Assumption 4.2.** Assume that \( |J_n| \rightarrow \infty \) when \( n \rightarrow \infty \) and let

\[
g^\theta(n) = \left\| G^\theta - \hat{G}^{(n), \theta} \right\|_2 \quad \text{and} \quad h^\theta(n) = \left\| \lambda^\theta - \hat{\lambda}^{(n), \theta} \right\|_2.
\]

Then each of the sequences \( g^\theta(n) \), \( h^\theta(n) \), and \( \sqrt{|J_n|} g^\theta(n) h^\theta(n) \) converge to zero uniformly over \( \Theta \) as \( n \rightarrow \infty \), that is,

\[
\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \max\{g^\theta(n), h^\theta(n), \sqrt{|J_n|} g^\theta(n) h^\theta(n)\} = 0.
\]

With this assumption, we can establish that the remainder terms also converge uniformly to the zero process.

**Proposition 4.4.** Under Assumptions 4.1 and 4.2, it holds that

\[
\sup_{t \in [0, 1]} \left| R_{i, t}^{(n), \theta} \right| \overset{P/\Theta}{\rightarrow} 0
\]

as \( n \rightarrow \infty \) for \( i = 1, 2, 3 \).

To control the asymptotic behavior of the LCM estimator in the alternative, we need to control the two terms \( D_1^{(n)} \) and \( D_2^{(n)} \).
Proposition 4.5. Let Assumptions 4.1 and 4.2 hold true.

(i) The stochastic process \( D_{1(n), \theta} - \sqrt{|J_n|} \gamma \) converges in distribution in \((C[0,1], \| \cdot \|_\infty)\) uniformly over \( \Theta \) as \( n \to \infty \).

(ii) If \( G_i^\theta = X_i^\theta - \Pi_i^\theta \) is the additive residual process, then \( D_{2(n), \theta} \xrightarrow{p/\Theta} 0 \) in \( D[0,1] \) as \( n \to \infty \).

We note that \( D_{2(n)} \) might not vanish without an assumption like \( G_t \) being the additive residual process, and it is not clear if \( D_{2(n)} \) will even converge in general. We will not pursue an analysis of the asymptotic behavior of \( D_{2(n)} \) in the general case. We note, however, that if we can estimate \( G \) with a parametric rate, that is, \( \sqrt{|J_n|} g(n) = O(1) \). Then it follows from the Cauchy–Schwarz inequality that \( D_{2(n)} \) is stochastically bounded, and \( D_{1(n)} \) still dominates in the alternative where \( \gamma \neq 0 \).

We can combine all of the propositions into a single theorem regarding the asymptotics of the LCM estimator, which we consider as our main result.

Theorem 4.6. Let Assumptions 4.1 and 4.2 hold true.

(i) It holds that
\[
\sqrt{|J_n|} \hat{\gamma}^{(n), \theta} \xrightarrow{D/\Theta_0} U^\theta
\]
in \( D[0,1] \) as \( n \to \infty \), where for each \( \theta \in \Theta_0 \), \( U^\theta \) is a mean zero continuous Gaussian martingale on \([0,1]\) with variance function \( V^\theta \).

(ii) For the additive residual process, it holds that for every \( \varepsilon > 0 \) there exists \( K > 0 \) such that
\[
\limsup_{n \to \infty} \sup_{\theta \in \Theta} \mathbb{P}(\sqrt{|J_n|} \cdot \| \hat{\gamma}^{(n), \theta} - \gamma^\theta \|_\infty > K) < \varepsilon.
\]

Thus, we have established the weak asymptotic limit of \( \sqrt{|J_n|} \hat{\gamma}^{(n)} \) under \( H_0 \). However, the variance function \( V \) of the limiting Gaussian martingale is unknown and must be estimated from data. We propose to use the empirical version of (25),
\[
\hat{V}_n(t) = \frac{1}{|J_n|} \sum_{j \in J_n} \int_0^t \left( \hat{G}_{j,s}^{(n)} \right)^2 dN_{j,s} = \frac{1}{|J_n|} \sum_{j \in J_n} \sum_{\tau \leq t: \Delta N_{j,s} = 1} \left( \hat{G}_{j,\tau}^{(n)} \right)^2,
\]
for which we have the following consistency result.

Proposition 4.7. Under Assumptions 4.1 and 4.2, it holds that
\[
\sup_{t \in [0,1]} | \hat{V}_n^\theta(t) - V^\theta(t) | \xrightarrow{p/\Theta} 0,
\]
as \( n \to \infty \).

We emphasize that \( V \) is only the asymptotic variance function of the LCM estimator under \( H_0 \). It is always the asymptotic variance function of \( U^{(n)} \), but in the alternative the asymptotic distribution of \( \hat{\gamma}^{(n)} \) also involves the asymptotic distribution of \( D_{1(n)} \) and is thus more complicated.

Tests of conditional local independence can now be constructed in terms of univariate functionals of \( \hat{\gamma}^{(n)} \) and \( \hat{V}_n \) that quantify the magnitude of the LCM. The asymptotics of such test statistics under \( H_0 \) are described in the following corollary, which is essentially an application of the continuous mapping theorem.
**Corollary 4.8.** Let $J : D[0, 1] \times D[0, 1] \to \mathbb{R}$ be a functional that is continuous on the closed subset $C[0, 1] \times \{ \mathcal{V}_\theta : \theta \in \Theta_0 \}$ with respect the uniform topology, that is, the topology generated by the norm $\| (f_1, f_2) \| = \max \{ \| f_1 \|_\infty, \| f_2 \|_\infty \}$ for $f_1, f_2 \in D[0, 1]$. Define the test statistic

$$\hat{D}_n = J(\sqrt{|J_n|}, \hat{\gamma}(n), \hat{\mathcal{V}}_n).$$

Under Assumptions 4.1 and 4.2, it holds that

$$\hat{D}_n \overset{D/\sigma_0}{\to} J(U^\theta, \mathcal{V}^\theta), \quad n \to \infty,$$

where $U^\theta$ is a mean zero continuous Gaussian martingale with variance function $\mathcal{V}^\theta$.

---

**5. The local covariance test.** In this section, we introduce a practically applicable test based on the LCM estimator. Using the asymptotic distribution of the LCM estimator, we show that the asymptotic distribution of our proposed test is independent of the sampling distribution under $H_0$ and has an explicit representation. We show, in addition, uniform asymptotic level of the test, and we give a uniform power result for the additive residual process. Finally, we modify the test to be based on a cross-fitted estimator of the LCM instead of using sample splitting, and we show uniform level of that test.

To construct a test statistic based on the LCM estimator, it is beneficial that its distributional limit does not depend on the variance function. As a simple example, consider the endpoint test statistic:

$$\hat{\gamma}(n)(1)$$

which under $H_0$ converges in distribution to $\mathcal{V}(1)^{-\frac{1}{2}} U_1$ by Corollary 4.8. The distribution of the latter is the standard normal distribution, and in particular it does not depend on $\mathcal{V}$.

Any test statistic constructed from $\hat{\gamma}(n)$ should capture deviations of $\gamma_t$ away from 0. The test statistic in (29) does, however, only consider the endpoint of the process, and since $\gamma$ is not necessarily monotone, $\gamma_t$ may deviate more from 0 for other $t \in [0, 1]$. Thus, in order to increase power against such alternatives we consider the test statistic

$$\hat{T}_n = \frac{\sqrt{|J_n|}}{\hat{\mathcal{V}}_n(1)} \sup_{0 \leq t \leq 1} |\hat{\gamma}(n)|.$$

We refer to $\hat{T}_n$ as the Local Covariance Test statistic (LCT statistic). We proceed to show that the LCT statistic can be turned into a test of $H_0$ with asymptotic level $\alpha$, and which has asymptotic power against any alternative with a nonzero LCM. This is the best we can hope for of any test based on the LCM estimator.

We note that it might be possible to establish similar results for other norms of the LCM, for example, a statistic based on a weighted $L_2$-norm. However, since other norms of the distributional limit $U$ will generally have a distribution with a complicated dependency on $\mathcal{V}$, we believe that the LCT statistic is the simplest to construct.

To establish uniform asymptotic level via Corollary 4.8 for tests based on test statistics such as (30), we need to assume that the asymptotic variances in $t = 1$ are uniformly bounded away from zero.

**Assumption 5.1.** There exists a $\delta_1 > 0$ such that for all $\theta \in \Theta$ it holds that $\mathcal{V}^\theta(1) \geq \delta_1$. 
5.1. Type I and type II error control. We proceed to show that under $H_0$, the LCT statistic is distributed as $\sup_{0 \leq t \leq 1} |B_t|$, where $(B_t)$ is a standard Brownian motion. From this point onward, we let $S$ denote a random variable with such a distribution and note that its CDF can be written as

\[ F_S(x) = \mathbb{P}(S \leq x) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left(-\frac{\pi^2(2k+1)^2}{8x^2}\right), \quad x > 0. \]  

(31)

See, for example, Section 12.2 in Schilling and Partzsch (2014) where the formula is derived from Lévy’s triple law.

The $p$-value for a test of $H_0$ equals $1 - F_S(\hat{T}_n)$, and since the series in (31) converges at an exponential rate, the $p$-value can be computed with high numerical precision by truncating the series.

Given a significance level $\alpha \in (0, 1)$, we set $z_1 - \alpha$ denote the $(1 - \alpha)$-quantile of $F_S$, which exists and is unique since the right-hand side of (31) is strictly increasing and continuous. The Local Covariance Test (LCT) with significance level $\alpha$ is then defined by

\[ \Psi_n = \Psi_n^{\alpha} = \mathbb{1}(F_S(\hat{T}_n) > 1 - \alpha) = \mathbb{1}(\hat{T}_n > z_1 - \alpha). \]  

(32)

From Theorem 4.6, we can now deduce the asymptotic properties of the LCT under the hypothesis of conditional local independence. Recall that $\xrightarrow{D/\Theta_0}$ denotes uniform convergence in distribution under $H_0$.

**Theorem 5.1.** Let Assumptions 4.1, 4.2 and 5.1 hold true. Then it holds that

\[ \hat{T}_n \xrightarrow{D/\Theta_0} S \]

as $n \to \infty$. As a consequence, for any $\alpha \in (0, 1)$,

\[ \limsup_{n \to \infty} \sup_{\theta \in \Theta_0} \mathbb{P}(\Psi_n^{\alpha, \theta} = 1) \leq \alpha. \]

In other words, the local covariance test defined in (32) has uniform asymptotic level $\alpha$.

In general, we cannot expect that the test has power against alternatives to $H_0$ for which the LCM is the zero function. This is analogous to other types of conditional independence tests based on conditional covariances, for example, GCM (Shah and Peters (2020)). However, we do have the following result that establishes power against local alternatives with $\|\gamma\|_{\infty}$ decaying at an order of at most $|J_n|^{-1/2}$.

**Theorem 5.2.** Let Assumptions 4.1 and 4.2 hold true. Using the additive residual process it holds that for any $0 < \alpha < \beta < 1$, there exists $c > 0$ such that

\[ \liminf_{n \to \infty} \inf_{\theta \in \Theta_0} \mathbb{P}(\Psi_n^{\alpha, \theta} = 1) \geq \beta, \]

where $\mathcal{A}_{c,n} = \{\theta \in \Theta | \|\gamma^\theta\|_{\infty} \geq c |J_n|^{-1/2}\}$.

5.2. Extension to cross-fitting. In Section 4, we considered sample splitting with observations indexed by $J_n^c$ used to estimate the two models and with observations indexed by $J_n$ used to estimate $\gamma$. Following Chernozhukov et al. (2018), we can improve efficiency by cross-fitting, that is, by flipping the roles of $J_n$ and $J_n^c$ to obtain a second equivalent estimator of $\gamma$. Heuristically, the two estimators are approximately independent, and thus their average should be a more efficient estimator. This procedure generalizes directly to a partition $J_n^1 \cup \cdots \cup J_n^K = \{1, \ldots, n\}$ of the indices into $K$ disjoint folds. The partition is assumed to have a uniform asymptotic density, meaning that $|J_n^k|/n \to 1/K$ as $n \to \infty$ for each $k$. 
We estimate $G$ and $\lambda$ using $(J^k_n)^c = \{1, \ldots, n\} \setminus J^k_n$ and subsequently estimate $\gamma$ using $J^k_n$. Then the $K$-fold Cross-fitted LCM estimator, abbreviated as X-LCM, is defined as the average LCM estimator over the $K$ folds, that is,

$$\hat{\gamma}^K_{t,(n)} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|J^k_n|} \sum_{j \in J^k_n} \int_0^t \hat{G}_{j,s}^{k,(n)} \, d\hat{M}_{j,s}^{k,(n)}, \quad (33)$$

where for each $j \in J^k_n$, the processes $\hat{G}_{j,s}^{k,(n)}$ and $\hat{M}_{j,s}^{k,(n)}$ are the model predictions of $G_j$ and $M_j$, respectively, based on training data indexed by $(J^k_n)^c$. We also define a $K$-fold version of the variance estimator:

$$\hat{V}^K_{t,(n)} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|J^k_n|} \sum_{j \in J^k_n} \int_0^t \left(\hat{G}_{j,s}^{k,(n)}\right)^2 \, dN_{j,s}. \quad (34)$$

Now, similar to the LCT statistic, the cross-fitted estimator can be used to construct a test statistic,

$$\hat{T}^K_n = \sqrt{\frac{n}{\hat{V}_{K,n}(1)}} \sup_{0 \leq t \leq 1} |\hat{\gamma}^K_{t,(n)}|, \quad (35)$$

from which we define the following test of conditional local independence.

**Definition 5.3.** Let $\alpha \in (0, 1)$ and let $\hat{T}^K_n$ be the test statistic from (35). The $K$-fold Cross-fitted Local Covariance Test (X-LCT) with significance level $\alpha$ is defined by

$$\Psi^K_n = \mathbb{1}(F_S(\hat{T}^K_n) > 1 - \alpha) = \mathbb{1}(\hat{T}^K_n > z_{1-\alpha}),$$

where $z_{1-\alpha}$ is the $(1 - \alpha)$-quantile of the distribution function $F_S$ given in (31).

We provide a summary of the computation of the X-LCT in Algorithm 2. The asymptotic analysis of $\hat{\gamma}^{(n)}$ generalizes to $\hat{\gamma}^K_{t,(n)}$, but we will refrain from restating all results for the $K$-fold cross-fitted estimator. For simplicity, we focus on the fact that the X-LCT has asymptotic level $\alpha$.

**Theorem 5.4.** Suppose that Assumption 4.2 is satisfied for every sample split $J^k_n \cup (J^k_n)^c$, $k = 1, \ldots, K$. Under Assumptions 4.1 and 5.1, the X-LCT statistic satisfies

$$\hat{T}^K_n \overset{D}{\rightarrow} S$$

for $n \to \infty$. In particular, the X-LCT has uniform asymptotic level $\alpha$.

Note that cross-fitting recovers full efficiency in the sense that the scaling factor is $\sqrt{n}$ rather than $\sqrt{|J^k_n|}$, which leads to a more powerful test. Moreover, the asymptotic distribution of $\hat{T}^K_n$ does not depend on the number of folds $K$, and any difference between various choices of $K$ can thus be attributed to finite sample errors. Larger values of $K$ will allocate more data to estimation of $G$ and $\lambda$, which intuitively should be the harder estimation problem. Following Remark 3.1 in Chernozhukov et al. (2018), we believe that a default choice of $K = 4$ or $K = 5$ should be reasonable in practice.
Algorithm 2: $K$-fold cross-fitted local covariance test (X-LCT)

1 input: processes $(N_j, X_j, Z_j)_{j=1,...,n}$, partition $J_1^n \cup \cdots \cup J_K^n$ of indices into $K$ folds;
2 options: historical regression methods for estimation of $\lambda$ and $G$ given $N$ and $Z$,
3 discrete time grid $\mathbb{T} \subset [0, 1]$, significance level $\alpha \in (0, 1)$;
4 begin
5 for $k = 1, \ldots, K$ do
6 apply Algorithm 1 on sample split $J_k^n \cup (J_k^n)^c$ to compute $\hat{\gamma}_{k,(n)}$ on grid $\mathbb{T}$;
7 use equation (27) on sample split $J_k^n \cup (J_k^n)^c$ to compute $\hat{\mathcal{V}}_{k,n}(1)$;
8 end
9 compute $\hat{\gamma}_{K,(n)} = \frac{1}{K} \sum_{k=1}^{K} \hat{\gamma}_{k,(n)}$ on grid $\mathbb{T}$;
10 compute $\hat{\mathcal{V}}_{K,n}(1) = \frac{1}{K} \sum_{k=1}^{K} \hat{\mathcal{V}}_{k,n}(1)$;
11 compute the X-LCT statistic $\hat{T}_{n}^K = \sqrt{n} \cdot \max_{t \in \mathbb{T}} |\hat{\gamma}_{n}^{K,(n)}(t)| / \sqrt{\hat{\mathcal{V}}_{K,n}(1)}$;
12 compute $p$-value $\hat{p} = 1 - F_S(\hat{T}_{n}^K)$ by truncating the series in equation (31);
13 end
14 output: the X-LCT $\hat{\Psi}^K = 1(\hat{p} < \alpha)$, and optionally the $p$-value $\hat{p}$;

6. Simulation study. In this section, we present the results from a simulation study based on the Cox example introduced in Section 2.2. We elaborate in Section 6.1 on the full model specification used for the simulation study, which will also illuminate how $\Pi$ and $\lambda$ can be modeled and estimated. The results from the simulation study focus on the distribution of the X-LCT statistic $\hat{T}_{n}^K$ and validate the asymptotic level and power of the X-LCT $\hat{\Psi}^K$. The latter is also compared to a hazard ratio test based on the marginal Cox model (11). The simulations were implemented in Python and the code is available.$^1$

6.1. Cox model continued. Consider the same setup as in Section 2.2. To fully specify the model, we need to specify the distribution of the processes $X$, $Y$ and $Z$. We suppose that $X$ and $Y$ can be written in terms of $Z$ as

$$
X_t = \int_0^t Z_s \rho_X(s,t) \, ds + V_t \quad \text{and} \quad Y_t = \int_0^t Z_s \rho_Y(s,t) \, ds + W_t,
$$

where $\rho_X$ and $\rho_Y$ are two functions defined on the triangle $\{(s,t) \in [0,1]^2 \mid s \leq t\}$, and where $V = (V_t)_{0 \leq t \leq 1}$ and $W = (W_t)_{0 \leq t \leq 1}$ are two noise processes with mean zero. The processes $Z$, $V$ and $W$ are assumed independent, which implies (10), and thus that $N$ is conditionally locally independent of $X$ given $\mathcal{F}_t = \mathcal{F}_t^{N,Z}$.

The specific dependency of $X$ and $Y$ on $Z$ is known as the historical functional linear model in functional data analysis (Malfait and Ramsay (2003)). Within this model,

$$
\Pi_t = \mathbb{E}(X_t | \mathcal{F}_{t-}) = \int_0^t Z_s \rho_X(s,t) \, ds,
$$

and on $(T \geq t)$,

$$
\mathbb{E}(e^{Y_t} | \mathcal{F}_t) = e^{\int_0^t Z_s \rho_Y(s,t) \, ds} \mathbb{E}(e^{W_t} | T \geq t) = e^{\tilde{\beta}_0(t) + \int_0^t Z_s \rho_Y(s,t) \, ds},
$$

\footnote{https://github.com/AlexanderChristgau/nonparametric-cli-test.}
where $\tilde{\beta}_0(t) = \log(\mathbb{E}(e^{W_t} \mid T \geq t))$. Since $\lambda_t = \mathbb{1}(T \geq t)\lambda_0^0 e^{\beta Z_t} \mathbb{E}(e^{Y_t} \mid F_t)$, it follows that on $(T \geq t)$,

$$\log(\lambda_t) = \beta_0(t) + \beta Z_t + \int_0^t Z_s \rho_Y(s, t) \, ds,$$

where the two baseline terms depending only on time have been merged into $\beta_0$.

The computations above suggest how the estimators $\hat{\lambda}^{(n)}$ and $\hat{\Pi}^{(n)}$ could be constructed. That is, $\hat{\lambda}^{(n)}$ could be based on estimates of $\beta$, $\beta_0$ and $\rho_Y$ from the observations $(T_j, Z_j)_{j \in J_n}$, and $\hat{\Pi}^{(n)}$ could be based on estimates of $\rho_X$ from $(X_j, Z_j)_{j \in J_n}$. We would then have

$$\hat{\Pi}^{(n)}_{j,t} = \int_0^t Z_{j,s} \hat{\rho}^{(n)}_X(s, t) \, ds$$

for $j \in J_n$ where $\hat{\rho}^{(n)}_X$ denotes the estimate of $\rho_X$, and similarly for $\hat{\lambda}^{(n)}$. Particular choices of estimators $\hat{\rho}^{(n)}_X$ and $\hat{\rho}^{(n)}_Y$ and their theoretical properties are reviewed in Section D in the supplement. Our conclusion from this review is that for the historical functional linear model, sufficient rate results should be possible but have not yet been established rigorously.

6.2. Sampling scheme. The actual time-discretized simulations and computations were implemented using an equidistant grid $T = (t_i)_{i=1}^q$ with $q = 128$ time points $0 = t_1 < \cdots < t_q = 1$. Inspired by Harezlak et al. (2007), we generated the processes as follows: let $\xi \in \mathbb{R}^3$ and $V, W, W \in \mathbb{R}^T$ be independent random variables such that $\xi \sim \mathcal{N}(0, I_3)$ and such that $V, W$ and $W$ are identically distributed with $V_{t_1}, V_{t_2} - V_{t_1}, \ldots, V_{t_q} - V_{t_{q-1}} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/q)$. Then the process $Z$ is determined by $Z_t = \xi_1 + \xi_2 t + \sin(2\pi \xi_3 t) + W_t$ for $t \in T$. The processes $X$ and $Y$ were then given by the historical linear model (36) with kernels $\rho_X$ and $\rho_Y$ being one of the following four kernels:

- zero: $(s, t) \mapsto 0$,
- constant: $(s, t) \mapsto 1$,
- Gaussian: $(s, t) \mapsto e^{-2(t-s)^2}$,
- sine: $(s, t) \mapsto \sin(4t - 20s)$.

To compute $X$ and $Y$, we evaluated the kernels on $(s, t) \in T^2 | s \leq t$ and approximated the integrals by Riemann sums. The full intensity for $N_t = \mathbb{1}(T \leq t)$ was specified with a Weibull baseline of the form $\lambda_t^{\text{full}} = \mathbb{1}(T \geq t)\beta_1 t^2 \exp(\beta_2 Z_t + Y_t)$, for $\beta_1 > 0$ and a choice of $\beta_2 \in \{-1, 1\}$. To sample $T$, we applied the inverse hazard method, which utilizes that $\Lambda_t^{\text{full}}$ is standard exponentially distributed. That is, we sampled $E \sim \text{Exp}(1)$ and numerically computed $T = \max\{t \in T | \Lambda_t^{\text{full}} < E\}$ as a discretized approximation. For any given parameter setting, the baseline coefficient $\beta_1$ was chosen sufficiently large to ensure that $\Lambda_t^{\text{full}} \geq E$ would occur before time $t = 1$ in more than $\frac{q-1}{q} \cdot n$ samples.

The simulation setting used to sample the data for Figures 3 and 4 was $\beta_2 = -1$ and $\rho_X = \rho_Y = \text{constant}$. With this setup, Assumption 4.1 is satisfied if $V, W$ and $W$ were bounded. Since we use the Gaussian distribution, they are technically not bounded, but they could be made bounded by introducing a lower and upper cap. Due to the light tails of the Gaussian distribution such caps would have no noticeable effect on the simulation results, and the results we report are generated without a cap.

The implementation details for the X-LCT and the hazard ratio test are given in Section G.1 in the supplement.
6.3. Distributions of p-values under $H_0$. We examine the distributional approximation $\tilde{T}_n^K \Rightarrow S$ (cf. Theorem 5.4) by comparing the p-values $1 - F_S(\tilde{T}_n^K)$ to a uniform distribution. Figure 5 shows the empirical distribution functions of the p-values computed from data simulated according to the scheme described in the previous section. The results are aggregated over the two choices of $\beta_2 \in \{-1, 1\}$ since these two settings were found to be similar. For more detailed results from the experiment see Figure G.1 in Section G in the supplement, which also includes the p-values corresponding to the endpoint test statistic.

For the hazard ratio test, Figure 5 shows that the p-values are subuniform for the zero kernel. In this case, the marginal Cox model is correct, and the nonuniformity of the p-values can be explained by the $L_2$-penalization. For the constant and Gaussian kernels, the hazard ratio test fails completely, whereas for the sine kernel, the mediated effect of $Z$ on $T$ through $Y$ is more subtle, and the model misspecification only becomes apparent for $n = 2000$. Overall, these results are consistent with the reasoning in the Section 2.2: a test based on the misspecified Cox model will wrongly reject the hypothesis of conditional local independence.

For the proposed X-LCT, Figure 5 shows that the associated p-values are slightly anti-conservative for $n = 100$. This is to be expected, and can be explained by the finite sample errors leading to more extreme values of $\tilde{T}_n^K$ than the approximation by $S$. As $n$ increases, these errors become smaller, and for $n = 2000$ the p-values actually seem to be subuniform. The subuniformity may be explained by the time discretization, since the maximum of the process is taken over $\mathbb{T}$ rather than $[0, 1]$. Figure G.3 in Section G in the supplement illustrates the asymptotic effect of the time discretization, which supports this claim. Another support of this claim is that the endpoint test does not appear to give subuniform p-values for large $n$; see Figure G.1. We finally note that the distributions of the p-values for our proposed test is largely unaffected by the kernel used to generate the data.

6.4. Power against local alternatives. To investigate the power of the X-LCT, we construct local alternatives to $H_0$ in accordance with the right graph in Figure 2 by replacing $Y_t$ by the process $Y_t + \rho_0 \sqrt{n} X_t$. That is, for $\rho_0 \neq 0$, blood pressure is then directly affected by pension savings, and $N_t$ is no longer conditionally locally independent of $X_t$ given $\mathcal{F}_t$. In
For each $\rho_0 \in \{0, 5, 10\}$, the lines show the average rejection rates of our proposed test X-LCT (blue) and the hazard ratio test (orange) as functions of sample size, with each average taken over 8 different settings. For each setting, the rejection rate is computed from 400 simulated data sets at a 5% significance level and the rejection rate is displayed with a dot.

In the leftmost panel, the data was generated under $H_0$ and the plot shows what we noted previously, namely that the X-LCT holds level for large $n$, whereas the hazard ratio test does not. For the local alternatives, $\rho_0 = 5$ and $\rho_0 = 10$, we note that the power of the hazard ratio test is quite sensitive to the simulation settings. For some settings it has no power, while for others it has some power.

In contrast, the proposed X-LCT has power against all of the local alternatives. The power increases with $n$ initially but stabilizes from around $n = 1000$. This is similar to the behavior observed under the null hypothesis and is not surprising. We expect that the sample size needs to be sufficiently large for the nonparametric estimators to work sufficiently well, and we expect the sufficient sample size to be mostly unaffected by the value of $\rho_0$. For fixed $n$, we also note that the power of $\hat{\Psi}_n^K$ is fairly robust with respect to the choice of $\beta_2$ and the choice of kernel. Overall, we find that the X-LCT is applicable in these settings with historical effects; it has consistent power against the $\sqrt{n}$ alternatives while controlling type I error for $n$ reasonably large.

In Section G.2 in the supplement, we provide additional numerical results for time-varying alternatives, and we compare the X-LCT with its endpoint counterpart.

7. Discussion. The LCM was introduced as a functional parameter that quantifies deviations from the hypothesis $H_0$ of conditional local independence. We showed how the parameter may be expressed in several ways, but that it is the representation in terms of the residual process that allows us to estimate the LCM with a $\sqrt{n}$-rate under $H_0$ without parametric model assumptions. The residual process was introduced as an abstract model of $X_t$ for each $t$ given the history up to time $t$, and we showed that such a residualization could be viewed as a form of orthogonalization. Similar ideas have been used recently for classical conditional independence testing, such as GCM (Shah and Peters (2020)), tests based on the partial copula (Petersen and Hansen (2021)) and GHCM (Lundborg, Shah and Peters...
It is, however, not possible to use any of these to test $H_0$, which cannot be expressed as a classical conditional independence. Our test based on the LCM is the first nonparametric test of conditional local independence with substantial theoretical support, and we propose to test $H_0$ in practice by using X-LCT based on the cross-fitted estimator of LCM.

Contrary to the tests of conditional independence mentioned above, we need sample splitting—even under $H_0$—to achieve our asymptotic results. We do not believe that this can be avoided. The standard argument to avoid this uses classical conditional independence in a crucial way, which does not translate into our framework—basically because we condition on information that changes with time. Our simulation study also indicates that sample splitting or cross-fitting is needed in practice for the LCM estimator to be unbiased under $H_0$.

While our cross-fitted estimator of the LCM, the X-LCM, share some of the general patterns of other double machine learning procedures—including the overall decomposition (17)—our analysis and results required a range of generalizations of known results and some novel ideas. The asymptotic distribution of the leading term, $U(n)$, is also a well-known consequence of Rebolledo’s CLT; see, for example, Section V.4 in (Andersen et al. (1993)) for related results in the context of survival analysis. However, we generalized this result to uniform convergence in the Skorokhod space $D[0,1]$, and we introduced new techniques for handling the remainder terms. These novel techniques are made necessary by the decomposition (17) being a decomposition of stochastic processes indexed by time. We outline below the three most important technical contributions we made.

First, to obtain uniform control of level and power, all asymptotic results in Section 4 are formulated in terms of uniform stochastic convergence. Since this notion of convergence had not previously been considered on general metric spaces, and especially not on the Skorokhod space, we had to develop the necessary theory. This development could be of independent interest, and we have collected the general definitions and main results on uniform stochastic convergence in metric spaces in Section B in the supplement. This framework also allowed us to show a uniform version of Rebolledo’s martingale CLT in Section C in the supplement.

Second, to establish distributional convergence under $H_0$, we need to control the remainder terms $R_{i,t}^{(n)}$ uniformly over $t$. The third term, $R_3^{(n)}$, is simple to bound, and by exploiting Doob’s submartingale inequality, the second term, $R_2^{(n)}$, can also be bounded. The most difficult first term, $R_1^{(n)}$, was controlled using stochastic equicontinuity via an exponential tail bound and the use of the chaining lemma. The necessary general uniform stochastic equicontinuity and chaining arguments are collected in Section C.3 in the supplement.

Third, to achieve rate results in the alternative, the processes $D_1^{(n)}$ and $D_2^{(n)}$ must be controlled. The process $D_1^{(n)}$ does, like $U^{(n)}$, not involve any estimation, and its distributional convergence follows from a general CLT argument for continuous stochastic processes. The term $D_2^{(n)}$ is more difficult to handle, as it may not have mean zero if $G_t$ is not the additive residual process. However, $X_t$ cancels out in $\hat{G}_t^{(n)} - G_t$ for the additive residual process, which makes the difference $\mathcal{F}_t$-predictable, and $D_2^{(n)}$ can then be bounded similar to $R_1^{(n)}$.

For a general residual process, it seems possible for $D_2^{(n)}$ to have a bias of order $\sqrt{|J_n|}g(n)$.

Our main result, Theorem 4.6, is stated under two assumptions. The second, Assumption 4.2, is a straightforward generalization to our setup of similar assumptions in the double machine learning literature on rates of convergence for the two estimators used. Both estimation errors are measured using a 2-norm, and it is plausible that we can relax one norm to a weaker form of convergence if we simultaneously strengthen the other norm. The first assumption, Assumption 4.1, requires uniform bounds on both $\lambda$ and $G$. This is a strong assumption but perhaps not particularly problematic from a practical viewpoint. Indeed, $G$ is a process we can choose, and we can thus make it bounded if necessary. And though many
theoretically interesting counting process models have unbounded intensities, a large cap on the intensity will make no difference in practice. We believe, nevertheless, that it is possible to relax Assumption 4.1 to a weaker form of control on the magnitudes of $\lambda$ and $G$ as functions of time, for example, moment bounds uniform in $\theta$. However, such a generalization will come at the expense of considerably more technical proofs, and we did not pursue this line of research.

A major practical question is whether we can estimate $\lambda$ and $G$ with sufficient rates, for example, $n^{-\frac{1}{4}+\epsilon}$. In Section D in the supplement, we give an overview of some known and some conjectured rate results for specific forms of $\lambda$ and $\Pi$. Beyond parametric models, we conclude that the existing rate results are scarce, and we regard it as an independent research project to establish rates for general historical regression methods.

Another question is whether we can replace the counting process $N$ by a more general semimartingale. Commenges and Gégout-Petit (2009) define conditional local independence for a class of special semimartingales, and Mogensen, Malinsky and Hansen (2018) and Mogensen and Hansen (2022) show global Markov properties for local independence graphs of certain Itô processes, which are, in particular, special semimartingales. Thus, conditional local independence is well-defined beyond counting processes, and we believe that most definitions and results of this paper would generalize beyond $N$ being a counting process. Besides some additional technical challenges, the major practical obstacle with such a generalization is that we cannot realistically assume to have completely observed sample paths of Itô processes, say. The discrete time nature of the observations should then be included in the analysis, and this is beyond the scope of the present paper.

Irrespectively of the remaining open problems, the simulation study demonstrated some important properties of our proposed test, the X-LCT. First, it was fairly simple to implement for the specific example considered using some standard estimation techniques that were not tailored to the specific model class. Second, it had good level and power properties and clearly outperformed the test based on the misspecified marginal Cox model. Third, both Neyman orthogonalization as well as cross-fitting were pivotal for achieving the good properties of the test.

Acknowledgments. The authors would like to thank the Associate Editor and the anonymous reviewers for constructive comments that lead to substantial improvements of the paper.

Funding. The work was supported by Novo Nordisk Foundation Grant NNF20OC0062897.

SUPPLEMENTARY MATERIAL

Supplementary article (DOI: 10.1214/23-AOS2323SUPPA; .pdf). The supplementary article (Christgau, Petersen and Hansen (2023)) contains proofs of results from the main text, auxiliary results, additional discussions and figures.

Computer code (DOI: 10.1214/23-AOS2323SUPPB; .zip). Computer code that reproduces numerical results for the simulation study in Section 6.

REFERENCES

AALEN, O. O. (1987). Dynamic modelling and causality. Scand. Actuar. J. 3–4 177–190. MR0943579 https://doi.org/10.1016/j.sser.2011.04.029

AALEN, O. O., RØYSLAND, K., GRAN, J. M. and LERDERGERBER, B. (2012). Causality, mediation and time: A dynamic viewpoint. J. Roy. Statist. Soc. Ser. A 175 831–861. MR2993496 https://doi.org/10.1111/j.1467-985X.2011.01030.x
Scheidegger, C., Hörmann, J. and Bühlmann, P. (2022). The weighted generalised covariance measure. *J. Mach. Learn. Res.* **23**, 1–68.

Schilling, R. L. and Partzsch, L. (2014). *Brownian Motion: An Introduction to Stochastic Processes*, 2nd ed. De Gruyter Graduate. de Gruyter, Berlin. MR3234570 https://doi.org/10.1515/9783110307306

Schweder, T. (1970). Composable Markov processes. *J. Appl. Probab.* **7**, 400–410. MR0264755 https://doi.org/10.2307/3211973

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48**, 1514–1538. MR4124333 https://doi.org/10.1214/19-AOS1857

Xu, H., Farajtabar, M. and Zha, H. (2016). Learning granger causality for Hawkes processes. In *Proceedings of the 33rd International Conference on Machine Learning* **48**, 1717–1726.

Zhou, K., Zha, H. and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multidimensional Hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*. 