COUNTERFACTUAL ANALYSES WITH GRAPHICAL MODELS BASED ON LOCAL INDEPENDENCE

KJETIL RØYSLAND

Abstract. We show that one can do causal inference in a natural way for continuous-time scenarios using tools from stochastic analysis. This provides new alternatives to the positivity condition for inverse probability weighting. The probability distribution that would govern the frequency of observations in the counterfactual scenario can be characterized in terms of a so-called martingale problem. The counterfactual and factual probability distributions may be related through a likelihood ratio given by a stochastic differential equation. We can do inference for counterfactual scenarios based on the original observations, re-weighted according to this likelihood ratio. This is possible if the solution of the stochastic differential equation is uniformly integrable, a property that can be determined by comparing the corresponding factual and counterfactual short-term predictions.

Local independence graphs are directed, possibly cyclic, graphs that represent short-term prediction among sufficiently autonomous stochastic processes. We show through an example that these graphs can be used to identify and provide consistent estimators for counterfactual parameters in continuous-time. This is analogous to how Judea Pearl uses graphical information to identify causal effects in finite state Bayesian networks.

Keywords: causal inference, stochastic analysis, event history analysis, marked point processes, change of probability measures, local independence.

1. Introduction

While randomized controlled trials are the gold standard for determining the effects of public health interventions or medical treatments, there are many situations where such trials are unethical and it is tempting to turn to registry data or observational studies for quality assessment of treatments. However, data from such sources is subject to various selection effects from drop-out due to underlying health problems to selection of the treatment itself. These problems have motivated the development of the field of causal inference, including in particular the area of marginal structural models, [RHB00] [Røy11], which have seen applications, for instance, in HIV cohort studies, [SHL$^+$05]. The underlying idea is that observational data can be used to mimic a relevant hypothetical controlled trial or counterfactual scenario.

In this paper, our primary concern is the possibility of estimating parameters in a model for the observations from a counterfactual scenario involving a relevant hypothetical randomized controlled trial. While the specification of an appropriate model for the counterfactual observations is an important topic in itself, we will focus solely on a situation in which such a counterfactual model has been specified correctly. It is common to re-weight the observational data in order to mimic observations coming from the counterfactual scenario. This is usually referred to as inverse probability weighting. Such re-weighting has occasionally been reported to be too unstable, even inconsistent, for various purposes, see [CH08]. It is therefore of great interest to understand when this strategy actually works. We will provide

some rigorous conditions for such re-weighting to be achievable. A similar exposition has not been carried out in the literature before, except partly in [Røy11] and [CH08].

A probability distribution on the underlying sample space that would govern the frequency of observations in the counterfactual scenario can be characterized in terms of a so called martingale problem. Short-term predictions provide dynamical characterizations of the various involved modules. A hypothetical direct intervention on a module would change its dynamics. The non-directly intervened modules on the other hand, should have the same dynamical characterization as in the factual scenario. Martingale problems have been thoroughly studied in stochastic analysis. To us this means that there exist well developed tools for determining feasibility of the previous re-weighting methods. An immediate application of these tools, yields for instance that the probability distribution that would govern the frequencies of events in the counterfactual situation is unique if it exists, see Theorem 4 in the Appendix.

If the re-weighting is feasible, is it then at all possible to estimate the parameters of interest in the counterfactual model from the re-weighted observations? In other words: are these parameters identifiable? Pearl's strategy [Pea09] is to take advantage of graphical structure, in terms of conditional independences, for identification of causal effects. It was shown in [SP06], [HV06] and [DFS11] that this strategy gives a complete theory in the simpler setting of finite state or Gaussian Bayesian networks. For more complicated settings, this problem is far from solved. Some results in this direction for time series were given in [ED10]. We show that it is possible to take advantage of local independence graphs for identification of causal effects in continuous-time settings. Note, as this general problem is very hard, we do not provide a complete theory for identification of causal effects, only an example which slightly extends [MVGH11].

The idea that the counterfactual situation can be assigned probabilities in a way that is consistent with a purely observational scheme, is not new. It has also been considered in the general context of marked point processes in [Røy11], [AP04], [Arj12], [PA99] and [Did00]. We choose a martingale based approach, similarly to [Røy11]. Note also that graphical models based on local independence and doubly stochastic Poisson processes were studied thoroughly in [Did08]. Continuous-time counterfactual interventions were also considered by J. Lok in [Lok08]. She considered structural nested models in continuous-time and applied ideas from structural equation modeling to survival data. Her strategy differs from ours in that we take a purely non-parametric point of view, through change of probability measures.

In Section 2 we describe models for the factual scenario. We then proceed in Section 3 with a description of counterfactual variables and distributions. In Section 4, we give a sufficient condition for such a counterfactual distribution to exist, and also a construction based on martingale methods. In Section 5, we introduce local independence graphs that play the same role as directed acyclic graphs usually do in the literature on causal inference. In Section 6, we consider an example where we can identify consistently estimate controlled direct effects in event history analysis. Finally, in the Appendix, we summarize som propoerties of dual predictable projections and consider unqiueness of counterfactual distributions.

2. The observational regime and autonomous modules

Eventually, we will consider statistical analyses based on observations of several i.i.d. individuals, but first we will consider models for one "generic" individual. We aim to investigate complex systems for each individual formed by finitely many autonomous modules that develop and influence each other throughout time. We will

not provide a detailed recipe for building appropriate models, but simply assume a stochastic model for a generic individual that has some specific properties.

2.1. **The underlying probability space and marked point processes.** We let $\mathcal{V}$ denote the finite set of modules that form the system of interest. The possible outcomes of these modules are supposed to be realized on a probability space $(\Omega, \mathcal{F}, Q)$ with some additional structure that we will now describe. Note that we do not assume that the actual frequencies of outcomes will be governed by the probability measure Q. This measure will only play a role as a "reference measure". The possible "initial" outcomes of each module V are given by the outcomes of a corresponding random variable V_0. The random variables in this family, which we denote by $\mathcal{V}_0$, are mutually independent with respect to Q. The intital outcome of each $V \in \mathcal{V}$ occurs at a, possibly unknown, time point $T(V_0) \leq 0$. The ordering of these time points is assumed to be known. We moreover let

$$p(V_0) := \{V_0' \in \mathcal{V}_0 | T(V_0') < T(V_0)\}, \tag{2.1}$$

and sometimes refer to this set as the *past* of V_0.

The outcomes in the follow-up are driven by a multivariate point process N [Jac75] on a finite time interval $[0, T]$. Let J denote the mark space of N. This space is supposed to be Lusin, i.e. a Borel subset in a compact metric space, and equipped with the Borel σ-algebra $\mathcal{J}$. We assume that for every module V, there exists a $J_V \in \mathcal{J}$ such that

$$V_t(\omega) = V_0(\omega) + \int_{J_V} \int_0^t h(\omega, s, x) N(\omega, ds, dx), \tag{2.2}$$

where h is a bounded process on $[0, T] \times J$ that is predictable with respect to the filtration generated by $N|_{J_V}$ and V_0. We also assume that $\mathcal{V}_0 \perp\!\!\!\perp_Q N$ and that $\coprod_{V \in \mathcal{V}} J_V$ defines a partition of J such that the restricted point processes $\{N|_{J_V}\}_{V \in \mathcal{V}}$ are mutually independent with respect to Q.

For each subset $\mathcal{W} := \{V^1, \ldots, V^d\} \subset \mathcal{V}$, let $\mathcal{F}_t^{\mathcal{W}}$ denote the filtration that is generated by V_0 and $N|_{J_V}$ for every $V \in \mathcal{W}$ and also satisfies the *usual conditions*, see [JS03]. We let $\mathscr{P}^{\mathcal{W}}$ denote the predictable σ-algebra generated by $\mathcal{F}_t^{\mathcal{W}}$ [JS03]. For notational simplicity, we will also write $\mathcal{F}_t^V$ or $\mathscr{P}^V$ instead of $\mathcal{F}_t^{\{V\}}$ or $\mathscr{P}^{\{V\}}$, as well as $\mathcal{F}_t$ or $\mathscr{P}$ instead of $\mathcal{F}_t^{\mathcal{V}}$ or $\mathscr{P}^{\mathcal{V}}$.

2.2. **The factual distribution.** The actual frequencies of outcomes in the model are not assumed to be governed by Q, but another probability measure P such that $P \ll Q$ and

$$V_0 \perp\!\!\!\perp_P T^{-1}T(V_0) \setminus \{V_0\} \mid p(V_0) \tag{2.3}$$

for every $V_0 \in \mathcal{V}_0$, i.e. every V_0 is independent w.r.t. its simultaneous variables, conditionally on the past. We will refer to the property (2.3) as *contemporaneous independence*, see [ED10]. This is useful to us since it provides at least one enumeration $\{V_0^1, \ldots, V_0^n\} = \mathcal{V}_0$ such that $T(V_0^i) \geq T(V_0^j)$ whenever $i > j$ and

$$E_P\big[f(V_0^k)\big|V_0^{k-1}, \ldots, V_0^1\big] = E_P\big[f(V_0^k)\big|p(V_0^k)\big], \tag{2.4}$$

whenever f is a bounded and measurable function and $1 \leq k \leq n$.

The processes in $\mathcal{V}$ are not necessarily mutually independent with respect to P, but are still sufficiently autonomous for our purpose. As an immediate manifestation of this autonomy, note that the modules may not "switch" states simultaneously P-a.s. The reason is that the processes in $\mathcal{V}$ are associated to disjoint subsets in the mark space J, which cannot occur simultaneously. We will refer to P as the *factual measure*. Note however, as some of the processes in $\mathcal{V}$ may be latent, the

factual measure P is also assumed to govern the frequency of events that may be unobserved.

2.3. **The factual likelihood ratio and its factorization.** The autonomy imposes a factorization of the likelihood ratio $\frac{dP}{dQ}$ that will prove to be important to us. First note that a repeated use of the Radon-Nikodym Theorem provides a family $\{Z_0^V\}_{V\in\mathcal{V}}$ of non-negative random variables such that each Z_0^V is $\mathcal{F}_0^{p(V)\cup\{V\}}$-measurable and

$$E_Q\big[Z_0^V\big|\mathcal{F}_0^{p(V)}\big]=1 \quad\text{and}\quad \frac{dP|_{\mathcal{F}_0}}{dQ|_{\mathcal{F}_0}}=\prod_{V\in\mathcal{V}} Z_0^V,\ Q\text{-a.s.} \tag{2.5}$$

There is a similar factorization of $\frac{dP}{dQ}$. Let U denote the dual predictable projection of N with respect to Q onto the filtration $\mathcal{F}_t$ as in [Jac75]. By Lemma 5 in the Appendix there exists a non-negative and $\mathscr{P}\otimes\mathcal{J}$-measurable process λ such that

$$E_P\Big[\int_J\int_0^T h(s,x)N(ds,dx)\Big]=E_P\Big[\int_J\int_0^T h(s,x)\lambda(s,x)U(ds,dx)\Big]$$

for every bounded and $\mathscr{P}\otimes\mathcal{J}$-measurable process h. As common practice, we mostly omit ω from equations in order to be notationally less overwhelming.

We now define the processes

$$H^V(t):=1+\frac{U(\{t\},J_V)-\int_{J_V}\lambda(t,x)U(\{t\},dx)}{1-U(\{t\},J_V)}$$

and

$$K_t^V:=\int_{J_V}\int_0^t \lambda(s,x)-H^V(s)\big(N(ds,dx)-U(ds,dx)\big). \tag{2.6}$$

By (7.3), we see that that $\{K^V\}_{V\in\mathcal{V}}$ defines a family of local Q-martingales with respect to the filtration $\mathcal{F}_t$ such that

$$\big[K^V,K^{V'}\big]=0,\ Q\text{-a.s. for } V\neq V'. \tag{2.7}$$

The solution of the SDE

$$Z_t=Z_0+\sum_{V\in\mathcal{V}}\int_0^t Z_{s-}dK_s^V \tag{2.8}$$

defines a Q-martingale with respect to the filtration $\mathcal{F}_t$ such that

$$Z_t=\frac{dP|_{\mathcal{F}_t}}{dQ|_{\mathcal{F}_t}},\ Q\text{-a.s.}$$

for every $t\in[0,T]$. This follows directly from [Jac75, Theorem 5.1].

We now obtain directly from Yor's additive formula, [Pro05, Theorem II 38] that

$$Z_t=\prod_{V\in\mathcal{V}} Z_t^V, \tag{2.9}$$

where each Z^V solves an SDE

$$Z_t^V:=Z_0^V+\int_0^t Z_{s-}^V dK_s^V. \tag{2.10}$$

3. Actions and counterfactual distributions

We assume that we may directly intervene on a subset of modules $\mathcal{A} \subset \mathcal{V}$ such that their outcomes are changed. This intervention does not directly affect the outcomes of the modules in $\mathcal{X} := \mathcal{V} \setminus \mathcal{A}$. The latter set of modules will only be affected indirectly: The conditional distributions of their short-term behavior, given the past, will remain the same, while the change of previous outcomes yields a change of the background these distributions depend on. We will limit our discussion to actions that are deterministically dependent on the past. These are sometimes referred to as conditional actions. Every conditional action will be represented by a measurable transformation θ of the generic state space $(\Omega, \mathcal{F})$. We think of $\theta(\omega)$ as the direct consequence in the "counterfactual universe" where the action θ was performed.

Whenever P' is a probability measure on $(\Omega, \mathcal{F})$, we let $\theta P'$ denote the push-forward measure over θ, i.e. $\theta P'(F) := P'\big(\theta^{-1}(F)\big)$ for every $F \in \mathcal{F}$. Whenever H is an $\mathcal{F}$-measurable random variable, we let $\theta^* H$ denote the transformed variable, where $\theta^* H(\omega) := H(\theta(\omega))$ for every $\omega \in \Omega$. We assume that θ is "continuous" in the sense that the reference measure Q is quasi-invariant with respect to θ, i.e.

$$\theta Q \ll Q. \tag{3.1}$$

3.1. **Actions and counterfactual distributions at baseline.** Let $V \in \mathcal{V}$ and suppose η is an $\mathcal{F}_0^V$-measurable random variable and h is a bounded and $\mathcal{F}_0^{p(V)}$-measurable random variable. We assume that the outcomes of the not directly intervened part of the system are left invariant by the transformation at baseline, i.e.

$$\theta^* \eta = \eta, \tag{3.2}$$

for every η and every $V \in \mathcal{X}$. We furthermore assume that the action depends deterministically on the past outcomes in the non-intervened system, i.e. whenever $V \in \mathcal{A}$ then

$$\theta^* \eta \text{ is } \mathcal{F}_0^{p(V) \cap \mathcal{X}}\text{-measurable} \tag{3.3}$$

for every η.

A probability distribution P_θ on $(\Omega, \mathcal{F})$ defines a *counterfactual distribution at baseline* if, whenever $V \in \mathcal{A}$, then

$$E_{P_\theta}\big[h\eta\big] = E_{P_\theta}\big[h\theta^*\eta\big], \tag{3.4}$$

and, whenever $V \in \mathcal{X}$, then

$$E_{P_\theta}\big[h\eta\big] = E_{P_\theta}\big[h\theta^* E_P[\eta|\mathcal{F}_0^{p(V)}]\big] \tag{3.5}$$

for every η.

Equation (3.5) means that the short-term behaviour of a directly intervened variable is simply given by the transformed variable. Its outcome is deterministically regulated by the past. Equation (3.5) means that the conditional distribution of an outcome of a not directly intervened variable in the counterfactual scenario, given its past, coincides with the corresponding distribution from the factual scenario.

Note that Pearl's $\mathrm{do}(X = x)$ may also be interpreted as a transformation on sample space that fixes X constantly equal to x and leaves the remaining variables invariant. This means that our characterization of probability measures on $(\Omega, \mathcal{F})$ that would govern the frequencies of events in our system if we, *contrary to the fact*, had applied the hypothetical intervention strategy, is a reformulation of Pearl's *do*-operator on Bayesian networks, [Pea09]. The present approach, however, translates more or less directly to continuous-time settings.

3.2. **Actions and counterfactual distributions in the follow-up period.** Whenever Z is a stochastic process on Ω, we let $\theta^* Z$ denote the process given by the transformed variables $\{\theta^* Z_t\}_{t\in[0,T]}$. We assume that $\theta^* N$ defines a marked point process that is adapted to the history $\{\mathcal{F}_t\}_{t\in[0,T]}$. The action θ is thought to force the outcomes $N|_{[0,T]\times J_{\mathcal{A}}}$ into the outcomes of $\theta^* N|_{[0,T]\times J_{\mathcal{A}}}$, which will only depend on the strictly previous behavior of the not directly intervened system, i.e. whenever $B \in J_{\mathcal{A}}$, then

$$\theta^* N_t(B) \text{ is predictable w.r.t. } \mathcal{F}_t^{\mathcal{X}}. \tag{3.6}$$

The outcomes of the not directly intervened part of the system are left invariant by the transformation during follow-up, i.e.

$$\theta^* N|_{[0,T]\times J_{\mathcal{X}}} = N|_{[0,T]\times J_{\mathcal{X}}}. \tag{3.7}$$

We will say that P_θ defines a *counterfactual distribution* if it defines a counterfactual distribution at baseline and if whenever X is process on the form (2.2) and Λ is an $\mathcal{F}_t$-predictable process of finite variation such that

$$E_P\left[\int_0^T h_s dX_s\right] = E_P\left[\int_0^T h_s d\Lambda_s\right]$$

for every bounded and $\mathcal{F}_t$-predictable process h, then

$$E_{P_\theta}\left[\int_0^T h_s dX_s\right] = E_{P_\theta}\left[\int_0^T h_s d\theta^*\Lambda_s\right] \text{ if } V \in \mathcal{X} \tag{3.8}$$

and

$$E_{P_\theta}\left[\int_0^T h_s dX_s\right] = E_{P_\theta}\left[\int_0^T h_s d\theta^* X_s\right] \text{ if } V \in \mathcal{A}. \tag{3.9}$$

Note that (3.8) means that $\theta^*\Lambda$ defines the compensator of X if $V \in \mathcal{X}$ and (3.9) means that $\theta^* X$ defines the compensator of X otherwise. This offers an analogous interpretation as in the baseline setting. Compensators provide a notion of short-term behaviour, analogously to the previous conditional distributions. The short-term behavior of a not directly intervened process in the counterfactual scenario, based on the past, coincides with the transformed short-term behaviour from the factual scenario. The short-term behavior of a directly intervened process is given entirely by the transformation.

Following [Pea10], we will say that a model consisting of a factual scenario, an action and a corresponding counterfactual distribution, defines a causal model if the counterfactual distribution would fit the actual corresponding counterfactual scenario. That P_θ actually would govern the frequency of observations for this hypothetical scenario is generally not testable and mostly comes down to the question of *no unmeasured confounding*, [Pea10].

4. Construction of counterfactual distributions

4.1. **Construction at baseline.** We will now construct the counterfactual distribtion in a situation with no follow-up period. The construction is then closely related to Pearl's framework, [Pea09]. The next result is important and says heuristically that if the conditional probability, given the past, of observing outcomes that coincide with counterfactually enforced ones are not too small, then there exists a counterfactual distribution. The equation (4.2) then offers a useful description of the distribution. Note that this is a measure theoretical version of the *truncated factorization formula* from [Pea09, (3.10)].

Theorem 1. *If there exists a non-negative $K \in L^1(\mathcal{F}_0, P)$ such that*

$$\frac{d\theta Q|_{\mathcal{F}_0}}{dQ|_{\mathcal{F}_0}} \leq K \prod_{V \in \mathcal{A}} Z_0^V, \ P\text{-a.s.}, \tag{4.1}$$

then

$$\prod_{V \in \mathcal{X}} Z_0^V \cdot \theta Q|_{\mathcal{F}_0} \tag{4.2}$$

defines a counterfactual distribution on $\mathcal{F}_0$ that is absolutely continuous with respect to $P|_{\mathcal{F}_0}$ and imposes contemporaneously independent outcomes.

Proof. First note that for every bounded $\mathcal{F}_0$-measurable random variable η,

$$\begin{aligned} E_P\left[\eta \frac{d\theta Q|_{\mathcal{F}_0}}{dQ|_{\mathcal{F}_0}} \prod_{V \in \mathcal{A}} \frac{1}{Z_0^V}\right] &= E_Q\left[\eta \frac{d\theta Q|_{\mathcal{F}_0}}{dQ|_{\mathcal{F}_0}} \prod_{V \in \mathcal{X}} Z_0^V\right] \\ &= E_{\theta Q}\left[\eta \prod_{V \in \mathcal{X}} Z_0^V\right] \\ &\leq E_P\big[\eta K\big]. \end{aligned}$$

This shows that (4.2) defines a finite measure P_θ on $\mathcal{F}_0$ such that $P_\theta \ll P|_{\mathcal{F}_0}$.

We choose an enumeration $V_1, \ldots, V_m$ of the variables in $\mathcal{X}$ such that $j < k$ implies that $T(V_j) \leq T(V_k)$. If $V_k \in \mathcal{X}$ and η is a bounded $\mathcal{F}_0^{\{V_k\} \cup p(V_k)}$-measurable random variable then

$$E_Q\big[\theta^* \eta \big| \mathcal{F}_0^{\{V_1, \ldots, V_{k-1}\}}\big] = \theta^* E_Q\big[\eta \big| \mathcal{F}_0^{p(V)}\big], \ Q\text{-a.s.} \tag{4.3}$$

To see this, we let η_1 be an $\mathcal{F}_0^{V_k}$-measurable and bounded random variable and let η_2 be an $\mathcal{F}_0^{p(V_k)}$-measurable and bounded variable and compute

$$\begin{aligned} E_Q\big[\theta^*(\eta_1 \eta_2) \big| \mathcal{F}_0^{\{V_1, \ldots, V_{k-1}\}}\big] &= E_Q\big[\eta_1 \big| \mathcal{F}_0^{\{V_1, \ldots, V_{k-1}\}}\big] \theta^* \eta_2 \\ &= \theta^*\big(E_Q\big[\eta_1 \big| \mathcal{F}_0^{\{V_1, \ldots, V_{k-1}\}}\big] \eta_2\big) \\ &= \theta^*\big(E_Q\big[\eta_1 \big| \mathcal{F}_0^{p(V_k)}\big] \eta_2\big) \\ &= \theta^* E_Q\big[\eta_1 \eta_2 \big| \mathcal{F}_0^{p(V_k)}\big], \ Q\text{-a.s.} \end{aligned}$$

The equation (4.3) now follows from the monotone class lemma. Especially, this means that for every $k \leq m$

$$E_Q\big[\theta^* Z_0^{V_k} \big| \mathcal{F}_0^{\{V_1, \ldots, V_{k-1}\}}\big] = \theta^* E_Q\big[Z_0^{V_k} \big| \mathcal{F}_0^{p(V_k)}\big] = 1, \ Q\text{-a.s.} \ , \tag{4.4}$$

and

$$\begin{aligned} E_{\theta Q}\big[Z_0^{V_1} \cdots Z_0^{V_k}\big] &= E_Q\big[\theta^* Z_0^{V_1} \cdots \theta^* Z_0^{V_{k-1}} E_Q\big[\theta^* Z_0^{V_k} \big| \mathcal{F}_0^{\{V_1, \ldots, V_{k-1}\}}\big]\big] \\ &= E_Q\big[\theta^* Z_0^{V_1} \cdots \theta^* Z_0^{V_{k-1}}\big] = E_{\theta Q}\big[Z_0^{V_1} \cdots Z_0^{V_{k-1}}\big]. \end{aligned}$$

That P_θ defines a probability measure on $\mathcal{F}_0$ follows by induction.

To see that (3.5) and (3.4) are satisfied, suppose $V_k \in \mathcal{X}$ and let η, h be bounded random variables such that η is $\mathcal{F}_0^{V_k}$-measurable and h is $\mathcal{F}_0^{p(V_k)}$-measurable. We

see that

$$\begin{aligned}
E_{P_\theta}\left[\eta h\right] &= E_{\theta Q}\left[\left(\prod_{j=1}^{k-1} Z_0^{V_j}\right)\eta h Z_0^{V_j}\right] \\
&= E_Q\left[\left(\prod_{j=1}^{k-1} \theta^* Z_0^{V_j}\right)\theta^* h E_Q\left[\theta^*\eta Z_0^{V_k}\middle|\mathcal{F}_0^{\{V_1,\dots,V_{k-1}\}}\right]\right] \\
&= E_Q\left[\left(\prod_{j=1}^{k-1} \theta^* Z_0^{V_j}\right)\theta^* h\theta^* E_Q\left[\eta Z_0^{V_k}\middle|\mathcal{F}_0^{p(V_k)}\right]\right] \\
&= E_{P_\theta}\left[h\theta^* E_P\left[\eta\middle|\mathcal{F}_0^{p(V)}\right]\right].
\end{aligned}$$

If $V_k \in \mathcal{A}$, then

$$\begin{aligned}
E_{P_\theta}\left[\eta h\right] &= E_{\theta Q}\left[\left(\prod_{j=1}^{k-1} Z_0^{V_j}\right)\eta h Z_0^{V_j}\right] \\
&= E_Q\left[\left(\prod_{j=1}^{k-1} \theta^* Z_0^{V_j}\right)\theta^* h\theta^*\eta E_Q\left[Z_0^{V_k}\middle|\mathcal{F}_0^{\{V_1,\dots,V_{k-1}\}}\right]\right] \\
&= E_Q\left[\left(\prod_{j=1}^{k-1} \theta^* Z_0^{V_j}\right)\theta^* h\theta^*\eta\right] \\
&= E_{P_\theta}\left[h\theta^*\eta\right].
\end{aligned}$$

□

4.2. **Construction for the follow-up period.** The condition (3.1) can be made somewhat more concrete if the processes, that may be directly intervened on, only are allowed to jump at a given finite sequence of predictable times. This behavior is very different from that of Poisson processes. More formally, we assume that there exists a bounded and $\mathcal{F}_t$-predictable multivariate counting measure $\tilde{U}^A$ on $[0,T]\times J_A$ such that

$$N|_{[0,T]\times J_A} \ll \tilde{U}^A \tag{4.5}$$

for every $A \in \mathcal{A}$. We can now show the reference measure Q is quasi-invariant if the probability of an outcome that coincides with the counterfactually enforced outcome at short-term is not too small.

Proposition 1. *Suppose that θ is an $\mathcal{F}$-measurable transformation on Ω that satisfies* (3.2)-(3.6) *and assume* (4.5)*. If there exists a bounded and $\mathscr{P}$-measurable process $\tilde{Y}$ such that*

(1) $\theta Q|_{\mathcal{F}_0} \ll Q|_{\mathcal{F}_0}$,
(2)

$$\int_{J_A}\int_0^T h(s,x)\theta^* N(ds,dx) = \int_{J_A}\int_0^T h(s,x)\tilde{Y}(s,x)U^A(ds,dx) \tag{4.6}$$

Q-a.s. for every $A \in \mathcal{A}$ and bounded and $\mathscr{P}$-measurable process h,

(3) *There exists a constant $c > 0$ such that*

$$1 - \theta^* N(\{s\}, J_A) \le c\cdot\left(1 - U^A(\{s\}, J_A)\right),\ Q\text{-a.s.} \tag{4.7}$$

for every $s \in [0,T]$,

then $\theta Q \ll Q$.

Proof. The integral equation

$$\begin{aligned}&\int_J \int_0^T h(s,x)U^\theta(ds,dx)\\ &= \sum_{A\in\mathcal{A}} \int_{J_A}\int_0^T h(s,x)\theta^* N(ds,dx) + \sum_{V\in\mathcal{X}} \int_{J_A}\int_0^T h(s,x)U^V(ds,dx)\end{aligned}$$

defines an $\mathcal{F}_t$- predictable random measure U^θ on $[0,T]\times J$.

Let $B\subset J$ be a measurable subset and define $N_t^B := \int_0^t \int_B N(ds,dx)$. If $B\subset J_A$ for an $A\in\mathcal{A}$ and S is a $\mathcal{F}_t$-adapted stopping time, then

$$E_{\theta Q}\big[N_S^B - U_S^\theta(B,[0,t])\big] = E_{\theta Q}\big[N_S^B - \theta^* N_S^B\big] = E_{\theta Q}\big[\theta^* N_S^B - \theta^* N_S^B\big] = 0.$$

This means that $N_t - U_t^\theta(B,[0,t])$ defines a local Q-martingale with respect to the filtration $\mathcal{F}_t$. Similarly, if $B\subset J_{\mathcal{X}}$, note that

$$\begin{aligned}E_{\theta Q}\Big[\int_0^T h_s dN_s^B\Big] =& E_Q\Big[\int_0^T \theta^* h_s dN_s^B\Big]\\ =& E_Q\Big[\int_0^T \theta^* h_s dU(ds,B)\Big]\\ =& E_{\theta Q}\Big[\int_0^T h_s dU^\theta(ds,B)\Big]\end{aligned}$$

for every bounded and $\mathcal{F}$-predictable process h. Now, $N([0,t],B) - U^\theta([0,t],B)$ defines a local θQ-martingale with respect to the filtration $\{\mathcal{F}_t\}_{t\in[0,T]}$. This means that

$$E_{\theta Q}\Big[\int_J\int_0^T h(s,x)N(ds,dx)\Big] = E_{\theta Q}\Big[\int_J\int_0^T h(s,x)U^\theta(ds,dx)\Big]$$

for every bounded and $\mathscr{P}\otimes\mathcal{J}$-measurable process h.

We define the processes

$$\begin{aligned}H^A(t,x) &:= \tilde{Y}(t,x) - 1 - \frac{U(\{t\},J_A) - \theta^* N(\{t\},J_A)}{1-U(\{t\},J_A)}\\ &\quad I(U(\{t\},J_A)\neq 1)\\ \zeta_t^A &:= \int_{J_A}\int_0^t H^A(s,x)\big(N(ds,dx) - U(ds,dx)\big)\end{aligned}$$

and let $\zeta := \sum_{A\in\mathcal{A}} \zeta^A$.

By [JS03, Proposition I 3.13] there exists a $\mathscr{P}$-measurable and non-negative stochastic process γ^A such that $\gamma^A \le 1$ and

$$\int_{J_A}\int_0^T h(s,x)U^A(ds,dx) = \int_{J_A}\int_0^T h(s,x)\gamma^A(s,x)\tilde{U}^A(ds,dx)$$

Q-a.s. for every bounded and $\mathscr{P}$-measurable stochastic process h.

A computation shows that the predictable variation process for ζ with respect to Q satisfies

$$\begin{aligned}\langle\zeta,\zeta\rangle_t &= \sum_{A\in\mathcal{A}} \langle\zeta^A,\zeta^A\rangle_t\\ &= \sum_{A\in\mathcal{A}} \int_{J_A}\int_0^t H^A(s,x)^2\gamma^A(s,x)(1-\gamma^A(s,x))\tilde{U}^A(ds,dx),\end{aligned}$$

which is Q-a.s. uniformly bounded. Now, [LM78, Theorem II.1] implies that the SDE:

$$\rho_t = \frac{d\theta Q|_{\mathcal{F}_0}}{dQ|_{\mathcal{F}_0}} + \int_0^t \rho_{s-} d\zeta_s \tag{4.8}$$

defines a uniformly integrable Q-martingale with respect to the filtration $\mathcal{F}_t$. This means that

$$\tilde{Q} := \rho_T \cdot Q$$

defines a probability measure on $(\Omega, \mathcal{F})$.

A computation shows that if $B \subset J_V$ for some $V \in \mathcal{V}$, then

$$N_t^B - U_t([0,t], B) - \int_0^t \rho_{s-}^{-1} d\langle N^B - U^B, \rho\rangle_s = N_t^B - U^\theta([0,t], B). \tag{4.9}$$

Girsanov's Theorem [JS03, Theorem III 1.21] implies that

$$E_{\tilde{Q}}\left[\int_J \int_0^T h(s,x) N(ds,dx)\right] = E_{\tilde{Q}}\left[\int_J \int_0^T h(s,x) U^\theta(ds,dx)\right]$$

for every bounded and $\mathscr{P} \otimes \mathcal{J}$-measurable process h. Finally, [Jac75, Theorem 3.4] implies that there exists only one probability measure which has U^θ as a dual predictable projection for N. Therefore $\theta Q = \tilde{Q} \ll Q$. □

The next result is important and says that if the probability of observing an outcome that coincides with the counterfactually enforced outcome at short-term is not too small, then there exists a counterfactual distribution for the follow-up period. The counterfactual distribution can then be obtained by re-weighting the factual distribution, i.e. $P_\theta \ll P$. Note that (4.12) provides a continuous-time analogy of the truncated factorization formula for Bayesian networks, [Pea09, (3.10)].

Theorem 2. *Suppose that the conditions of Theorem 1 are satisfied and that there exists a bounded and $\mathscr{P}$-measurable process Y such that*

(1)

$$\int_{J_A} \int_0^T h(s,x)\theta^* N(ds,dx) = \int_{J_A} \int_0^T h(s,x) Y(s,x)\lambda(s,x) U(ds,dx) \tag{4.10}$$

P-a.s. for every $A \in \mathcal{A}$ and bounded and $\mathscr{P}$-measurable process h,

(2) *there exists a constant $c > 0$ such that*

$$1 - \theta^* N(\{s\}, J_A) \le c\big(1 - \lambda \cdot U(\{s\}, J_A)\big), \ P\text{-a.s.} \tag{4.11}$$

for every $s \in [0,T]$.

Then there exists a counterfactual distribution P_θ such that $P_\theta \ll P$. We also have that $P_\theta \ll \theta Q$ and

$$X_t := \prod_{V \in \mathcal{X}} Z_t^V, \tag{4.12}$$

where Z^V is the process defined in (2.10), defines a θQ-martingale with respect to the filtration $\{\mathcal{F}_t\}$ that satisfies the SDE

$$X_t = \prod_{V \in \mathcal{X}} Z_0^V + \sum_{V \in \mathcal{X}} \int_0^t X_{s-} dK_s^V \tag{4.13}$$

and

$$\frac{dP_\theta}{d\theta Q} = X_T. \tag{4.14}$$

Proof. We follow the proof of Proposition 1 and define the processes

$$G^A(t,x) := Y(t,x) - 1 - \frac{\lambda \cdot U(\{t\}, J_A) - \theta^* N(\{t\}, J_A)}{1 - \lambda \cdot U(\{t\}, J_A)} I(\lambda \cdot U(\{t\}, J_A) \neq 1)$$
$$\xi_t^A := \int_{J_A} \int_0^t G^A(s,x)\big(N(ds,dx) - \lambda \cdot U(ds,dx)\big)$$
$$\xi := \sum_{A \in \mathcal{A}} \xi^A$$

By [JS03, Proposition I 3.13] there exists a $\mathscr{P}$-measurable and non-negative stochastic process γ^A such that $\gamma^A \leq 1$ and

$$\int_{J_A} \int_0^T h(s,x)\lambda(s,x)U(ds,dx) = \int_{J_A} \int_0^T h(s,x)\gamma^A(s,x)\tilde{U}^A(ds,dx)$$

Q-a.s. for every bounded and $\mathscr{P}$-measurable stochastic process h.

A computation shows that the predictable variation process for ξ with respect to P satisfies

$$\langle \xi, \xi \rangle_t = \sum_{A \in \mathcal{A}} \langle \xi^A, \xi^A \rangle_t$$
$$= \sum_{A \in \mathcal{A}} \int_{J_A} \int_0^t G^A(s,x)^2 \gamma^A(s,x)(1 - \gamma^A(s,x))\tilde{U}^A(ds,dx),$$

which is Q-a.s uniformly bounded. Now, [LM78, Theorem II.1] implies that the SDE

$$W_t = \frac{dP_\theta|_{\mathcal{F}_0}}{dP|_{\mathcal{F}_0}} + \int_0^t W_{s-} d\xi_s \tag{4.15}$$

defines a uniformly integrable P-martingale with respect to the filtration $\mathcal{F}_t$. This means that

$$P_\theta := Z_T \cdot P$$

defines a probability measure on $(\Omega, \mathcal{F})$.

The integral equation

$$\int_J \int_0^T h(s,x)\nu^\theta(ds,dx) \tag{4.16}$$
$$= \int_{J_\mathcal{X}} \int_0^T h(s,x)\lambda(s,x)U(ds,dx) + \int_{J_\mathcal{A}} \int_0^T h(s,x)\theta^* N(ds,dx) \tag{4.17}$$

defines a predictable and non-negative random measure ν^θ on $[0,T] \times J$ such that

$$\xi_t = \int_J \int_0^t \left(Y(s,x) - 1 - \frac{U\lambda \cdot (\{s\}, J) - \nu^\theta(\{s\}, J)}{1 - \lambda \cdot U(\{s\}, J)} I\big(\lambda \cdot U(\{s\}, J) \neq 1\big) \right) N(ds,dx) - \lambda \cdot U(ds,dx).$$

We obtain from [Jac75, Theorem 5.2] that

$$E_{P_\theta}\left[\int_J \int_0^T h(s,x)N(ds,dx) \right] = E_{P_\theta}\left[\int_J \int_0^T h(s,x)\nu^\theta(ds,dx) \right]$$

for every bounded and $\mathscr{P} \otimes \mathcal{J}$-measurable process h, i.e. P_θ defines a counterfactual distribution.

We may compute that

$$\begin{aligned}\Delta\zeta_s^A = &\int_{J_A} \tilde{Y}(s,x)N(\{s\},dx) - \theta^* N(\{s\},J_A) \\ &+ \Big(\tilde{U}(\{s\},J_A) - \theta^* N(\{s\},J_A) I\big(U(\{s\},J_A) \neq 1\big)\Big) \\ &\times \Big(\tilde{U}(\{s\},J_A) - N(\{s\},J_A)\Big)\end{aligned}$$

and that

$$\begin{aligned}\Delta\xi_s^A = &\int_{J_A} Y(s,x)N(\{s\},dx) - \theta^* N(\{s\},J_A) \\ &+ \Big(\tilde{U}(\{s\},J_A) - \theta^* N(\{s\},J_A) I\big(\lambda \cdot U(\{s\},J_A) \neq 1\big)\Big) \\ &\times \Big(\tilde{U}(\{s\},J_A) - N(\{s\},J_A)\Big).\end{aligned}$$

We moreover define a process χ as follows

$$\chi_t := \sum_{s \leq t} \frac{\Delta\xi_s - \Delta\zeta_s}{\Delta\zeta_s + 1} I\big(\Delta\zeta_s \neq -1\big).$$

One can show that χ only jumps at the jump times of $\tilde{U}$ and that $\Delta\chi$ is uniformly bounded. This means that the SDE

$$\pi_t := \frac{d\theta Q|_{\mathcal{F}_0}}{dQ|_{\mathcal{F}_0}} \prod_{A \in \mathcal{A}} \frac{1}{Z_0^A} + \int_0^t \pi_{s-} d\chi_s. \tag{4.18}$$

defines a P semi-martingale with respect to the filtration $\mathcal{F}_t$. Note that $\Delta\zeta_s = -1$ implies that $\Delta\xi_s = -1$, so

$$\zeta + [\zeta, \chi] + \chi = \xi, \ P\text{-a.s.} \tag{4.19}$$

Yor's additive formula [Pro05, Theorem II 38] then implies that

$$\pi_t \rho_t = \frac{dP_\theta|_{\mathcal{F}_0}}{dP|_{\mathcal{F}_0}} + \int_0^t \pi_{s-} \rho_{s-} d\xi_s. \tag{4.20}$$

This implies that $W = \rho\pi$, hence

$$E_{P_\theta}\big[h\big] = E_P\big[hW_T\big] = E_Q\big[hZ_T\rho_T\pi_T\big] = E_{\theta Q}\big[hZ_T\pi_T\big]$$

for every bounded and $\mathcal{F}_T$- measurable random variable h, so $P_\theta \ll \theta Q$. Finally [Jac75, Theorem 5.1], shows that the likelihood ratio $\frac{dP_\theta}{d\theta Q}$ is given by the SDE (4.13), hence Yor's additive formula provides the identity (4.12).

□

Note that since $P_\theta \ll \theta Q = \theta^2 Q$, the counterfactual distribution P_θ is actually invariant with respect to the action θ, i.e.

$$\theta P_\theta = P_\theta. \tag{4.21}$$

5. Local independence

5.1. **Identifiability and short-term dependence.** A causal effect is identifiable if it can be uniquely obtained from the factual distribution of the observable variables. This is generally very hard to determine and may also require further parametric assumptions. We show that it is possible to take advantage of graphical structure, in terms of local independence graphs to do this. Such graphs are useful when deciding which situations causal effects are identifiable, and also which factors we might adjust for.

We will say that $V \in \mathcal{V}$ is *locally independent* of a subset $\mathcal{B} \subset \mathcal{V}$ at baseline, conditionally on $\mathcal{V}' \subset \mathcal{V}$, if the conditional density of V_0, given the past, does not depend on the baseline information from $\mathcal{B}$. More precisely, for every integrable and $\mathcal{F}_0^V$-measurable random variable η, there exists a random variable $\tilde{\eta}$ that is $\mathcal{F}_0^{p(V)\cap(\mathcal{V}'\setminus\mathcal{B})}$-measurable and such that if h is $\mathcal{F}_0^{p(V)\cap\mathcal{V}'}$-measurable, then

$$E_P\big[\eta h\big] = E_P\big[\tilde{\eta} h\big]. \tag{5.1}$$

A process $V \in \mathcal{V}$ is locally independent of $\mathcal{B} \subset \mathcal{V}$ during follow-up, conditionally on $\mathcal{V}'$, if for every process X on the form (2.2) there exists an $\mathcal{F}_t^{\{V\}\cup\mathcal{V}'\setminus\mathcal{B}}$-predictable process Λ with finite variation such that

$$E_P\bigg[\int_0^T h_s dX_s\bigg] = E_P\bigg[\int_0^T h_s d\Lambda_s\bigg] \tag{5.2}$$

for every bounded and $\mathcal{F}_t^{\{V\}\cup\mathcal{V}'}$-predictable process h. If V is locally independent of $\mathcal{B}$, conditionally on $\mathcal{V}'$, both at baseline and during follow-up, we will say that V is locally independent of $\mathcal{B}$, conditionally on $\mathcal{V}'$. This will sometimes be written $\mathcal{B} \nrightarrow V|\mathcal{V}'$. A *local independence graph* is a directed graph $G = (\mathcal{V}', \mathcal{E})$ for $\mathcal{V}' \subset \mathcal{V}$ such that the absence of an arrow from a subset $\mathcal{B} \subset \mathcal{V}'$ to a process $V \in \mathcal{V}'$ means that $\mathcal{B} \nrightarrow V|\mathcal{V}'$. Note that local independence graphs are also refered to as local independence graphs, see [Did08] [ARGL12], and were introduced in [Sch70].

Given time points $\{T(V_0)\}_{V\in\mathcal{V}}$ at baseline and a local independence graph $G = (\mathcal{V}, \mathcal{E})$, we can pick a linear ordering of $\mathcal{V}_0$ that satisfies (2.4) and therefore yields

$$V_0^i \perp\!\!\!\perp_P \{V_0^1, V_0^2, \ldots, V_0^{i-1}\}\big|\mathcal{F}_0^{\mathrm{pa}(V^i)}, \tag{5.3}$$

for every $i \leq n$. The property (5.3) is known as the ordered directed Markov Property and was shown to be equivalent to the local directed Markov property in [Lau01, Theorem 2.11]. This means that Bayesian networks and local independence graphs are two descriptions of the same structure when the nodes correspond to single variables. Note that local independence graphs, where the nodes are allowed to be families of variables or processes, are allowed to be cyclic.

5.2. **Measurability of intensities.** Local independence during the follow-up is closely related to the measurability of intensities.

Lemma 1. *Suppose that V is locally independent of $\mathcal{B}$ at baseline, conditionally on $\mathcal{V}'$, then $\mathcal{B} \nrightarrow V|\mathcal{V}'$ if and only if there exists a non-negative and $\mathscr{P}^{\{V\}\cup\mathcal{V}'\setminus\mathcal{B}}$-measurable process λ^V such that*

$$E_P\bigg[\int_{J_V}\int_0^T h(s,x)N(ds,dx)\bigg] = E_P\bigg[\int_{J_V}\int_0^T h(s,x)\lambda^V(s,x)U(ds,dx)\bigg] \tag{5.4}$$

for every bounded and $\mathscr{P}^{\{V\}\cup\mathcal{V}'}$-measurable process h.

Proof. If there exists a process λ^V as in (5.4), then $\mathcal{B} \nrightarrow V|\mathcal{V}'$ follows directly. Conversely, suppose that $\mathcal{B} \nrightarrow V|\mathcal{V}'$ and let $D \subset J_V$ be a measurable subset. Now, $N_t^D := N([0,t], D)$ defines a processes on the form (2.2), so there must exist a corresponding predictable increasing process Λ^D of finite variation such that

$$E_P\bigg[\int_0^T h_s dN_s^D\bigg] = E_P\bigg[\int_0^T h_s d\Lambda_s^D\bigg]$$

for every bounded and $\mathcal{F}_t^{\{V\}\cup\mathcal{V}'}$-predictable process h.

The Radon-Nikodym theorem now provides an $\mathcal{F}^{\{V\}\cup\mathcal{V}'\setminus\mathcal{B}}$-measurable and non-negative process $\lambda^{(D)}$ such that

$$E_P\Big[\int_0^T h_s dN_s^D\Big] = E_P\Big[\int_0^T h_s \lambda_s^{(D)} U(ds, D)\Big] \tag{5.5}$$

for every bounded and $\mathcal{F}_t^{\{V\}\cup\mathcal{V}'}$-measurable process h.

Since J is a Lusin space, we may construct a non-negative and $\mathscr{P}^{\{V\}\cup\mathcal{V}'}$-measurable process λ^V that satisfies (5.4) as a limit of processes that are finite linear combinations of processes on the form $f \cdot J_D$ where D is a measurable subset in J_V and f is a bounded $\mathcal{F}_t^{\{V\}\cup\mathcal{V}'\setminus\mathcal{B}}$-measurable process.

□

5.3. **Markovian factorization property.** The local Markov property implies the *Markovian factorization property*, see [Pea09, (1.33)] and [Lau01, (2.10)]. We will now see that a local independence graph yields a similar factorization for the follow-up period. We use the following notation from graph theory: whenever $V \in \mathcal{V}$, let $cl(V) \subset \mathcal{V}$ denote the set formed by V and its parents in G.

Theorem 3. *If $G = (\mathcal{V}, \mathcal{E})$ is a local independence graph with respect to P then there exists an $\mathcal{F}_t^{cl(V)}$-adapted P-indistiguishable version of each process Z^V from Theorem 2.9 where*

$$Z = \prod_{V\in\mathcal{V}} Z^V, P\text{ -a.s.}$$

Proof. Let $\mathcal{F}_0^{\mathrm{pa}(V)} := \bigvee_{V'\in\mathrm{pa}(V)} \mathcal{F}_0^{V'}$ and $\mathcal{F}_0^{\mathrm{cl}(V)} := \bigvee_{V'\in\mathrm{cl}(V)} \mathcal{F}_0^{V'}$ and let

$$Y^V := \frac{dP|_{\mathcal{F}_0^{\mathrm{pa}(V)}}}{dQ|_{\mathcal{F}_0^{\mathrm{pa}(V)}}}.$$

Now

$$P|_{\mathcal{F}_0^{\mathrm{cl}(V)}} \ll Y^V \cdot Q|_{\mathcal{F}_0^{\mathrm{cl}(V)}}, \tag{5.6}$$

so there exists, by the Radon-Nikodym theorem, an $\mathcal{F}_0^{\mathrm{cl}(V)}$-measurable random variable $\tilde{Z}_0^V$ such that

$$P|_{\mathcal{F}_0^{\mathrm{cl}(V)}} = \tilde{Z}_0^V Y^V \cdot Q|_{\mathcal{F}_0^{\mathrm{cl}(V)}}. \tag{5.7}$$

We then have, for every bounded and measurable function h, that

$$\begin{aligned} E_P\big[h(V_0)\big|\mathcal{F}_0^{p(V)}\big] &= E_P\big[h(V_0)\big|\mathcal{F}_0^{pa(V)}\big] = E_Q\big[h(V_0)\tilde{Z}_0^V\big|\mathcal{F}_0^{pa(V)}\big] \\ &= E_Q\big[h(V_0)\tilde{Z}_0^V\big|\mathcal{F}_0^{p(V)}\big]. \end{aligned}$$

The contemporaneous independence at baseline and a simple monotone class argument shows that

$$E_P[\eta] = E_Z[\eta \prod_{V\in\mathcal{V}} \tilde{Z}_0^V] \tag{5.8}$$

for every bounded and $\mathcal{F}_0$-measurable random variable η.

For the follow-up, note that by Lemma 1 there exists a non-negative and $\mathscr{P}^{\mathrm{cl}(V)}$-measurable process λ^V such that

$$E_P\Big[\int_{J_V}\int_0^T h(s,x)N(ds,dx)\Big] = E_P\Big[\int_{J_V}\int_0^T h(s,x)\lambda^V(s,x)U(ds,dx)\Big]$$

for every bounded and $\mathscr{P}$-measurable process h.

We may now form K^V, Z^V and Z as in Theorem 2.9 using λ^V instead of λ. Following the short argument in [Bré81, Theorem II T12], we see that any other

choice of a non-negative and $\mathscr{P}$-measurable process λ that satisfies the previous equation would necessarily give

$$\int_{J_V}\int_0^T I(\lambda(s,x)\neq\lambda^V(s,x))N(ds,dx)=0, P\text{-a.s.} \tag{5.9}$$

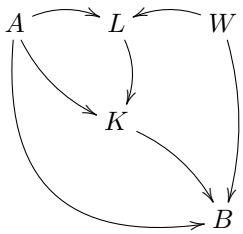
This means that the corresponding versions of the process K^V from (2.6) would be P indistinguishable. Furthermore, this also means that the version corresponding to λ^V provides an $\mathcal{F}_t^{\mathrm{cl}(V)}$-adapted solution of the SDE (2.10) which is P-indistinguishable version from Z^V.

□

6. An example: Controlled direct effects

We now illustrate how local independence graphs can be used to identify causal effects by an example with cancer patients. Suppose each patient is offered one of two different surgical treatments a_1 or a_2. The patient is subject to an examination after surgery where some measurements are taken. These measurements might depend on the chosen surgical procedure and some underlying health condition that is not directly observed. After the surgery, the patient is given further treatment in order to prevent relapse. The chosen post surgery treatment strategy might depend on the surgical procedure and the measurements.

We consider a generic model for the patients in this scenario. The relevant outcomes are provided by the family of random variables $\mathcal{V}=\{W,A,L,K,B\}$. As in section 2, we consider a probability measure Q such that these variables are independent and a probability measure P that governs the frequency of outcomes in the factual scenario and such that $P\ll Q$. Let the random variable A denote the choice of surgery, let W denote the latent health condition, let L take the value of the measurements after surgery, let K denote the post surgery treatment strategy and let B denote the status of relapse. We furthermore assume that $T(W)<T(A)<T(L)<T(K)<T(B)$ and that the following local independencies are satisfied:



How much of the treatment effect is due to the choice of surgical procedure alone, i.e. not due to the choice of post surgery treatment? Judea Pearl, [Pea09, 4.5.3] showed that it is possible to identify the controlled direct effect from surgery on the risk of relapse, even without any observations of W. We rephrase his argument slightly:

Proposition 2. *If θ^*K is $\mathcal{F}_0^L$-measurable, θ^*A is constant, L, W and B are θ-invariant, there exists a constant $c>0$ such that*

$$P\big(A=\theta^*A\big)>0 \quad \textit{and} \quad P\big(K=\theta^*K\big|A=\theta^*A,L\big)\geq c,\ P\textit{-a.s} \tag{6.1}$$

and h is a bounded and measurable function, then there exists a unique counterfactual distribution P_θ such that $P_\theta\ll P$ and

$$E_{P_\theta}\big[h(B)\big]=\theta^*E_P\big[\theta^*E_P\big[h(B)|\mathcal{F}_0^{\{L,A,K\}}\big]\big|\mathcal{F}_0^A\big],\ P_\theta\textit{-a.s.}\ . \tag{6.2}$$

Let $\tilde{\mathcal{F}}_0 := \mathcal{F}_0^{\{L,A,K,B\}}$ *and suppose that* $\tilde{Z}^B$ *is a non-negative and* $\tilde{\mathcal{F}}_0$*-measurable random variable and* $\tilde{Z}^L$ *is a non-negative* $\mathcal{F}_0^A$*-measurable random variable such that*

$$E_P\big[h(B)\big|\mathcal{F}_0^{\{A,L,K\}}\big] = E_Q\big[h(B)\tilde{Z}^B\big|\mathcal{F}_0^{\{A,L,K\}}\big]$$
$$E_P\big[h(L)\big|\mathcal{F}_0^{A}\big] = E_Q\big[h(L)\tilde{Z}^L\big|\mathcal{F}_0^{A}\big]$$

P-a.s. Now,

$$E_{P_\theta}\big[H\big] = E_{\theta Q}\big[H\tilde{Z}^L\tilde{Z}^B\big] \tag{6.3}$$

for every $\tilde{\mathcal{F}}_0$*-measurable random variable* H*, i.e.*

$$\frac{dP_\theta|_{\tilde{\mathcal{F}}_0}}{d\theta Q|_{\tilde{\mathcal{F}}_0}} = \tilde{Z}^L\tilde{Z}^B. \tag{6.4}$$

Proof. Note that (6.1) means that (4.1) is satisfied, i.e. we obtain a counterfactual distribution P_θ from Theorem 1.

Whenever h_1, h_2 are bounded and measurable functions, then

$$\begin{aligned}
E_{P_\theta}\big[h_1(B)h_2(L)\big] &= E_P\big[W_0 h_1(B)h_2(L)\big] \\
&= E_P\Big[W_0 E_P\big[h(B)\big|\mathcal{F}_0^{A,K,L}\big]h(L)\Big] \\
&= E_{P_\theta}\Big[E_P\big[h(B)\big|\mathcal{F}_0^{A,K,L}\big]h(L)\Big] \\
&= E_{P_\theta}\Big[\theta^* E_P\big[h(B)\big|\mathcal{F}_0^{A,K,L}\big]h(L)\Big] \text{ by (4.21).}
\end{aligned}$$

This shows that $E_{P_\theta}\big[h_1(B)\big|\mathcal{F}_0^{L}\big] = \theta^* E_P\big[h_1(B)\big|\mathcal{F}_0^{\{A,L,K\}}\big]$ P_θ-a.s. Moreover, note that

$$\begin{aligned}
E_{P_\theta}\big[h_2(L)\big] &= E_{P_\theta}\Big[\theta^* E_P\big[h_2(L)\big|\mathcal{F}_0^{A,W}\big]\Big] \\
&= E_{P_\theta}\Big[\theta^* E_P\big[h_2(L)\big|\mathcal{F}_0^{A}\big]\Big] = \theta^* E_P\big[h_2(L)\big|\mathcal{F}_0^{A}\big],\ P_\theta\text{-a.s.}
\end{aligned}$$

Combining these computations, we obtain

$$\begin{aligned}
E_{P_\theta}\big[h(B)\big] &= E_{P_\theta}\big[E_{P_\theta}\big[h(B)\big|\mathcal{F}_0^{L}\big]\big] = E_{P_\theta}\big[\theta^* E_P\big[h(B)|\mathcal{F}_0^{L,A,K}\big]\big] \\
&= \theta^* E_P\big[\theta^* E_P\big[h(B)|\mathcal{F}_0^{L,A,K}\big]\big|A\big]
\end{aligned}$$

P_θ-a.s. for every bounded and measurable function h.

To see that the equation (6.3) is satisfied, note that by the monotone class lemma:

$$\begin{aligned}
E_{P_\theta}\big[H\big] &= \theta^* E_P\big[\theta^* E_P\big[H|\mathcal{F}_0^{\{A,L,K\}}\big]\big|\mathcal{F}_0^{A}\big] \\
&= \theta^* E_Q\big[\theta^* E_Q\big[H\tilde{Z}^B|\mathcal{F}_0^{\{A,L,K\}}\big]\tilde{Z}^L\big|\mathcal{F}_0^{A}\big] \\
&= E_{\theta Q}\big[E_Q\big[H\tilde{Z}^B|\mathcal{F}_0^{\{A,L,K\}}\big]\tilde{Z}^L\big] \\
&= E_{\theta Q}\big[H\tilde{Z}^B\tilde{Z}^L\big].
\end{aligned}$$

□

If we consider actions θ_1 and θ_2 such that $\theta_i^* A = a_i$ and $\theta_1^* K = \theta_2^* K$, Q-a.s. then the relative direct risk of relapse is given by

$$\frac{P_{\theta_1}\big(B=1\big)}{P_{\theta_2}\big(B=1\big)} = \frac{E_P\big[\theta_1^* E_P\big[h(B)|\mathcal{F}_0^{\{A,L,K\}}\big]\big|A=a_1\big]}{E_P\big[\theta_2^* E_P\big[h(B)|\mathcal{F}_0^{\{A,L,K\}}\big]\big|A=a_2\big]}. \tag{6.5}$$

6.1. **Incomplete observations and time dependent treatments.** We have not yet taken into account that the patient observations could be censored during the follow-up period. There might be several reasons for such censoring. This might be due to end of study period, drop-out due to the underlying health or because of other reasons. The risk of having an observed relapse will typically be smaller than the risk of having a relapse. We will work in the framework of event history analysis in order to provide a reasonable effect measure subject to such incomplete observations. This will also allow us to consider time dependent post surgery strategies K.

6.1.1. *A dynamic model.* We proceed with the previous setup, but where B and K are represented by processes and every patient may be censored during the follow-up period. The factors A, L and W are as in the previous example. B is represented by a counting process that jumps from 0 to 1 at the time of the event. The censoring of the individual is represented by a counting process C that jumps from 0 to 1 at the time of censoring.

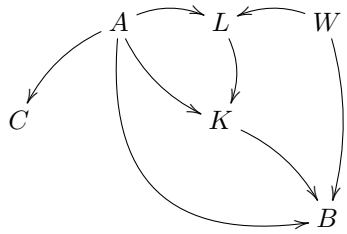
We suppose that the baseline treatment A may be of two different types, hence A takes value in $\{0, 1\}$. Moreover, we suppose that additional post-surgery treatment is given to the patient at the jumps of the counting process K. This treatment may be given recursively, but only at a series of $\mathcal{F}_t$-predictable times, i.e. (4.5) must be satisfied. We furthermore suppose that $\theta^* K_s$ is constant for every s P-a.s. and suppose that $B_0 = 0$, $K_0 = 0$ and $C_0 = 0$ P-a.s.

Let $T_1, \ldots, T_n$ denote the potential post-treatment times and let $U_t^K := \sum_i I(T_i \leq t)$. The counting process U^K is predictable and $\nu_t^K = \int_0^t P\big(\Delta K_s \neq 0 \big| \mathcal{F}_{s-}\big) dU_s^K$. By Theorem 2, we see that there exists a counterfactual distribution if $P\big(A = \theta^* A\big) > 0$ and there exist $c_1, c_2 > 0$ such that

$$1 - c_1 P\big(\Delta K_s = 0 \big| \mathcal{F}_{s-}\big) \leq \Delta\theta^* K_s \leq c_2 P\big(\Delta K_s \neq 0 \big| \mathcal{F}_{s-}\big) \tag{6.6}$$

for every s P-a.s.

We suppose that the following local independence graph is satisfied with respect to the factual distribution P



Especially, this means that the short-term behavior of the censoring may not depend on other variables than A.

6.1.2. *Restriction to Aalen's additive hazard model.* If we assume that the event process satisfies Aalens's additive hazard model[ABGK93], it is actually possible to identify, and also consistently estimate the direct effect from surgery. Every outcome after the time of censoring is supposed to be unobserved. In addition, we assume that we are not able to observe the variable W.

We consider the censored process

$$\tilde{B}_t := B_0 + \int_0^t (1 - C_{s-}) dB_s$$

and let $\tilde{\mathcal{F}}_t$ denote the filtration that is generated by A, K, L, C and B. Furthermore let Y_t denote the factual "at-risk" process, i.e. $Y_t = I(B_{t-} = C_{t-} = 0)$. We assume

that there exist measurable and bounded functions ψ^0, ψ^K, ψ^L and ψ^A such that

$$E_P\Big[\int_0^T h_s d\tilde{B}_s\Big] = E_P\Big[\int_0^T h_s Y_s\big(\psi_s^0 + A\psi_s^A + L\psi_s^L + \tilde{K}_{s-}\psi_s^K\big)ds\Big] \tag{6.7}$$

for every bounded and $\tilde{\mathcal{F}}_t$-predictable process h.

We are now able to identify the controlled direct effect from surgery. Note that this is just a slight variation of the model considered in [MVGH11].

Lemma 2. *If σ^1 and σ^2 are two $\mathcal{F}_0^A \vee \mathcal{F}_t^{\tilde{B}}$-predictable processes such that*

$$E_P\Big[L\int_0^T h_t Y_t \exp\big(\int_0^t K_{-s}\psi_s^K ds\big)dt\Big] = E_P\Big[\int_0^T h_t\sigma_t^1 dt\Big]$$
$$E_P\Big[\int_0^T h_t Y_t \exp\big(\int_0^t K_{-s}\psi_s^K ds\big)dt\Big] = E_P\Big[\int_0^T h_t\sigma_t^2 dt\Big]$$

for every $\mathcal{F}_0^{A,B,C}$-predictable and bounded process h, then

$$E_{P_\theta}\Big[\int_0^T g_t Y_t dB_t\Big] = E_{P_\theta}\Big[\int_0^T g_t Y_t\big(\psi_t^0 + \psi_t^L\theta^*\frac{\sigma_t^1}{\sigma_t^2} + \theta^* A\psi_t^A + \theta^* K_{t-}\psi_t^K\big)dt\Big] \tag{6.8}$$

for every $\mathcal{F}_t^{B,C}$-predictable and bounded process g.

Sketch of proof. By Theorem 1, there exist an $\mathcal{F}_0^A$-measurable random variable W_0^1 and an $\mathcal{F}_0^{A,K,L}$-measurable random variable W_0^2 such that

$$\frac{dP_\theta|_{\mathcal{F}_0}}{dP|_{\mathcal{F}_0}} = W_0^1 W_0^2 \text{ and } \frac{dP_\theta|_{\mathcal{F}_0^{L,A}}}{dP|_{\mathcal{F}_0^{L,A}}} = W_0^1.$$

If H_1 is $\mathcal{F}_0^L$-measurable, $\tilde{H}_1 := E_P\big[H_1\big|\mathcal{F}_0^A\big]$ and H_2 is $\mathcal{F}_0^A$-measurable then

$$E_{P_\theta}\big[H_1H_2\big] = E_P\big[H_1H_2W_0^1\big] = E_P\big[\tilde{H}_1H_2W_0^1\big] = E_{P_\theta}\big[\tilde{H}_1H_2\big]. \tag{6.9}$$

Similarly, let h be a bounded and $\tilde{\mathcal{F}}_t$-predictable process and let $\mu_s^B := \tilde{Y}_s\big(\psi_s^0 + A\psi_s^A + L\psi_s^L + K_{s-}\psi_s^K\big)$ and note that

$$\begin{aligned}
E_{P_\theta}\Big[\int_0^T h_s dB_s\Big] &= E_P\Big[\int_0^T h_s dB_s W_T\Big] \\
&= E_P\Big[\int_0^T h_s W_{s-} dB_s\Big] + E_P\Big[\int_0^T h_s d[B,W]_s\Big] \\
&= E_P\Big[\int_0^T h_s W_{s-} dB_s\Big] \\
&= E_P\Big[\int_0^T h_s W_{s-}\mu_s^B ds\Big] \\
&= E_P\Big[\int_0^T h_s \mu_s^B ds W_T\Big] \text{ by [JS03, Proposition I 3.14]} \\
&= E_{P_\theta}\Big[\int_0^T h_s\mu_s^B ds\Big].
\end{aligned}$$

One can show that there exists an intermediate probability measure $\tilde{P}$ on $\tilde{\mathcal{F}}_T$ such that

(1)
$$P_\theta|_{\tilde{\mathcal{F}}_T} \ll \tilde{P} \ll P|_{\tilde{\mathcal{F}}_T}$$

(2) for every bounded and Borel-measurable function h:
- $E_{\tilde{P}}\big[h(A)\big] = h(\theta^* A)$, $\tilde{P}$-a.s.,
- $E_{\tilde{P}}\big[h(L)\big|\mathcal{F}_0^A,\big] = E_P\big[h(L)\big|\mathcal{F}_0^A\big]$,

- $E_{\tilde{P}}\big[h(K_0)\big|\mathcal{F}_0^{A,L}\big] = E_P\big[h(K_0)\big|\mathcal{F}_0^{A,L}\big]$,
- $E_{\tilde{P}}\big[h(B_0)\big|\mathcal{F}_0^{A,L,K}\big] = E_P\big[h(B_0)\big|\mathcal{F}_0^{A,L,K}\big]$.

(3) Whenever h is a bounded and $\tilde{\mathcal{F}}_t$-predictable process, then

- $$E_{\tilde{P}}\Big[\int_0^T h_s dB_s\Big] = E_{\tilde{P}}\Big[\int_0^T h_s \mu_s^B ds\Big],$$
- If μ^K and μ^C are $\tilde{\mathcal{F}}_t$-predictable processes such that

$$E_P\Big[\int_0^T h_s dK_s\Big] = E_P\Big[\int_0^T h_s \mu_s^K dU_s^K\Big]$$
$$E_P\Big[\int_0^T h_s dC_s\Big] = E_P\Big[\int_0^T h_s \mu_s^C dU_s^C\Big],$$

then

$$E_{\tilde{P}}\Big[\int_0^T h_s dK_s\Big] = E_{\tilde{P}}\Big[\int_0^T h_s \mu_s^K dU_s^K\Big]$$
$$E_{\tilde{P}}\Big[\int_0^T h_s dC_s\Big] = E_{\tilde{P}}\Big[\int_0^T h_s \mu_s^C dU_s^C\Big].$$

Note that by [Jac75, Proposition 4.3], there exists an $\mathcal{F}_t^{A,L,B}$-adapted $\tilde{P}$-martingale Ξ such that

$$\Xi_T = \frac{dP_\theta|_{\mathcal{F}_T^{A,L,B,C}}}{d\tilde{P}|_{\mathcal{F}_T^{A,L,B,C}}},\ \big[B,\Xi\big] = 0$$

and

$$Y\Xi_- = Y\exp\big(-\int_0^\cdot \theta^* K_r \psi_r^K dr\big). \tag{6.10}$$

Bayes formula with predictable projections shows that

$$E_{\tilde{P}}\Big[L\int_0^T Y_s h_s ds\Big] = E_{\tilde{P}}\Big[\int_0^T Y_s h_s \frac{\sigma_s^1}{\sigma_s^2} ds\Big] \tag{6.11}$$

for every bounded and $\mathcal{F}_t^{A,B,C}$-predictable process h. Now,

$$\begin{aligned}
E_{P_\theta}\Big[\int_0^T Lh_s Y_s ds\Big] &= E_{\tilde{P}}\Big[\int_0^T Lh_s Y_s ds \Xi_T\Big] \\
&= E_{\tilde{P}}\Big[\int_0^T \Xi_{s-} Lh_s Y_s ds\Big] \\
&= E_{\tilde{P}}\Big[\int_0^T \Xi_{s-} h_s Y_s \frac{\sigma_s^1}{\sigma_s^2} ds\Big] \text{ by (6.10)} \\
&= E_{\tilde{P}}\Big[\int_0^T h_s Y_s \frac{\sigma_s^1}{\sigma_s^2} ds \Xi_T\Big] \\
&= E_{P_\theta}\Big[\int_0^T h_s Y_s \frac{\sigma_s^1}{\sigma_s^2} ds\Big],
\end{aligned}$$

for every bounded $\mathcal{F}_t^{A,B,C}$-predictable process h, which implies that (6.9) holds. □

6.1.3. *Consistency of the Modified Sequential G-Estimator.* We are now able to show that the Modified Sequential G-Estimator suggested in [MVGH11] is uniformly consistent, also when we consider a time-dependent mediating treatment K. Let θ_1, θ_2 be two actions as in the previous proposition, but where $\theta_1^* A = 0$ and $\theta_2^* A = 1$ and consider corresponding $\mathcal{F}_t^{A,B,C}$-predictable processes γ^1 and γ^2 as the fractions in (6.11). Furthermore, we assume that our observations consist of the event histories for n independent equally distributed individuals, following the current generic model. We will also slightly misuse the notation and let N, from now on, denote the corresponding counting process that is aggregated over the n independent individuals.

Lemma 3. *Let $\widehat{\Psi}^0, \widehat{\Psi}^A, \widehat{\Psi}^L$ and $\widehat{\Psi}^K$ denote the usual additive regression estimators of Aalen, let $\tilde{Y} := Y_t^B Y_t^C$ and define*

$$\tilde{M}_t := N_t^B - \int_0^t \mu_s^B ds,\ \overline{\gamma}_t := \begin{pmatrix} \gamma_t^1 \\ \gamma_t^2 \end{pmatrix},\ \Gamma_t := \begin{pmatrix} \Psi_t^0 + \int_0^t \rho_s^1 d\Psi_s^L \\ \Psi_t^A + \int_0^t \rho_s^2 - \rho_s^1 d\Psi_s^L \end{pmatrix}$$

$$\widehat{H}_t := diag \begin{pmatrix} \tilde{Y}_t^1 \exp\big(\int_0^t K_{s-}^1 d\widehat{\Psi}_s^K\big) \\ \vdots \\ \tilde{Y}_t^n \exp\big(\int_0^t K_{s-}^n d\widehat{\Psi}_s^K\big) \end{pmatrix}\ H_t := diag \begin{pmatrix} \tilde{Y}_t^1 \exp\big(\int_0^t K_{s-}^1 d\Psi_s^K\big) \\ \vdots \\ \tilde{Y}_t^n \exp\big(\int_0^t K_{s-}^n d\Psi_s^K\big) \end{pmatrix}$$

$$Z_t := \tilde{Y}_t \cdot \begin{pmatrix} 1 & A_1 \\ \vdots & \\ 1 & A_n \end{pmatrix},\ Z_{s-}^{\widehat{H}} := \big(Z_{s-}^T \widehat{H}_{s-} Z_{s-}\big)^{-1} Z_{s-}^T \widehat{H}_{s-},$$

$$Z_{s-}^H := \big(Z_{s-}^T H_{s-} Z_{s-}\big)^{-1} Z_{s-}^T H_{s-} \ \ and\ \widehat{\Gamma}_t := \int_0^t Z_{s-}^{\widehat{H}} dN_s - \int_0^t Z_{s-}^{\widehat{H}} K_{s-} d\widehat{\Psi}_s^K.$$

We have that

$$\lim_{n\to\infty} P\big(\sup_{t\le T} \big|\widehat{\Gamma}_t - \Gamma_t\big| \ge \delta\big) = 0 \tag{6.12}$$

for every $\delta > 0$.

Proof. Note that

$$\widehat{\Gamma}_t - \Gamma_t = \int_0^t Z_{s-}^{\widehat{H}} dN_s - \int_0^t Z_{s-}^{\widehat{H}} K_{s-} d\widehat{\Psi}_s^K - \Gamma_t \tag{6.13}$$

$$= \int_0^t Z_{s-}^{\widehat{H}} - Z_{s-}^H dN_s + \int_0^t \big(Z_{s-}^H - Z_{s-}^{\widehat{H}}\big) K_{s-} d\Psi_s^K \tag{6.14}$$

$$+ \int_0^t Z_{s-}^H d\tilde{M}_s + \int_0^t Z_{s-}^{\widehat{H}} K_{s-} d\big(\Psi_s^K - \widehat{\Psi}_s^K\big) \tag{6.15}$$

$$+ \int_0^t Z_{s-}^H \begin{pmatrix} Z_{s-} & L_{s-} \end{pmatrix} d \begin{pmatrix} \Psi_s^0 \\ \Psi_s^A \\ \Psi_s^L \end{pmatrix} - \Gamma_t \tag{6.16}$$

Let

$$V = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

We have that $V^T Z_{s-}^T H_{s-} Z_{s-} V = S_{t-}$ where S_{t-} is a 2×2-diagonal matrix. Moreover, $\big(Z_{s-}^T H_{s-} Z_{s-}\big)^{-1} = V S_{t-} V^T$.

Note that $\big|\int_0^\cdot Z^H_{s-}d\tilde{M}_s\big|_2^2$ is Lenglart dominated by $\mathrm{Tr}\langle\int_0^\cdot Z^H_{s-}d\tilde{M}_s\rangle$ and

$$\begin{aligned}&\mathrm{Tr}\langle\int_0^\cdot Z^H_{s-}d\tilde{M}_s\rangle_T\\ =&\int_0^T \mathrm{Tr}\big(Z^T_{s-}H_{s-}Z_{s-}\big)^{-1}Z^T_{s-}H_{s-}\mathrm{diag}\mu H_{s-}Z_{s-}\big(Z^T_{s-}H_{s-}Z_{s-}\big)^{-1}ds\\ \leq&\int_0^T \mathrm{Tr}\big(Z^T_{s-}H_{s-}Z_{s-}\big)^{-1}\|\mathrm{diag}\mu H_{s-}\|_{\mathrm{op}}ds\end{aligned}$$

which converges in probability to 0. By Lenglart's inequality[JS03], we obtain that $\int_0^\cdot Z^H_{s-}d\tilde{M}_s$ converges uniformly to 0 in probability with respect to P.

Since

$$\lim_{\delta\to\infty}P\Big(\sup_s|Z_sK_s|\geq\delta\Big)=0 \text{ and } \lim_{\delta\to\infty}P\Big(\sup_s|\widehat{Z}_sK_s|\geq\delta\Big)=0$$

and $\widehat{\Psi}^K$ converges uniformly in probability to Ψ^K, we also have that

$$\int_0^t Z^{\widehat{H}}_{s-}K_{s-}d\big(\Psi^K_s-\widehat{\Psi}^K_s\big) \text{ and } \int_0^t\big(Z^H_{s-}-Z^{\widehat{H}}_{s-}\big)K_{s-}d\Psi^K_s$$

converge uniformly in probability to 0 w.r.t. P. This shows that (6.15) converges uniformly in probability to 0 w.r.t. P as well.

We have that

$$Z^H_{s-}L_{s-}=VS^{-1}_{s-}\begin{pmatrix}\sum_{i=1}^n H^i_{s-}L^i_{s-}(1-A^i)\\ \sum_{i=1}^n H^i_{s-}L^i_{s-}A^i\end{pmatrix}=V\begin{pmatrix}\frac{\sum_{i=1}^n H^i_{s-}L^i_{s-}(1-A^i)}{\sum_{i=1}^n H^i_{s-}(1-A^i)}\\ \frac{\sum_{i=1}^n H^i_{s-}L^i_{s-}A^i}{\sum_{i=1}^n H^i_{s-}A^i}\end{pmatrix}.$$

The law of large numbers implies that $Z^H_{s-}L_{s-}$ converges in P-probability to $V\overline{\gamma}(s)$. Now, (6.16) equals

$$\int_0^t Z^H_{s-}L_{s-}-V\gamma(s)d\Psi^L_s \tag{6.17}$$

and

(6.18)

$$E_P\big[\sup_t\big|\int_0^t Z^H_{s-}L_{s-}-V\gamma(s)d\Psi^L_s\big|\big]\leq\int_0^T E_P\big[\big|\big(Z^H_{s-}L_{s-}-V\gamma(s-)\big)\big|\big]|\psi^L_s|ds.$$

Therefore (6.16) converges uniformly in probability w.r.t. P.

A computation shows that $|\int_0^\cdot Z^{\widehat{H}}_{s-}-Z^H_{s-}dN_s|_1$ is Lenglart dominated by

$$\begin{aligned}\|V\|_1\int_0^\cdot\sum_j\Big|&\frac{\widehat{H}^j_{s-}\big(1-A_j\big)\mu^j_s}{\sum_i\widehat{H}^i_{s-}\big(1-A_i\big)}-\frac{H^j_{s-}\big(1-A_j\big)\mu^j_s}{\sum_i H^i_{s-}\big(1-A_i\big)}\Big|\\ &+\Big|\frac{\widehat{H}^j_{s-}A_j\mu^j_s}{\sum_i\widehat{H}^i_{s-}A_i}-\frac{H^j_{s-}A_j\mu^j_s}{\sum_i H^i_{s-}A_i}\Big|ds.\end{aligned}$$

This process converges uniformly in probability to 0, so we see that (6.14) also converges uniformly in probability to 0. This means that $\widehat{\Gamma}-\Gamma$ converges uniformly in probability to 0, so $\widehat{\Gamma}$ actually converges to Γ in the similar sense.

□

The cumulative P_{θ_i}-hazard of $\tilde{B}$ is given by

$$\Lambda^{\theta_i}_t=\int_0^t\psi^0_s+\theta^*_iA\psi^A_s+\theta^*_iK_s\psi^K_s+\gamma_s\psi^L_sds \tag{6.19}$$

Since stochastic integrals are continuous with respect to uniform convergence on compacts in probability, we see that

$$\lim_{\delta\to 0} P\left(\sup_t \Big| \int_0^t \begin{pmatrix} 1, & \theta_i^* A, & \theta_i^* K_{s-} \end{pmatrix} \begin{pmatrix} d\widehat{\Gamma}_s \\ d\widehat{\Psi}_s^K \end{pmatrix} - \Lambda_t^{\theta_i} \Big| \geq \delta \right) = 0,$$

i.e. we obtain a consistent estimator of $\Lambda_t^{\theta_i}$. A consistent estimator for the controlled direct effect of A on B is given by the second component of $\widehat{\Gamma}$.

7. Discussion

The primary concern in this paper is the possibility of estimating parameters for the counterfactual situation from the observational data, given that the counterfactual model is correct. This comes mainly down to whether the counterfactual probability is absolutely continuous with respect to the factual probability and whether the counterfactual parameters of interest are identifiable. The previously mentioned related works by Arjas and Parner, [Arj12] and [AP04] construct counterfactual probability distributions by piecing stochastic intervals together as in [Jac75, Section 3]. Unlike Parner and Arjas, we take a more martingale oriented approach, also based on the seminal paper [Jac75]. This enables us to apply directly already well established methods from stochastic analysis and martingale theory. In fact, surprisingly much causal inference can be well understood in terms of martingale measures, Bayes' rule and Girsanov's Theorem. This approach translates directly the problem about data re-weighting into a thoroughly studied problem in the literature, i.e. whether the stochastic exponential of a local martingale defines a martingale, see [LM78] and [KS02].

Another difference from the work of Arjas and Parner is that we consider an explicit intervention in terms of a transformation θ on sample space. While not being absolutely necessary, it still provides additional clarification as it makes the notion of counterfactual outcomes more explicit, or perhaps even demystified. The notation $\mathrm{do}(X = x)$, [Pea09], is simply interpreted as the measurable transformation on the sample space that forces every outcome of X into x and leaves the remaining observations unchanged. When the action becomes more complex than just forcing a variable into a fixed value, this interpretation becomes even more appealing.

The introduction of the transformation θ sheds some light on another aspect: One may in fact think of a causal inference problem as a stochastic control problem, or a decision problem, where the assumptions about the model are kept as modest as possible. The main objective in stochastic control theory is to find an optimal intervention strategy and compute the corresponding expected payoff. Causal inference appears as a special case of this, in the sense that there one mostly considers only one intervention strategy, namely the transformation θ, and aims to compute the expected payoff.

One is often confronted with latent factors in epidemiological settings. This lack of information typically yields non-identifiable effects. In special situations, one can use graphical arguments to ensure identifiability of counterfactual parameters and also provide exact formulas for these. Such examples are the *back-door formula*, *front-door formula* and *sequential back-door formula*, [Pea09, 3.3.1,3.3.2,4.4.3] and [ED10]. We show that we may take advantage of the local independence graphs to identify causal effects in event-history analysis.

When the counterfactual effect is possibly unidentifiable, one may try to compute upper and lower bounds for this. This can also be thought of as a control problem where "the nature" is allowed to control the latent factors in order to maximize or minimize counterfactual effects. This corresponds to an optimization problem under

constraints. The latent variables may only be altered in such a way that the observable factors maintain the same joint distribution and also such that some given directed graph constantly defines a local independence graph. Let $\mathcal{S}$ denote the set of counterfactual distributions corresponding to these constraints. The "causal effect" would then be sandwiched by $\inf_{P'\in\mathcal{S}} E_{P'}\big[\eta\big] \leq E_{P_\theta}\big[\eta\big] \leq \sup_{P''\in\mathcal{S}} E_{P''}\big[\eta\big]$.

The set $\mathcal{S}$ may have a somewhat complicated geometry. If one instead considers the convex hull, we obtain other, not necessarily, tight bounds:

$$\inf_{P'\in\mathrm{conv}(\mathcal{S})} E_{P'}\big[\eta\big] \leq E_{P_\theta}\big[\eta\big] \leq \sup_{P''\in\mathrm{conv}(\mathcal{S})} E_{P''}\big[\eta\big].$$

These bounds may be computed by allready developed linear programing techniques. This approach was for instance taken in [BP97], but is likely to generalize to more complicated continuous-time scenarios as well.

Appendix

Uniqueness of counterfactual distributions.

Lemma 4. *There exists at most one counterfactual distribution P_θ on $\mathcal{F}_0$ that imposes contemporaneously independent outcomes.*

Proof. Let $T_1,\dots,T_m$ be an enumeration of $\{T(V)\}_{V\in\mathcal{V}}$ such that $j<k$ implies $T_j<T_k$.

Assume that P' and P'' are two counterfactual distributions that have contemporaneously independent outcomes and η is an $\mathcal{F}_0^{V_k}$-measurable random variable. Let $\{X_i\}_i$ be an enumeration of $\{V\in\mathcal{X}|T(V)=T_1\}$ and let $\{A_j\}_j$ be an enumeration of $\{V\in\mathcal{A}|T(V)=T_1\}$. Whenever $\{h_i\}_i$ and $\{g_j\}_j$ are two families of bounded and measurable functions, then

$$\begin{aligned}
&E_{P'}\big[\prod_i h_i(X_i)\prod_j g_l(A_j)\big] = E_{P'}\big[\prod_i h_i(X_i)\big]E_{P'}\big[\prod_j g_j(A_j)\big]\\
=&\prod_i E_{P'}\big[h_i(X_i)\big]E_{P'}\big[\prod_j g_j(A_j)\big] = \prod_i E_{P''}\big[h_i(X_i)\big]E_{P''}\big[\prod_j g_j(A_j)\big]\\
=&E_{P''}\big[\prod_i h_i(X_i)\big]E_{P''}\big[\prod_j g_j(A_j)\big] = E_{P''}\big[\prod_i h_i(X_i)\prod_j g_j(A_j)\big]
\end{aligned}$$

This shows that if η is a bounded random variable that only depends on the information at T_1, then $E_{P'}[\eta]=E_{P''}[\eta]$. We continue with an induction argument and assume that $E_{P'}\big[\eta\big]=E_{P''}\big[\eta\big]$ for every bounded and random variable η that only depends on $\{V\in\mathcal{V}|T(V)<T_k\}$ and aim to prove that this also holds if η depends on the information at time T_k. Let $\{X_i\}_i$ be an enumeration of $\{V\in\mathcal{X}|T(V)=T_k\}$ and let $\{A_j\}_j$ be an enumeration of $\{V\in\mathcal{A}|T(V)=T_k\}$. Whenever $\{h_i\}_i$ and $\{g_j\}_j$

are two families of bounded and measurable functions, then

$$\begin{aligned}
&E_{P'}\big[\eta\prod_i h_i(X_i)\prod_j g_j(A_j)\big]\\
=&E_{P'}\big[\eta E_{P'}\big[\prod_i h_i(X_i)\big|\mathcal{F}_0^{p(V_1)}\big]\prod_j \theta^* g_j(A_j)\big]\\
=&E_{P'}\big[\eta\prod_i E_{P'}\big[h_i(X_i)\big|\mathcal{F}_0^{p(V_1)}\big]\prod_j \theta^* g_j(A_j)\big]\\
=&E_{P''}\big[\eta\prod_i E_{P''}\big[h_i(X_i)\big|\mathcal{F}_0^{p(V_1)}\big]\prod_j \theta^* g_j(A_j)\big]\\
=&E_{P''}\big[\eta E_{P''}\big[\prod_i h_i(X_i)\big|\mathcal{F}_0^{p(V_1)}\big]\prod_j \theta^* g_j(A_j)\big]\\
=&E_{P''}\big[\eta\prod_i h_i(X_i)\prod_j g_j(A_j)\big]
\end{aligned}$$

This proves the induction hypothesis, i.e. $E_{P'}[\eta] = E_{P''}[\eta]$ whenever η depends on $\{V \in \mathcal{A}|T(V) \leq T_k\}$.

□

Theorem 4. *There exists at most one probability measure on $\mathcal{F}_T$ that simultaneously satisfies* (3.4), (3.5), (3.8) *and* (3.9).

Proof. Recall the definition (4.16). (3.8) and (3.9) imply that

$$E_{P_\theta}\Big[\int_J\int_0^T h(s,x)N(ds,dx)\Big] = E_{P_\theta}\Big[\int_J\int_0^T h(s,x)\nu^\theta(ds,dx)\Big]. \tag{7.1}$$

Now [Jac75, Theorem 3.4] implies that there exists at most one probability measure on $\mathcal{F}_T$ that coincides with P_θ on $\mathcal{F}_0$ and satisfies (7.1).

□

Dual predictable projections.

Lemma 5. *Let U denote the dual predictable projection of N with respect to Q onto the filtration $\mathcal{F}_t$.*

(1) *If h is a bounded and $\mathscr{P}^V$ measurable processes then*

$$\int_{J_V}\int_0^\cdot h(s,x)U(ds,dx)$$

defines an $\mathcal{F}_t^V$-predictable process of finite variation.

(2) *If h and h' are bounded and $\mathscr{P}\otimes\mathcal{J}$ measurable processes then*

$$\Big[\int_{J_V}\int_0^\cdot h(s,x)U(ds,dx), \int_{J_{V'}}\int_0^\cdot h'(s,x)U(ds,dx)\Big] = 0 \tag{7.2}$$

$$\Big[\int_{J_V}\int_0^\cdot h(s,x)U(ds,dx), \int_{J_{V'}}\int_0^\cdot h'(s,x)N(ds,dx)\Big] = 0 \tag{7.3}$$

Q-a.s. whenever $V \neq V'$.

(3) *There exists a non-negative and $\mathscr{P}\otimes\mathcal{J}$-measurable process λ such that*

$$E_P\Big[\int_J\int_0^T h(s,x)N(ds,dx)\Big] = E_P\Big[\int_J\int_0^T h(s,x)\lambda(s,x)U(ds,dx)\Big]$$

for every bounded and $\mathscr{P}\otimes\mathcal{J}$-measurable process h.

Proof. The integral equation

$$\int_J \int_0^T h(s,x) N^V(ds,dx) = \int_{J_V} \int_0^T h(s,x) N(ds,dx) \tag{7.4}$$

defines a multivariate point process N^V with mark space J which only jumps at marks in J_V. Theorem [Jac75, 2.1] provides a dual predictable projection U^V of N^V with respect to the reference measure Q onto the filtration $\mathcal{F}_t^V$.

Let h be a bounded and $\mathscr{P} \otimes \mathcal{J}$ measurable process. Theorem [JS03, I 2.2.ii] and a monotone class argument provides a bounded and $\mathscr{P}^V$-measurable process h^V such that

$$\tilde{h}(\cdot,\cdot) = E_Q\big[h(\cdot,\cdot)\big|\mathcal{F}_T^V\big],\ Q\text{-a.s.}$$

Now,

$$\begin{aligned}
E_Q\Big[\int_{J_V}\int_0^T h(s,x)U(ds,dx)\Big] &= E_Q\Big[\int_{J_V}\int_0^T h(s,x)N(ds,dx)\Big] \\
&= E_Q\Big[\int_J\int_0^T h(s,x)N^V(ds,dx)\Big] \\
&= E_Q\Big[\int_J\int_0^T \tilde{h}(s,x)N^V(ds,dx)\Big] \\
&= E_Q\Big[\int_J\int_0^T \tilde{h}(s,x)U^V(ds,dx)\Big] \\
&= E_Q\Big[\int_J\int_0^T h(s,x)U^V(ds,dx)\Big]
\end{aligned}$$

which proves the first claim.

To prove (7.2), let $W \subset J_V$ and $W' \subset J_{V'}$ be measurable subsets and consider the corresponding counting processes

$$N_t^W := N([0,t],W) \text{ and } N_t^{W'} := N([0,t],W')$$

and let

$$U_t^W := U([0,t],W) \text{ and } U_t^{W'} := U([0,t],W').$$

Following [Jac75, Proposition 2.3], we see that

$$\Delta U_s^W = E_Q\big[\Delta N_s^W\big|\mathcal{F}_{s-}\big] \text{ and } \Delta U_s^{W'} = E_Q\big[\Delta N_s^{W'}\big|\mathcal{F}_{s-}\big],\ Q\text{-a.s.}$$

Now,

$$\begin{aligned}
0 \le E_Q\Big[\big[U^W,U^{W'}\big]_T\Big] &= E_Q\Big[\sum_{s\le T}\Delta U_s^W \Delta U_s^{W'}\Big] \\
&\le \sum_{s\le T} E_Q\Big[\Delta U_s^W \Delta U_s^{W'}\Big] \text{ by Fatou's Lemma} \\
&= \sum_{s\le T} E_Q\Big[E_Q\big[\Delta N_s^W\big|\mathcal{F}_{s-}\big]E_Q\big[\Delta N_s^{W'}\big|\mathcal{F}_{s-}\big]\Big] \\
&= \sum_{s\le T} E_Q\Big[\Delta N_s^W \Delta N_s^{W'}\Big] \\
&= 0,
\end{aligned}$$

so $[U^W,U^{W'}] = 0$, Q-a.s.

Whenever f and f' are bounded and $\mathcal{F}_t$-predictable processes we have

$$\Big[\int_0^\cdot f_s dU_s^W, \int_0^\cdot f'_s dU_s^{W'}\Big] = \int_0^\cdot f_s f'_s d\big[U^W,U^{W'}\big]_s = 0,\ Q\text{-a.s.} \tag{7.5}$$

The equation (7.2) is therefore satisfied in the special case with $h = f \cdot \chi_W$ and $h' = f' \cdot \chi_{W'}$. The general case now follows from an application of the Monotone class Theorem. The equation (7.3) follows from an almost similar argument.

For the last claim, let ν denote the dual predictable projection of N with respect to P onto the filtration $\mathcal{F}_t$ and note that $\nu \ll U$ since $P \ll Q$. The existence of λ then follows directly from [Jac75, Theorem 4.1]. □

Acknowledgements

Supported by the Research Council of Norway. Project: 191460/V50. The author would like to thank Prof. Odd O. Aalen and Prof. Torben Martinussen for very helpful discussions on this project.

References

[Aal87] Odd O. Aalen. Dynamic modelling and causality. *Scand. Actuar. J.*, (3-4):177–190, 1987.

[ABGK93] Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

[AP04] Elja Arjas and Jan Parner. Causal reasoning from longitudinal data. *Scand. J. Statist.*, 31(2):171–187, 2004.

[ARGL12] O.O. Aalen, K. Røysland, J.M. Gran, and B. Ledergerber. Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2012.

[Arj12] Elja Arjas. *Causality: Statistical Perspectives & Applications*, chapter Causal inference from longitudinal data. Wiley-Blackwell, 2012.

[BP97] Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92:1171–1177, 1997.

[Bré81] Pierre Brémaud. *Point processes and queues*. Springer-Verlag, New York, 1981. Martingale dynamics, Springer Series in Statistics.

[CH08] S.R. Cole and M.A. Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656, 2008.

[DFS11] Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *Ann. Statist.*, 39(2):865–886, 2011.

[Did00] Vanessa Didelez. *Graphical models for event history analysis based on local independence*. Logos Verlag Berlin, Berlin, 2000. Dissertation, Universität Dortmund, Dortmund, 2000.

[Did08] Vanessa Didelez. Graphical models for marked point processes based on local independence. *J.R Statist. Soc. B*, 70(1):245–264, 2008.

[ED10] Michael Eichler and Vanessa Didelez. On Granger causality and the effect of interventions in time series. *Lifetime Data Anal.*, 16(1), 2010.

[HV06] Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence*, pages 217–224. Citeseer, 2006.

[Jac75] Jean Jacod. Multivariate point processes: predictable projection, Radon-Nikodým derivatives, representation of martingales. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 31:235–253, 1974/75.

[JS03] Jean Jacod and Albert N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003.

[KS02] Jan Kallsen and Albert N. Shiryaev. The cumulant process and Esscher's change of measure. *Finance Stoch.*, 6(4):397–428, 2002.

[Lau01] Steffen L. Lauritzen. Causal inference from graphical models. In *Complex stochastic systems (Eindhoven, 1999)*, volume 87 of *Monogr. Statist. Appl. Probab.*, pages 63–107. Chapman & Hall/CRC, Boca Raton, FL, 2001.

[LM78] Dominique Lépingle and Jean Mémin. Sur l'intégrabilité uniforme des martingales exponentielles. *Z. Wahrsch. Verw. Gebiete*, 42(3):175–203, 1978.

[Lok08] Judith J. Lok. Statistical modeling of causal effects in continuous time. *Ann. Statist.*, 36(3):1464–1507, 2008.

[MVGH11] Torben Martinussen, Stijn Vansteelandt, Mette Gerster, and Jacob Hjelmborg. Estimation of direct effects for survival data using the aalen additive hazards model. *To appear in: Journal of the Royal Statistical Society: Series B*, 2011.

[PA99] J. Parner and E. Arjas. *Causal reasoning from longitudinal data*. Rolf Nevanlinna Inst., Univ. of Helsinki, 1999.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009. Models, reasoning, and inference.

[Pea10] J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):7, 2010.

[Pro05] Philip E. Protter. *Stochastic integration and differential equations*, volume 21 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2005. Second edition. Version 2.1, Corrected third printing.

[RHB00] J.M. Robins, M.A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, Sep 2000.

[Røy11] K. Røysland. A martingale approach to continuous time marginal structural models. *Bernoulli*, 17(3):895–915, 2011.

[Sch70] Tore Schweder. Composable Markov processes. *J. Appl. Probability*, 7:400–410, 1970.

[SHL+05] J.A. Sterne, M.A. Hernán, B. Ledergerber, K. Tilling, R. Weber, P. Sendi, M. Rickenbach, J.M. Robins, and M. Egger. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *Lancet*, 366:378–384, 2005.

[SP06] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial intelligence*, 2006.

University of Oslo, Institute of Basic Medical Sciences, Department of Biostatistics, Boks 1122 Blindern, 0317 Oslo, Norway

E-mail address: kjetil.roysland@medisin.uio.no