Biometrika (2023), **00**, 0, *pp.* 1–20 https://doi.org/10.1093/biomet/asad005
Advance Access Publication 23 January 2023

An instrumental variable method for point processes: generalized Wald estimation based on deconvolution

By ZHICHAO JIANG

School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong 510275, China
jiangzhch7@mail.sysu.edu.cn

SHIZHE CHEN

Department of Statistics, University of California, Davis, One Shields Avenue, Davis, California 95616, U.S.A.
szdchen@ucdavis.edu

and PENG DING

Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, U.S.A.
pengdingpku@berkeley.edu

Summary

Point processes are probabilistic tools for modelling event data. While there exists a fast-growing literature on the relationships between point processes, how such relationships connect to causal effects remains unexplored. In the presence of unmeasured confounders, parameters from point process models do not necessarily have causal interpretations. We propose an instrumental variable method for causal inference with point process treatment and outcome. We define causal quantities based on potential outcomes and establish nonparametric identification results with a binary instrumental variable. We extend the traditional Wald estimation to deal with point process treatment and outcome, showing that it should be performed after a Fourier transform of the intention-to-treat effects on the treatment and outcome, and thus takes the form of deconvolution. We refer to this approach as generalized Wald estimation and propose an estimation strategy based on well-established deconvolution methods.

Some key words: Causal inference; Identification; Intensity; Principal stratification; Unmeasured confounding.

1. Introduction

Point processes have long been used for modelling event data. The past decade has witnessed a surge of interest in point process models in many fields, including neuroscience, finance and the social sciences. In this paper, we consider the analysis of neural data as a concrete motivation. Modern technologies allow neuroscientists to simultaneously record neural spike trains, i.e., arrays of timestamps of when neurons fire, across the brain. With these data, one can hope to gain insight into the mechanisms of neural computing. The essence of such scientific questions is the inference of causal effects.

Current technologies, however, present a major challenge for causal inference with neural data. Except for experiments on very simple animals, even state-of-the-art technologies can record only a very small fraction of neurons in chosen regions of the nervous system, leaving the vast majority unobserved. These unmeasured neural activities inevitably lead to the issue of unmeasured confounding; that is, unmeasured activities could be the common causes of observed neural activities. As a result, any relationship inferred based on the partially observed system may reflect, not the true causal relationship, but rather a spurious association.

Fortunately, advances in optogenetics have created new opportunities for dealing with unmeasured confounding. Neuroscientists are able to instigate neural activities in a living brain via optical stimulation, which alters the activity of any chosen neuron with high spatial and temporal precision (Mardinly et al., 2018; Carrillo-Reid et al., 2019). From a causal inference perspective, such interventions can serve as instrumental variables for inferring the causal relationships between neurons, as they affect the outcome neuron only through the treatment neuron while introducing exogenous variation in the treatment neuron.

Instrumental variable methods are powerful tools for inferring causal effects in the presence of unmeasured confounding between the treatment and the outcome. The seminal paper of Angrist et al. (1996) clarified the role of a binary instrumental variable in identifying the causal effect of a binary treatment for an unmeasured subgroup, known as the complier average causal effect. Angrist et al. (1996) proposed two crucial identification assumptions, monotonicity and exclusion restriction. Under these assumptions, they showed that the complier average causal effect is identified by the Wald estimator (Wald, 1940; Ridder & Moffitt, 2007), which equals the ratio of the differences in means of the outcome and the treatment when the instrumental variable changes from 0 to 1.

Most existing work on instrumental variables considers nondynamic settings, and the instrumental variable methods in survival analysis focus mainly on a scalar treatment and a nonrecurrent outcome (e.g., Li et al., 2015; Martinussen et al., 2017; Richardson et al., 2017; Jiang et al., 2018). To the best of our knowledge, there is no formal instrumental variable framework for point processes that addresses nonparametric identification.

We propose an instrumental variable method for causal inference in the case where both the treatment and the outcome take the form of point processes. We define several causal quantities for the effect of the treatment on the outcome over time. Using a binary instrumental variable, we establish the nonparametric identification of causal effects allowing for the unmeasured treatment-outcome confounding. The identification assumptions hold as long as the impact of the unmeasured confounders on the outcome is additive. Our identification result implies that the causal effects can be obtained by solving a convolution equation. This extends the Wald estimation in traditional instrumental variable methods so that it takes the form of deconvolution, leading to the proposed generalized Wald estimation. We also examine several commonly used models within our framework, studying the identification of the causal effects and the causal interpretation of the model parameters with a binary instrumental variable. When the unmeasured confounders are additive on the outcome, the causal effects are identifiable without any distributional assumptions on the confounders based on the proposed generalized Wald estimation. Our findings justify the identifiability of many commonly used models, such as the Hawkes process, broadening their applicability with fewer assumptions.

We use the following notation. Let $A \perp\!\!\!\perp B \mid C$ denote conditional independence of A and B given C. Let $\mathbb{R}$ denote the set of real numbers and $\mathcal{B}(\mathbb{R})$ the Borel σ-algebra of the whole real line. Let $L^1(\mathbb{R})$ denote the set of functions $f(x)$ such that $\int_{-\infty}^{\infty} |f(x)| \, \mathrm{d}x < \infty$.

Unless specified otherwise, in this paper we assume that all functions belong to $L^1(\mathbb{R})$. Let Ψ denote the Fourier transform, i.e., for any $f(x) \in L^1(\mathbb{R})$ and $v \in \mathbb{R}$, define

$$(\Psi f)(v) = \int_{-\infty}^{\infty} f(x) \exp(-\mathrm{i}2\pi v x)\, \mathrm{d}x,$$

where $\mathrm{i} = \sqrt{(-1)}$. Let Ψ^{-1} denote the inverse Fourier transform.

2. An instrumental variable framework for point processes

2.1. *A brief review of the binary instrumental variable model*

We begin by reviewing the binary instrumental variable model in the context of noncompliance (Angrist et al., 1996). For unit i, let Z_i be the binary treatment assigned, N_i the actual treatment received and Y_i the outcome of interest. Let N_{iz} be the potential value of the treatment received if the assigned treatment condition is z, and let Y_{izn} be the potential value of the outcome if the assigned treatment is z and the actually received treatment is n. The joint values of N_{i1} and N_{i0} define the unmeasured compliance type $U_i = (N_{i1}, N_{i0})$. Units with $(N_{i1} = 1, N_{i0} = 0)$ are compliers who take the treatment assigned, units with $(N_{i1} = 1, N_{i0} = 1)$ are always-takers who always take treatment 1, units with $(N_{i1} = 0, N_{i0} = 0)$ are never-takers who always take treatment 0, and units with $(N_{i1} = 0, N_{i0} = 1)$ are defiers who take the treatment opposite to that assigned.

Angrist et al. (1996) make three assumptions: (i) exclusion restriction, that the treatment assigned affects the outcome only through the treatment received, i.e., $Y_{izn} = Y_{iz'n}$ for all z, z' and n; (ii) randomization, that Z_i is independent of N_{iz} and Y_{izn} for $z, n = 0, 1$; (iii) monotonicity, that the assigned treatment does not negatively affect the treatment received for all units, i.e., $N_{i1} \geqslant N_{i0}$. Exclusion restriction simplifies Y_{izn} to Y_{in}. Randomization rules out confounding between the treatment assignment and the treatment received, as well as confounding between the treatment assignment and the outcome. Monotonicity rules out defiers. Under these assumptions, Angrist et al. (1996) introduce the complier average causal effect as the average effect of the treatment received on the outcome for compliers, $\textsc{cace} = E(Y_{i1} - Y_{i0} \mid N_{i1} = 1, N_{i0} = 0)$, and show that it is identified by

$$\textsc{cace} = \frac{E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)}{E(N_i \mid Z_i = 1) - E(N_i \mid Z_i = 0)}. \tag{1}$$

In this model, the treatment assignment Z_i is the instrumental variable. The expression in (1) suggests the Wald estimator (Wald, 1940) for the cace, i.e., the ratio of the differences in means of the outcome and the treatment received when the treatment assigned changes from 0 to 1.

Angrist et al. (1996) identify only the treatment effect in the complier subpopulation. For extrapolation to the whole population, we can invoke the homogeneity assumption (cf., Heckman, 1996; Chen et al., 2009) that the treatment effect is the same across compliance groups:

$$E(Y_{i1} - Y_{i0} \mid N_{i1}, N_{i0}) = E(Y_{i1} - Y_{i0}). \tag{2}$$

Under the assumption in (2), the treatment effect in the whole population equals the cace.

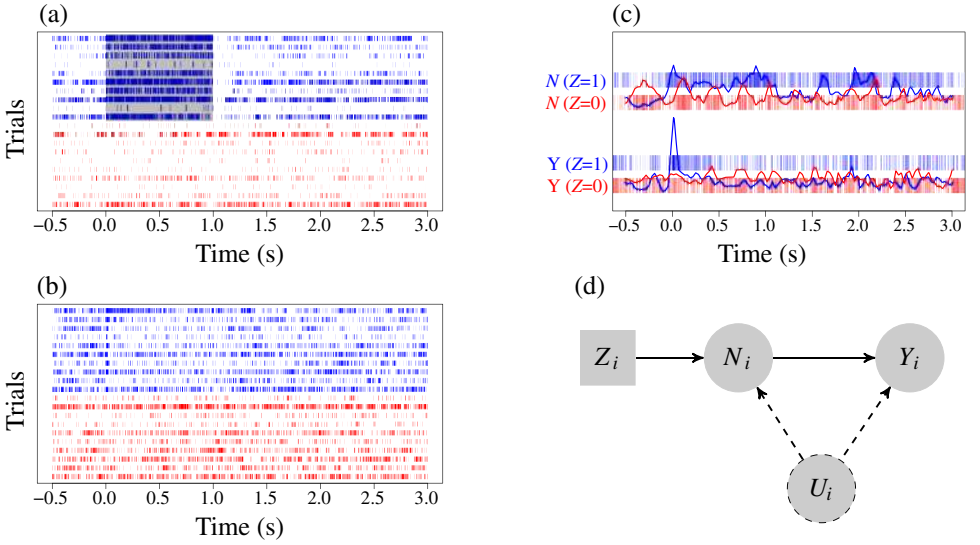

Fig. 1. Neural data from Bolding & Franks (2018a) and the causal diagram depicting the relationships among the variables in the instrumental variable framework: (a) spike trains collected in the mouse olfactory bulb in the stimulated (blue) and unstimulated (red) trials, where each row represents a spike train in the olfactory bulb in one trial (N_i) and the shaded area represents the duration of the light pulse; (b) spike trains collected in the mouse piriform cortex in the stimulated (blue) and unstimulated (red) trials, where each row represents a spike train in the piriform cortex in one trial (Y_i); (c) zoomed-in view of two randomly selected trials, where the solid curves are the smoothed intensities; (d) causal diagram for the relationships among the variables, where variables in the solid square and circles are observed, the variable in the dashed circle is unobserved, and the subscript i denotes the ith unit.

2.2. *Notation and basic assumptions with point process treatment*

We now consider a setting where both the treatment N_i and the outcome Y_i are point processes. We will establish a ratio relationship similar to that in (1) for the point process treatment and outcome, but in the frequency domain. As a concrete example, we consider the neuroscience application from Bolding & Franks (2018a). In this application, the treatment and the outcome are the neural activities of the mouse olfactory bulb and piriform cortex within each brain region, respectively. In experiments at the single-cell resolution, one can also model single-neuron activities as the treatment and outcome. As shown in Figs. 1(a) and (b), these data take the form of spike trains that are commonly modelled as point processes. Bolding & Franks (2018a) applied light pulses to randomly selected trials to stimulate the olfactory bulb without affecting other brain regions. Therefore, the light pulse serves as an instrumental variable. To formally discuss causal inference, we need to generalize the model in Angrist et al. (1996) to account for the point process treatment and outcome.

Let $i = 1, \ldots, m$ index the units. We use point processes to describe the neuronal activities; see, e.g., Cox & Isham (1980, Ch. 2) or Daley & Vere-Jones (2003, Ch. 3). Define the treatment point process $N_i(\cdot)$ as a family of random nonnegative integers $\{N_i(A)\}_{A \in \mathcal{B}(\mathbb{R})}$ that count the number of events in each set A. Let $\mathrm{d}N_i(t) \equiv N_i\{[t, t + \mathrm{d}t)\}$. Throughout this paper we consider point processes that are simple, $\mathrm{pr}\{\mathrm{d}N_i(t) = 0 \text{ or } 1 \text{ for all } t\} = 1$, and with bounded intensity, $\mathrm{pr}\{\mathrm{d}N_i(t) = 1\}/\mathrm{d}t < \infty$. In a similar manner, we introduce the outcome point process $Y_i(\cdot)$. We focus on a binary instrumental variable, $Z_i \in \{0, 1\}$, and present results for a discrete instrumental variable in the Supplementary Material. Without loss of generality, we assume that the onset of the instrumental variable is at time 0, and that $N_i(\cdot)$ and $Y_i(\cdot)$ are observed from 0 to T. To avoid cumbersome bookkeeping, we constrain the processes to the observed period and ignore the history before time 0, and we write $N_i([0, t])$ as simply $N_i(t)$.

For ease of discussion, we first consider the case where the treatment $N_i(\cdot)$ is a point process with at most one event. We refer to $N_i(\cdot)$ as a single-point process if $N_i(T) \leqslant 1$. We will extend the method to a general point process in § 3.2. We can characterize a single-point process $N_i(\cdot)$ using its event time: define $\mathcal{T}_i = T^+$ if $N_i(T) = 0$ and $\mathcal{T}_i = \tau$ if $N_i(t) = 1$ for $t \geqslant \tau$ and $N_i(t) = 0$ for $t < \tau$.

We adopt the potential outcomes framework under the following stable unit treatment value assumption (Rubin, 1980).

Assumption 1. There is no interference between units and there are no different versions of the instrumental variable and the treatment process.

Assumption 1 rules out the spillover effect of other units' instrumental variable on one's treatment process, and that of other units' instrumental variable and treatment process on one's outcome process. It also requires that there be only one version of the instrument and the treatment process. In our motivating example, one unit corresponds to one trial, and trials conducted at different times may use the same mouse. The no-interference assumption would be violated if the neural dynamics of a mouse were to adapt to stimulation over time, causing activities in one trial to depend on previous trials. This phenomenon is known as neural plasticity. To restrict spillover between trials, adequate washout periods are incorporated to separate trials sufficiently apart. As a result, we can reasonably neglect spillover effects. Furthermore, uniform stimulation is employed to ensure that there is only one version of the instrumental variable. For the treatment process, we follow the common practice in neural data analysis of focusing on the effect of the timings of spikes, ignoring the variation in spike intensities (cf. Brillinger, 1988; Yu et al., 2009; Wu et al., 2017; Zhao & Park, 2017).

Assumption 1 allows us to define the potential values as a function of a unit's own instrumental variable and treatment process. Let $N_{iz}(\cdot)$ and $Y_{iz}(\cdot)$ be the potential processes of the treatment and outcome, respectively, and let $\mathcal{T}_{iz}$ be the potential event time of the treatment process if the instrumental variable were set to $Z_i = z$. Also, define $Y_{iz\tau}(\cdot)$ as the potential process of the outcome if the instrumental variable were set to $Z_i = z$ and the event time were set to $\mathcal{T}_i = \tau$. By definition, the two versions of the potential outcome process satisfy $Y_{iz}(\cdot) = Y_{iz,\mathcal{T}_{iz}}(\cdot)$. The observed treatment process is $N_i(\cdot) = Z_i N_{i1}(\cdot) + (1 - Z_i) N_{i0}(\cdot)$, and the observed outcome process can be written as $Y_i(\cdot) = Z_i Y_{i1}(\cdot) + (1 - Z_i) Y_{i0}(\cdot)$ or $Y_i(\cdot) = Z_i Y_{i1\tau}(\cdot) + (1 - Z_i) Y_{i0\tau}(\cdot)$ if $\mathcal{T}_i = \tau$. We assume that $\{Z_i, N_{iz}(\cdot), Y_{iz\tau}(\cdot) : z = 0, 1; \tau \in [0, T] \cup T^+\}_{i=1}^m$ are independent and identically distributed, and hence the observables $\{Z_i, N_i(\cdot), Y_i(\cdot)\}_{i=1}^m$ are also independent and identically distributed. We abbreviate $N_i(\cdot)$ to N_i and $Y_i(\cdot)$ to Y_i when no confusion is likely to arise. In our motivating example, the experiment is carefully designed to ensure that the trials are independent and identically distributed. For instance, the optical stimulation is targeted at the same location with the same chosen power to eliminate unintentional variability, and thus ensures an identical distribution condition; the trials are separated with adequate washout periods to ensure independence between units; and the power of optical stimulation and the length of the experiment are limited to avoid physical damage to the neural circuits.

In addition to Assumption 1, we impose the following three assumptions throughout the paper. First, we generalize the exclusion restriction assumption of Angrist et al. (1996).

Assumption 2 (*Exclusion restriction*). We have that $Y_{iz'\tau} = Y_{iz\tau}$ for $z, z' = 0, 1$ and all i.

Assumption 2 means that the instrumental variable affects the outcome only through the treatment. It holds in optogenetic experiments since only the targeted neurons respond to

optical stimulation. Under Assumption 2, we can simplify $Y_{iz\tau}$ to $Y_{i\tau}$. There are two ways of describing the potential outcome processes under Assumption 2, namely Y_{iz} and $Y_{i\tau}$. We will use Y_{i1} and Y_{i0} to represent the potential processes if the instrumental variable were set to $Z_i = 1$ and $Z_i = 0$, respectively, and use $Y_{i\tau}$ to represent the potential process if the event time of N_i were set to $\mathcal{T}_i = \tau$.

Second, the following independence assumption holds automatically because trials are randomly selected for optical stimulation.

Assumption 3 (*Randomization*). We have that $Z_i \perp\!\!\!\perp \{N_{iz}(\cdot), Y_{i\tau}(\cdot) : z \in \{0, 1\}, \tau \in [0, T] \cup T^+\}$.

Assumption 3 implies that $Z_i \perp\!\!\!\perp \{\mathcal{T}_{iz}, Y_{iz}(t) : z \in \{0, 1\}, t \in [0, T]\}$ under Assumption 2. It allows for the identification of the intention-to-treat effects of the instrumental variable on the treatment and the outcome. However, it is insufficient for identifying the effect of the treatment on the outcome, owing to the possibility of unmeasured confounders.

Finally, we make the following no-anticipation assumption because the event time of N_i at a later time cannot affect Y_i at a previous time.

Assumption 4 (*No anticipation*). We have that $Y_{i\tau}(t) = Y_{i\tau'}(t)$ for $\tau, \tau' \geqslant t$ and all i.

Assumption 4 is well known in causal inference with time series data (see, e.g., Bojinov & Shephard, 2019). The use of a nonstrict inequality sign, instead of a strict inequality in Assumption 4, indicates that the effect of N_i on Y_i is not instantaneous. Replacing the nonstrict inequality with a strict inequality would allow $Y_{i\tau}(\tau) \neq Y_{i\tau'}(\tau)$ for $\tau < \tau'$, i.e., the event at time τ has an effect on the outcome at the same time. The nonstrict inequality in Assumption 4 also implies $Y_{iT^+} = Y_{iT}$ because the event at time T does not have an effect on Y_i in $[0, T]$.

Under Assumptions 1–4, the relationships among Z_i, N_i, Y_i and the unmeasured confounder U_i can be illustrated by the causal diagram in Fig. 1(d). The randomized stimulation Z_i affects the treatment N_i, which in turn affects the outcome Y_i. Because the treatment N_i is not randomized, unmeasured confounders U_i may exist between N_i and Y_i.

2.3. *Definitions of causal effects with point process treatment and outcome*

We are now ready to define the causal quantities of interest. First, we define the average causal effect, ACE, of the instrumental variable on the treatment and outcome processes at time t as $\text{ACE}_N(t) = E\{N_{i1}(t) - N_{i0}(t)\}$ and $\text{ACE}_Y(t) = E\{Y_{i1}(t) - Y_{i0}(t)\}$, respectively. Although $\text{ACE}_N(t)$ and $\text{ACE}_Y(t)$ are possible quantities of interest in the experiment, they do not directly answer the question of how the treatment N_i affects the outcome Y_i. Therefore, we define the ACE of the treatment process on the outcome process as

$$\text{ACE}(t; \tau_1, \tau_2) = E\{Y_{i\tau_1}(t) - Y_{i\tau_2}(t)\}, \quad \tau_1, \tau_2 \in [0, T] \cup T^+, \ \tau_1 \geqslant \tau_2, \tag{3}$$

which characterizes how the change in the event time of N_i from τ_2 to τ_1 affects Y_i at time t. A positive $\text{ACE}(t; \tau_1, \tau_2)$ with $\tau_1 \geqslant \tau_2$ implies that a later event in the treatment process increases the expected outcome process at time t. This effect varies over time t and depends on the two event times τ_1 and τ_2. We define the average causal effect rate, ACER, of N_i on Y_i as

$$\text{ACER}(t; \tau) = \lim_{\Delta\tau \to 0+} \frac{\text{ACE}(t; \tau + \Delta\tau, \tau)}{\Delta\tau} = \frac{\partial E\{Y_{i\tau}(t)\}}{\partial \tau}.$$

The ACER measures how fast $E\{Y_{i\tau}(t)\}$ changes given an infinitesimal change in the event time τ. This concept is similar to the infinitesimal shift function defined in Lok (2008). Under Assumption 4, we have

$$\text{ACE}(t;\tau_1,\tau_2) = \begin{cases} \text{ACE}(t;\tau_1,\tau_2), & \tau_2 < \tau_1 < t, \\ \text{ACE}(t;t,\tau_2), & \tau_2 < t \leqslant \tau_1, \\ 0, & t \leqslant \tau_2 \leqslant \tau_1, \end{cases}$$

and thus $\text{ACER}(t;\tau) = 0$ if $t \leqslant \tau$. When the treatment is a single-point process, we have the following relationship between the ACE and ACER of the treatment:

$$\text{ACE}(t;\tau_1,\tau_2) = \int_{\tau_2}^{\tau_1} \text{ACER}(t;\tau)\,\mathrm{d}\tau. \tag{4}$$

Therefore, we can focus on the ACER because it determines the ACE.

Under Assumption 3, the ACEs of the instrumental variable on the treatment and outcome processes can be identified by the observed differences between the stimulated and unstimulated groups,

$$\text{ACE}_N(t) = f(t), \quad f(t) = E\{N_i(t) \mid Z_i = 1\} - E\{N_i(t) \mid Z_i = 0\}, \tag{5}$$

$$\text{ACE}_Y(t) = h(t), \quad h(t) = E\{Y_i(t) \mid Z_i = 1\} - E\{Y_i(t) \mid Z_i = 0\}. \tag{6}$$

However, Assumption 3 is insufficient for identification of the ACE and ACER of the treatment process on the outcome process, because the treatment process is not randomized.

3. Nonparametric identification and estimation

3.1. *Nonparametric identification with a single-point process treatment*

We begin by generalizing the monotonicity assumption of Angrist et al. (1996).

Assumption 5 (*Monotonicity*). For each i, the potential event times of N_i satisfy $\mathcal{T}_{i1} \leqslant \mathcal{T}_{i0}$.

Assumption 5 requires that the potential event time of N_i under stimulation be no later than that without stimulation. Under Assumption 5, the ACE of the instrumental variable on the treatment process at time τ equals the proportion of a subpopulation defined by the joint potential event times of N_i, i.e.,

$$\text{ACE}_N(\tau) = \text{pr}(\mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}), \quad \tau \in [0, T].$$

Units in this subpopulation would have the event time of the treatment process before or equal to τ with stimulation and after τ without stimulation. Thus, these can be viewed as the compliers whose treatment is positively affected by the stimulation. With a point process treatment, the definition of compliers is time-dependent. Similarly, the other three subpopulations, $\mathcal{T}_{i0} \leqslant \tau < \mathcal{T}_{i1}$, $\max(\mathcal{T}_{i1}, \mathcal{T}_{i0}) \leqslant \tau$ and $\tau < \min(\mathcal{T}_{i1}, \mathcal{T}_{i0})$, generalize the defiers, always-takers and never-takers in the binary instrumental variable model, respectively.

We cannot validate Assumption 5 since it depends on unit-level potential outcomes. However, Assumption 5 implies a testable condition that can be checked using the observed data.

PROPOSITION 1. *Under Assumption 3, Assumption 5 implies that for all* $\tau \in [0, T]$,

$$\text{pr}(\mathcal{T}_i > \tau \mid Z_i = 1) \leqslant \text{pr}(\mathcal{T}_i > \tau \mid Z_i = 0).$$

Proposition 1 states the stochastic dominance of the survival function of $\mathcal{T}_i$ under stimulation over that without stimulation. We can assess Assumption 5 by comparing the empirical survival functions of $\mathcal{T}_i$ in the stimulated and unstimulated groups. If the two curves cross, then the testable condition in Proposition 1 is violated, which in turn falsifies Assumption 5. Therefore, our identification results will consider scenarios both with and without Assumption 5.

Angrist et al. (1996) showed that the effect of the instrumental variable on the outcome equals the product of the effect of the instrumental variable on the treatment and the effect of the treatment on the outcome. The following theorem generalizes their result to our setting.

THEOREM 1. *Suppose that* N_i *is a single-point process and Assumptions 1–4 hold. For any* $t \in [0, T]$,

$$\begin{aligned}\text{ACE}_Y(t) = &\int_0^T E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i0} \leqslant \tau < \mathcal{T}_{i1}\} \,\text{pr}(\mathcal{T}_{i0} \leqslant \tau < \mathcal{T}_{i1})\, \mathrm{d}\tau \\ &- \int_0^T E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\} \,\text{pr}(\mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0})\, \mathrm{d}\tau. \end{aligned} \tag{7}$$

If, in addition, Assumption 5 holds, then for any $t \in [0, T]$ *we have*

$$\text{ACE}_Y(t) = -\int_0^T E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\} \,\text{ACE}_N(\tau)\, \mathrm{d}\tau. \tag{8}$$

In Theorem 1, $\partial Y_{i\tau}(t)/\partial\tau$ is a generalized derivative that may consist of Dirac delta functions (Lax, 2002, Appendix B). Since the conditional set $\mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}$ depends on τ, it is important to be aware that $E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\} = \partial E\{Y_{i\tau'}(t) \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\}/\partial\tau'|_{\tau'=\tau}$, which is generally not equal to $\partial E\{Y_{i\tau}(t) \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\}/\partial\tau$ that takes into account the change of the conditional set.

By rewriting $\text{ACER}(t;\tau)$ as $E\{\partial Y_{i\tau}(t)/\partial\tau\}$, we can view $E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\}$ as the ACER in the subpopulation $\mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}$. In a sense, $E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\}$ generalizes the complier average causal effect in the binary instrumental variable model. Similarly, $E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i0} \leqslant \tau < \mathcal{T}_{i1}\}$ generalizes the average causal effect for the defiers. These conditional expectations may not equal the ACER because $Y_{i\tau}(t)$ and $(\mathcal{T}_{i1}, \mathcal{T}_{i0})$ may not be independent, because of the unmeasured confounding between N_i and Y_i.

The formula in (7) shows that the average causal effect of the instrumental variable on the outcome process, $\text{ACE}_Y(t)$, equals the difference between the weighted averages of the two subpopulation ACERs over the timeline. The weights rely on the joint distribution of $(\mathcal{T}_{i1}, \mathcal{T}_{i0})$. When Assumption 5 holds, the first term on the right-hand side of (7) vanishes and the weight $\text{pr}(\mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0})$ is equal to $\text{ACE}_N(\tau)$. As a result, (7) reduces to (8) under monotonicity.

Under Assumption 3, $\text{ACE}_Y(t)$ and $\text{ACE}_N(t)$ are identifiable. Thus, we can view (7) and (8) as integral equations for the subgroup ACERs (Newey & Powell, 2003). Unfortunately, these subpopulation ACERs are not identifiable without additional assumptions. To provide some intuition, consider (8) under monotonicity. Based on the observed data, (5) and (6) give the identification formulas for $\text{ACE}_Y(t)$ and $\text{ACE}_N(\tau)$ for all $t, \tau \in [0, T]$ under Assumption 3.

So (8) is an integral equation for the unknown quantity defined as $\gamma(\tau, t) = E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\}$. Consider a discrete approximation of $\gamma(\tau, t)$ by evaluating its values over a $K_1 \times K_2$ two-dimensional grid of (τ, t). Equation (8) generates only K_2 equations by considering the K_2 grid of t, which cannot provide identification of the $K_1 \times K_2$ unknown values of $\gamma(\cdot, \cdot)$. Consequently, identification of the ACERs is infeasible without additional assumptions.

To address this problem, we make the following identification assumption.

Assumption 6 (*Stationarity*). We have that $ace(t; \tau_1, \tau_2) = ace(t-\tau_2; \tau_1-\tau_2, 0)$ for $\tau_1 \geqslant \tau_2$ and all t.

Assumption 6 states that the ACE of the treatment on the outcome is invariant under timeline shifts. The left-hand side is the effect of the treatment when the event time is τ_1 versus τ_2 on the outcome at time t. In contrast, the right-hand side represents the same effect, but with the timeline shifted forward by τ_2. Therefore, Assumption 6 means that the ACE of the treatment is invariant regardless of the absolute time. Under Assumption 6, we have $\text{ACER}(t; \tau) = \text{ACER}(t - \tau; 0)$, and thus can simplify $\text{ACER}(t; \tau)$ to $\text{ACER}(t - \tau)$ with $\text{ACER}(t - \tau) = 0$ if $t \leqslant \tau$. We show in the Supplementary material that, if $E\{Y_{iT^+}(t)\}$ is constant for all t, then, Assumption 6 leads to

$$\text{ACER}(t; 0) = -\partial E\{Y_{i0}(t)\}/\partial t. \tag{9}$$

The formula in (9) offers a more natural interpretation of the ACER, that is, $-\text{ACER}(t; 0)$ describes the expected rate of change in the potential outcome at time t when the event in N_i happens at time 0. The next theorem gives sufficient conditions for identifying the ACER.

THEOREM 2. *Suppose that N_i is a single-point process and that Assumptions 1–4 and 6 hold. Furthermore, suppose that either of the following conditions holds:*

(i) *Assumption 5 holds and for all $t, \tau \in [0, T]$,*

$$E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\} = \partial E\{Y_{i\tau}(t)\}/\partial\tau; \tag{10}$$

(ii) *for all $t, \tau \in [0, T]$,*

$$E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}\} = E\{\partial Y_{i\tau}(t)/\partial\tau \mid \mathcal{T}_{i0} \leqslant \tau < \mathcal{T}_{i1}\} = \partial E\{Y_{i\tau}(t)\}/\partial\tau. \tag{11}$$

Then the ACER *satisfies*

$$\text{ACER}(t; \tau) = \begin{cases} \text{ACER}(t - \tau; 0), & t > \tau, \\ 0, & t \leqslant \tau \end{cases} \tag{12}$$

and

$$h(t) = -\int_0^T \text{ACER}(t - \tau; 0) f(\tau)\, \mathrm{d}\tau \tag{13}$$

for $t \in [0, T]$. If, further, $(\Psi f)(\nu) \neq 0$ for all $\nu \in \mathbb{R}$, then the ACER *is identified by $\text{ACER}(t; \tau) = -\Psi^{-1}(G)(t - \tau)$ for $t > \tau$ and $\text{ACER}(t; \tau) = 0$ for $t \leqslant \tau$, where*

$$G(\nu) = \frac{(\Psi h)(\nu)}{(\Psi f)(\nu)}, \quad \nu \in \mathbb{R}. \tag{14}$$

The condition in (10) means that the ACERs are homogenous across subpopulations defined by $\mathcal{T}_{i1} \leqslant \tau < \mathcal{T}_{i0}$ with different values of τ, generalizing the homogeneity assumption in (2). Without monotonicity, the condition in (11) further requires that the ACERs be homogenous across subpopulations defined by $\mathcal{T}_{i0} \leqslant \tau < \mathcal{T}_{i1}$. As with the instrumental variable methods in survival analysis (e.g., Li et al., 2015; Tchetgen Tchetgen et al., 2015), these conditions are satisfied as long as the impact of the confounders on the outcome is additive. With a binary treatment and a scalar outcome, Wang & Tchetgen Tchetgen (2018) imposed a similar condition, assuming no additive interaction between the treatment and the unmeasured confounders on the outcome. We will study this condition in detail under several commonly used outcome models in § 4.

The deconvolution problem (13) belongs to the family of Wiener–Hopf equations; see Noble (1959) and other references. It is essentially the same as the well-studied deconvolution of densities in statistics (e.g., Fan, 1991; Diggle & Hall, 1993; Pensky & Vidakovic, 1999; Johannes, 2009; Dattner et al., 2011, 2016). From the Paley–Wiener–Schwartz theorem, we know that $(\Psi f)(\nu) \neq 0$ for all $\nu \in \mathbb{R}$ if $f(t) = E\{N_i(t) \mid Z_i = 1\} - E\{N_i(t) \mid Z_i = 0\}$ is a nonzero function with bounded support. This is true as long as the effect of the instrumental variable Z_i on the treatment process N_i vanishes in finite time. The nonzero condition on $(\Psi f)(\nu)$ is also employed in the nonparametric deconvolution problem; see, for example, Fan (1991).

In the binary instrumental variable model with the homogeneity assumption, the effect of the treatment on the outcome equals the ratio of the effects of the instrumental variable on the treatment and the outcome. Theorem 2 shows that this ratio relationship also holds for point process treatment and outcome, but in the frequency domain. The well-known convolution theorem tells us that the Fourier transform of a convolution of two functions is equal to the product of their Fourier transforms. Therefore, by applying the Fourier transform to each term of the convolution equation in (13), we can obtain the generalized Wald estimation formula (14) in Theorem 2.

3.2. *Treatment with multiple events*

We generalize the identification result in § 3.1 to a treatment process with possibly multiple events. We begin by generalizing the definition of potential values and causal effects. Let $Y_{i,n(\cdot)}(\cdot)$ be the potential process of the outcome if the treatment were set to a fixed process $n(\cdot)$. The observed outcome process is $Y_i(\cdot) = Y_{i,n(\cdot)}(\cdot)$ if $N_i(\cdot) = n(\cdot)$. Then we can define the ACE of the treatment $n(\cdot)$ versus $n'(\cdot)$ on the outcome as

$$\text{ACE}\{t; n(\cdot), n'(\cdot)\} = E\{Y_{i,n(\cdot)}(t) - Y_{i,n'(\cdot)}(t)\}. \tag{15}$$

For single-point process treatments, (15) reduces to the definition in (3). Using the linearity of expectation, we can write

$$\begin{aligned}\text{ACE}\{t; n(\cdot), n'(\cdot)\} &= E\{Y_{i,n(\cdot)}(t) - Y_{i,n'(\cdot)}(t)\} \\ &= E\{Y_{i,n(\cdot)}(t) - Y_{iT+}(t)\} - E\{Y_{i,n'(\cdot)}(t) - Y_{iT+}(t)\},\end{aligned}$$

where $E\{Y_{i,n(\cdot)}(t) - Y_{iT+}(t)\}$, and $E\{Y_{i,n'(\cdot)}(t) - Y_{iT+}(t)\}$ are the effects of $n(\cdot)$ and $n'(\cdot)$ versus a null process with no events in $[0, T]$, respectively. As in § 3.1, we can characterize the treatment process using event times. Suppose that $n(\cdot)$ has l events at times $\tau_1, \ldots, \tau_l$.

Then $Y_{i,n(\cdot)}(t)$ can be written as $Y_{i,\tau_1,\ldots,\tau_l}(t)$, so its expectation decomposes as

$$E\{Y_{i,\tau_1,\ldots,\tau_l}(t)\} = E\{Y_{iT+}(t)\} + \sum_{s=1}^{l} E\{Y_{i,\tau_1,\ldots,\tau_s}(t) - Y_{i,\tau_1,\ldots,\tau_{s-1}}(t)\} \tag{16}$$

with $Y_{i,\tau_1,\ldots,\tau_{s-1}}(t) = Y_{iT+}(t)$ for $s = 1$. The following assumption simplifies the decomposition by assuming away the interactive effects of the event times in the potential outcome process.

Assumption 7 (*Additivity*). We have that $E\{Y_{i,\tau_1,\ldots,\tau_s}(\cdot) - Y_{i,\tau_1,\ldots,\tau_{s-1}}(\cdot)\} = E\{Y_{i,\tau_s}(\cdot) - Y_{iT+}(\cdot)\}$ for any $s \geqslant 1$ and any event times $(\tau_1, \ldots, \tau_s)$ satisfying $\tau_1 < \tau_2 < \cdots < \tau_s$.

Point processes with event times $(\tau_1, \ldots, \tau_s)$ and $(\tau_1, \ldots, \tau_{s-1})$ have the same trajectory up to time τ_{s-1}, where the former has an additional event at τ_s. Assumption 7 means that the effect of the process with event times $(\tau_1, \ldots, \tau_s)$ versus that with event times $(\tau_1, \ldots, \tau_{s-1})$ does not depend on their common trajectory up to time τ_{s-1}. Hence, the causal effect remains the same when the first $s-1$ events are removed from both processes. Under Assumption 7, (16) simplifies to $E\{Y_{i,\tau_1,\ldots,\tau_l}(t) - Y_{iT+}(t)\} = \sum_{s=1}^{l} E\{Y_{i,\tau_s}(t) - Y_{iT+}(t)\}$, which means that the effect of each event time on the outcome process is additive. In the Supplementary Material we show that Assumption 7 holds under the Hawkes process (Hawkes, 1971) and under Aalen's additive hazard model (Aalen, 1980) for the potential outcome process. Assumption 7 may be violated because of the interactive effect of the event times in the treatment process. For instance, neural ensembles are famous for their plasticity in the long term, the ability of neural connections to reorganize themselves in response to stimulation, which clearly violates Assumption 7. Such violations of Assumption 7 are sometimes of scientific interest. We leave the investigation of their effects for future research.

Under Assumption 7, we can separately study the effect of each event in N_i. The following proposition generalizes (4) to treatment processes with multiple events.

PROPOSITION 2. *Under Assumptions* 1, 2, 4 *and* 7, *we have*

$$\text{ACE}\{t; n(\cdot), n'(\cdot)\} = -\int_0^t \text{ACER}(t;\tau)\{n(\tau) - n'(\tau)\}\,\mathrm{d}\tau.$$

Based on Proposition 2, we can focus on the identification of the ACER. The next theorem generalizes Theorem 2 to treatment processes with multiple events.

THEOREM 3. *Suppose that Assumptions* 1–4, 6 *and* 7 *hold. If for all* $t \in [0, T]$ *and any fixed processes* $n(\cdot)$ *and* $n'(\cdot)$,

$$E\{Y_{i,n(\cdot)}(t) - Y_{i,n'(\cdot)}(t) \mid N_{i1}(\cdot) = n(\cdot), N_{i0}(\cdot) = n'(\cdot)\} = E\{Y_{i,n(\cdot)}(t) - Y_{i,n'(\cdot)}(t)\}, \tag{17}$$

then the ACER *satisfies* (12) *and* (13). *If* $(\Psi f)(v) \neq 0$ *for all* $v \in \mathbb{R}$, *then the* ACER *is identified by* $\text{ACER}(t;\tau) = -\Psi^{-1}(G)(t-\tau)$ *for* $t > \tau$ *and* $\text{ACER}(t;\tau) = 0$ *for* $t \leqslant \tau$, *where* $G(v)$ *is as defined in* (14).

When N_i is a single-point process, Theorem 3 does not require Assumption 7, and the condition (17) reduces to (11) in Theorem 2. As a result, Theorem 3 reduces to Theorem 2 when N_i has at most one event. Similar to Theorem 2, the condition in (17) means that the ACERs are homogenous across subpopulations defined by N_{i1} and N_{i0}.

3.3. *Estimation*

We consider estimation of the ACER based on identification results from Theorems 2 and 3. This is essentially the deconvolution problem commonly studied in the literature; for more discussion see, for instance, Diggle & Hall (1993), Pensky & Vidakovic (1999), Johannes (2009) and Dattner et al. (2011, 2016). Since deriving an optimal estimation procedure is not the focus of this paper, we only provide a simple regression-based procedure to estimate the ACER. To be specific, we use a two-step procedure by first obtaining the estimates of f and h, and then solving for the ACER from the empirical version of the convolution equation in (13).

Let $\hat{f}$ and $\hat{h}$ denote the estimators of f and h defined in (5) and (6), which equal the empirical mean differences in the treatment and outcome processes in the stimulated and unstimulated groups. We approximate the true ACER with truncated basis expansions: for $\Delta \in [0, T]$,

$$\text{ACER}(\Delta; 0) \approx \sum_{j=1}^{J} \psi_j(\Delta)\beta_j,$$

where J is a tuning parameter for the number of bases and $\{\psi_j(\cdot) : j = 1, 2, \ldots, J\}$ is a set of prespecified basis functions. Here the support of $\text{ACER}(\cdot\,; 0)$ can be determined by prior knowledge. We then estimate $\beta = (\beta_1, \ldots, \beta_J)$ by minimizing the following penalized ℓ_2-distance based on the convolution equation (13):

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^J} \left\| \hat{h} + \sum_{j=1}^{J} (\psi_j * \hat{f})\beta_j \right\|_2^2 + \eta \|\beta\|_2^2, \tag{18}$$

where $*$ denotes the convolution between two functions and we have introduced the ridge penalty to reduce boundary effects. An analytic solution for $\hat{\beta}$ is available since the objective function in (18) is quadratic in β. Recalling that we consider independent trials, we can choose the values of the parameters J and η using cross-validation or based on prior knowledge such as the smoothness of the ACER. Denoting the selected parameter values by $\hat{J}$ and $\hat{\eta}$, the final estimator is $\widehat{\text{ACER}}(\cdot\,; 0) = \sum_{j=1}^{\hat{J}} \psi_j(\cdot)\hat{\beta}_{j,\hat{\eta}}$. We can then construct the confidence band for the function $\widehat{\text{ACER}}(\cdot\,; 0)$ using the bootstrap. The asymptotic properties for $\widehat{\text{ACER}}$ as the sample size m increases follow from the standard theory assuming independent samples. We leave the rigorous discussion for future analysis, as it is beyond the scope of the present work.

4. THE ROLE OF MODELS: CAUSAL INTERPRETABILITY AND IDENTIFIABILITY

4.1. *Conditional intensity*

In this section, we study several commonly used models for point process outcomes in applied research when an instrumental variable is available, allowing for the presence of unmeasured confounders. We do not impose any distributional assumptions on the unmeasured confounders. Consequently, it is difficult to study the identifiability of the model parameters themselves. We take an alternative route by connecting the model parameters to the causal effects, and considering the identifiability and estimation of the causal effects directly. With a binary instrumental variable, we show that the ACER is identifiable and can be estimated using generalized Wald estimation under many commonly used models. This estimation strategy does not rely on the identification or estimation of the model parameters, as long as the unmeasured confounding is additive in the underlying outcome model.

We begin by introducing some additional notation to characterize a point process. Let $U_i(\cdot)$ denote the unmeasured confounding process on $\mathbb{R}$. We use $\mathcal{H}_{it-}$ to represent the σ-algebra induced by the history up to, but not including, time t. We define the conditional intensity of Y_i as

$$\lambda_Y(t) = E\{\mathrm{d}Y_i(t)/\mathrm{d}t \mid \mathcal{H}_{it-}\}. \tag{19}$$

The conditional intensity, or intensity, is the conditional mean of the event rate of Y_i in an infinitesimal time interval $[t, t + \mathrm{d}t)$, which is analogous to the conditional mean of the outcome in the binary instrumental variable model. It fully characterizes the probabilistic structure of a point process, and is closely related to the hazard function in survival analysis; see Daley & Vere-Jones (2003, Ch. 7) for more discussion of the intensity.

In (19), the conditional intensity could depend on the history of Y_i. When the outcome Y_i describes recurrent events, it is common practice to allow the conditional intensity to depend on past events of Y_i (e.g., Hawkes, 1971; Brillinger, 1988; Lawrence, 2004; Kulkarni & Paninski, 2007; Yu et al., 2009; Gao et al., 2015; Macke et al., 2015; Gao et al., 2016; Wu et al., 2017; Zhao & Park, 2017; Pandarinath et al., 2018). As concrete examples, in the context of neural data, the dependence on past events captures the known phenomena that a single neuron cannot fire consecutively in a very short period of time, and that activities in a region may trigger inhibitory circuits to stabilize the activity on a longer time scale.

An inherent constraint on the intensity is that it must be nonnegative for the probabilistic model to be well-defined. A similar constraint is well acknowledged in modelling the hazard function in survival analysis. This constraint on the intensity creates a schism in the modelling of point processes, between whether to employ a linear working model (Aalen, 1980) or a nonnegative generative model (Cox, 1972). In either model, since U_i is unobserved, existing methods usually impose strong parametric assumptions on U_i in order to estimate the parameters. A common assumption is that U_i is a Gaussian process (see, e.g., Yu et al., 2009; Zhao & Park, 2017), primarily owing to its simplicity for the Bayesian computation. However, the analysis can be sensitive to these parametric assumptions. We study a linear model in § 4.2 and a nonlinear model in § 4.2.

Before specifying $\lambda_Y(t)$, we assume the following conditions on the relationships among N_i, Y_i and U_i, which are commonly used in instrumental variable methods when outcome models are employed; see Tchetgen Tchetgen et al. (2015) for an example in survival analysis.

Assumption 8. The following conditions hold:

(i) $Z_i \perp\!\!\!\perp \{N_{iz}(t), Y_{i\tau}(t), U_i(t) : z \in \{0,1\}, t \in [0,T], \tau \in [0,T] \cup T^+\}$;

(ii) for $t \in [0,T]$, $Y_{i,n(\cdot)}(t) \perp\!\!\!\perp N_i(\cdot) \mid \{\mathcal{H}^{*\setminus N_i}_{it-}, N_i(s) = n(s), s \in [0,t)\}$ and any fixed point process $n(\cdot)$, where $\mathcal{H}^{*\setminus N_i}_{it-}$ denotes the σ-algebra induced by all potential processes including U_i, except for N_i up to time t.

See Lok (2008) for the measure-theoretic description of the independence given histories of point processes. Assumption 8(i) is a restatement of Assumption 3 with the addition of U_i. It holds because Z_i is randomized. Assumption 8(ii) generalizes the latent sequential ignorability in Ricciardi et al. (2020) to a continuous-time setting with point process treatment and outcome. It assumes that U_i fully characterizes the confounding between the treatment and the outcome, so that the treatment is independent of the potential outcome at time t given the histories of N_i, Y_i and U_i. Under Assumption 8 we have that for any

$t \in [0, T]$,

$$E\{\mathrm{d}Y_{i,n(\cdot)}(t)/\mathrm{d}t \mid \mathcal{H}_{it-}\} = E\{\mathrm{d}Y_i(t)/\mathrm{d}t \mid N_i(\cdot) = n(\cdot), \mathcal{H}_{it-}^{*\backslash N_i}\}, \tag{20}$$

which links the potential processes to the conditional intensity of the observed outcome. Therefore, the discussion in § 4.2 will focus on the models for the observed outcome. Under each of the models, we will connect the model parameters with the ACER and study its identification.

4.2. *Identification of causal effects with linear additive unmeasured confounding*

We start with linear models for the intensity. This type of model has been widely used in different contexts because of its mathematical tractability (e.g., Hawkes, 1971; Aalen, 1980; Tchetgen Tchetgen et al., 2015; Jiang et al., 2018). In particular, consider a linear Hawkes process $Y_i(\cdot)$ with intensity

$$\lambda_Y(t) = \mu_Y + \int_0^t g(t-s)\,\mathrm{d}N_i(s) + \int_0^t \omega(t-s)\,\mathrm{d}Y_i(s) + \psi_{U_i}(t), \tag{21}$$

where $g(\Delta) = \omega(\Delta) = 0$ for $\Delta \leqslant 0$ and $\psi_{U_i}(t)$ represents any function of $\{U_i(s) : s \in [0, t)\}$. The next proposition connects the ACER with the parameters in (21) and shows the identification.

PROPOSITION 3. *Suppose that Assumptions 1–3 and 8 hold and that the underlying outcome model satisfies (21). Then the following hold:*

(i) *we have that*

$$\mathrm{ACER}(t;\tau) = \mathrm{ACER}(t-\tau;0) = -(\Psi^{-1}\tilde{G})(t-\tau),$$

where $\tilde{G}(v) = \{1 + (\Psi\omega)(v)\}^{-1}(\Psi g)(v)$ *if* $1 + (\Psi\omega)(v) \neq 0$ *for all* $v \in \mathbb{R}$;

(ii) *when* $(\Psi f)(v) \neq 0$ *for all* $v \in \mathbb{R}$, *the* ACER *is identified by* $\mathrm{ACER}(t;\tau) = -(\Psi^{-1}G)(t-\tau)$ *for* $t > \tau$ *and* $\mathrm{ACER}(t;\tau) = 0$ *for* $t \leqslant \tau$, *with* $G(v)$ *as defined in* (14).

In practice, the function $g(\cdot)$ is often interpreted as the effect of an event in N_i on the outcome Y_i conditional on the history up to time t. Proposition 3(i) expresses the ACER in terms of the model parameters, showing that $g(\cdot)$ and $\omega(\cdot)$ jointly characterize the ACER of N_i on Y_i. When the dependence on past Y_i does not exist, i.e., $\omega(\cdot) \equiv 0$, we have $\mathrm{ACER}(t;\tau) = -g(t-\tau)$. From Proposition 3(i), we can obtain the ACER if we can estimate the model parameters in (21). However, this requires specifying the distribution of $U_i(\cdot)$. Fortunately, Proposition 3(ii) shows that, when a binary instrumental variable is available, we can identify the ACER without any distributional assumption on U_i and hence estimate it using the method in § 3.3. This broadens the applicability of (21) with fewer parametric assumptions.

Proposition 3(ii) is an application of Theorem 2 under (21). The linearity in (21) plays a key role in the causal interpretation of the model parameters and nonparametric identification of the ACER. The linear terms of N_i and Y_i connect the ACER with $g(\cdot)$ and $\omega(\cdot)$, and the linear term of U_i implies Assumption 6 and the condition in (11).

4.3. *Identification of causal effects with nonlinear additive unmeasured confounding*

We now consider the following nonlinear model, which is similar to models in survival analysis with an instrumental variable (e.g., MacKenzie et al., 2014; Li et al., 2015; Tchetgen Tchetgen et al., 2015):

$$\lambda_Y(t) = \phi\left\{\mu_Y + \int_0^t g(t-s)\,\mathrm{d}N_i(s)\right\} + \psi_{U_i}(t). \tag{22}$$

Model (22) generalizes model (21) by allowing for a nonlinear relationship between N_i and Y_i through the link function ϕ while requiring the unmeasured confounding effect to be additive. For (22), the following proposition characterizes the causal effect and its identifiability.

PROPOSITION 4. *Suppose that Assumptions 1–3 and 8 hold, N_i is a single-point process, and the underlying outcome model satisfies (22). Then the following hold:*

(i) *for $t, \tau \in [0, T]$,*

$$\text{ACER}(t;\tau) = \phi(\mu_Y) - \phi\{\mu_Y + g(t-\tau)\};$$

(ii) *when $(\Psi f)(\nu) \neq 0$ for all $\nu \in \mathbb{R}$, the ACER is identified by $\text{ACER}(t;\tau) = -(\Psi^{-1}G)(t-\tau)$ for $t > \tau$ and $\text{ACER}(t;\tau) = 0$ for $t \leqslant \tau$, where $G(\nu)$ is as defined in (14).*

From Proposition 4(i), the causal interpretation of $g(\cdot)$ depends on μ_Y and the link function $\phi(\cdot)$ under (22). As a result, even with the same link function, the interpretation of $g(\cdot)$ differs in populations with different values of μ_Y. This tells us to beware of interpreting $g(\cdot)$ as some causal effects. Proposition 4(ii) is an application of Theorem 2. As with (21), we can use the method in § 3.3 to estimate the ACER without knowledge of $\phi(\cdot)$ or U_i. Therefore, Proposition 4 suggests directly targeting the ACER instead of the model parameter $g(\cdot)$. This circumvents the daunting task of identifying, estimating and interpreting the model parameters in (22), broadening its applicability with fewer parametric assumptions.

Critically, although (22) allows for nonlinearity, it restricts the effect of the unmeasured confounder to be additive. Relaxing this modelling assumption is challenging. In the Supplementary Material we show that, when the confounding effect on Y_i is nonadditive, the ACER would depend on the distribution of the confounder, making the identification impossible without a distributional assumption on U_i.

5. NUMERICAL ANALYSIS

5.1. *Simulation*

We use simulation to illustrate the numerical performance of the proposed nonparametric estimation procedure. In this simulation study, we generate the treatment N_i and outcome Y_i from the model

$$\lambda_N(t) = \mu_N + \phi_{\beta_0}\{\alpha(t; a_N, b_N)Z_i + U_i(t)\}, \tag{23}$$

$$\lambda_Y(t) = \phi_{\beta_2}\left[\phi_{\beta_1}\left\{\mu_Y + \int_0^t \alpha(\Delta; a_Y, b_Y)\,\mathrm{d}N_i(t-\Delta)\right\} + \phi_{\beta_1}\{U_i(t-d_U)\}\right], \tag{24}$$

where $\alpha(\cdot\,; a, b) = ba^2 t\exp(-at)$ is the alpha function (see, e.g., Ermentrout & Terman, 2010, Ch. 7) and $\phi_\beta(x) = x^\beta$ is the link function. The confounding variable U_i is generated

as a Gaussian process with mean zero and a squared exponential kernel $\text{cov}\{U_i(t), U_i(t+d)\} = \sigma_U^2 \exp\{-d^2/(2l_U^2)\}$. The parameters in (23) and (24) are set to $\mu_N = \mu_Y = 0.2$, $a_N = 10, b_N = 0.5$, $a_Y = 8, b_Y = 1$, $d_U = 0.5$, $\sigma_U = 0.2$ and $l_U = 0.1$. We consider five scenarios in this simulation.

Scenario 1a ($\beta_0 = \beta_1 = \beta_2 = 1$): a linear model for Y_i with a single-point process N_i, which is achieved by suppressing the intensity (23) to zero after the first event in N_i is generated.

Scenario 1b ($\beta_1 = 2$, $\beta_0 = \beta_2 = 1$): an additive confounding model for Y_i with a single-point process N_i.

Scenario 2a ($\beta_0 = \beta_1 = \beta_2 = 1$): a linear model for Y_i with multiple events in N_i.

Scenario 2b ($\beta_0 = 3$, $\beta_1 = 2$, $\beta_2 = 1$): a linear model for Y_i with multiple events in N_i and nonadditive confounding effects on N_i.

Scenario 3 ($\beta_0 = \beta_1 = 1$, $\beta_2 = 3$): a nonadditive confounding model for Y_i.

For each scenario, we generate m trials with m ranging from 40 to 800. In each simulated dataset, half of the trials are set to have $Z_i = 1$ and the other half $Z_i = 0$. The processes N_i and Y_i are generated from 0 to $T = 3$ using a thinning process, while the unmeasured confounding process U_i is generated from -1 to 3 to account for its delayed effect on Y_i.

For Scenario 3, identification of the ACER is difficult with nonlinear confounding effects, as illustrated in the Supplementary Material. Therefore we use the Monte Carlo method to calculate the ACER to show its dependence on the distribution of the unmeasured confounder. For Scenarios 1a to 2b, we apply the proposed generalized Wald estimation procedure in §3.3. We estimate the function $h(t)$ as the difference between the empirical cumulative intensities of Y_i in the treatment group, $Z_i = 1$, and the control group, $Z_i = 0$. The function $f(t)$ is estimated in a similar manner. We approximate the ACER using a cubic B-spline with six knots evenly spaced in $[0, 1]$, where the mass of $\alpha(\cdot\,; a_Y, b_Y)$ resides. The tuning parameter of the ridge penalty η is set to m^{-1} to reduce boundary effects from the nonparametric approximation. To measure the performance, we calculate the proportion of integrated squared errors with respect to the true ACER, that is,

$$r = \frac{\int_0^1 \{\widehat{\text{ACER}}(\Delta) - \text{ACER}(\Delta; 0)\}^2 \, \text{d}\Delta}{\int_0^1 \text{ACER}^2(\Delta; 0) \, \text{d}\Delta}, \tag{25}$$

where the true ACER is calculated from (24) based on Propositions 3 and 4.

Figure 2 shows the simulation results averaged over 1000 replicates. Panels (a) and (b) show that the performances of the estimators improve as the number of trials increases in Scenarios 1a, 1b, 2a and 2b. In particular, the proportion of integrated squared error in Scenario 1a is larger than in Scenario 2a, despite their using the same model for Y_i. This reveals a feature of point process treatments, namely that given the same number of trials, more events in N_i contribute more information for recovering the causal effects of N_i on Y_i. The four curves in Fig. 2(a) and (b) converge slowly towards zero due to the existence of approximation error in the basis expansion and the bias from penalization. Panels (c) and (d) of Fig. 2 show the calculated true ACER in Scenario 3 under two different distributions of U_i; in each panel the five curves correspond to τ being $0, 0.25, 0.5, 0.75$ and 1. Even with the same t, the shape of $\text{ACER}(t; \tau)$ varies as τ changes in both panels, implying that $\text{ACER}(t; \tau)$ does not equal $\text{ACER}(t - \tau)$. Moreover, the contrast between Figs. 2(c)

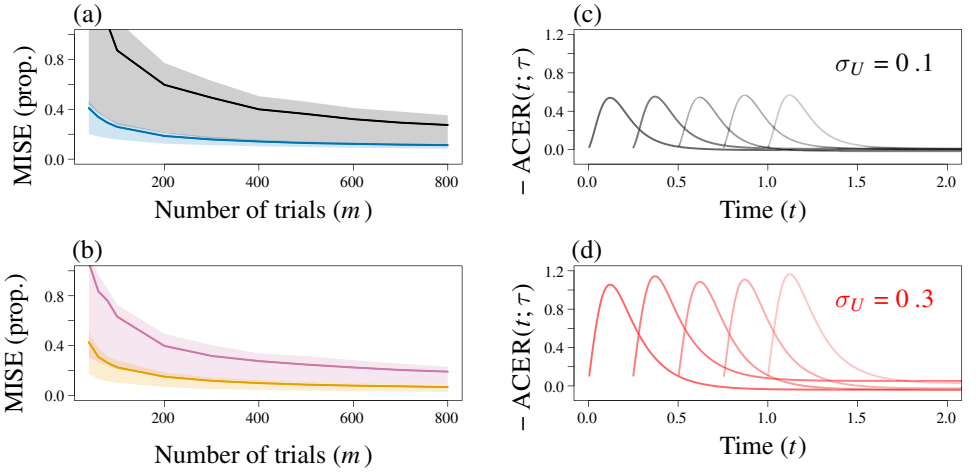

Fig. 2. Identification of the ACER and performance of the generalized Wald estimation averaged across 1000 replicates. Panel (a) shows the performance of the proposed estimation procedures in Scenarios 1a (black) and 1b (blue), and panel (b) shows the performance of the proposed estimation procedures in Scenarios 2a (purple) and 2b (orange). The horizontal axes represent the number of trials, m, and the vertical axes the measure r defined in (25). The shaded areas are the interquartile bands from the 1000 replicates. Estimation performances in the four scenarios are not directly comparable, given the huge differences between their data-generating mechanisms. Based on the interpretation in (9), panels (c) and (d) show the true value of $-\text{ACER}(t;\tau)$ with nonadditive confounding effects on Y, in Scenario 3. The curves in (c) are calculated with $\sigma_U = 0.1$, and those in (d) with $\sigma_U = 0.3$. In each panel the five curves correspond to τ being $0, 0.25, 0.5, 0.75$ and 1, respectively, and are not shift-invariant.

and (d) shows that the ACER depends on the distribution of the unmeasured process U_i, demonstrating the sensitivity of the ACER to distributional assumptions on U_i.

5.2. *Empirical analysis*

We apply the proposed method to the neural data from Bolding & Franks (2018a). First, we provide some basic background to aid understanding of the experiment; for more details, see Bolding & Franks (2018a) and the Supplementary Material. Bolding & Franks (2018a) conducted an experiment to understand how a mouse brain maintains stationarity in odour detection regardless of odour concentration. Specifically, it is known that neural activities in the olfactory bulb, OB, increase in response to a higher concentration of odour particles, and that a spike in the OB triggers neural activities of principal neurons, PN, in the piriform cortex, where the odour is perceived by the brain. To avoid other neural processes that normalize odour responses, Bolding & Franks (2018a) used optogenetics to stimulate neurons in the OB with one-second light pulses, which meet the requirements of an instrumental variable in our framework. Bolding & Franks (2018a) also took an optogenetic approach to circumventing the contribution of centrifugal inputs and other intrabulbar processes, which effectively cuts of the feedback from the PN to the OB. Figure 3(d) shows a causal diagram of the relationship between the stimulation, OB, and the PN.

The dataset contains spike trains recorded in the OB and PN during the experiment. A total of 160 trials were conducted on eight mice, with each mouse having 10 trials without stimulation, $Z_i = 0$, and 10 trials with a one-second light pulse at 20 mW mm^{-2}, $Z_i = 1$. The light pulse, if present, has onset time 0 and ends at 1 s. In our analysis, we consider the first 3.5 seconds of a trial, from -0.5 to 3, as there are hardly any residual effects afterwards. We regard the treatment N_i as the process of events in the OB and the outcome Y_i as the process of events in the PN. Each recorded event in N_i is a spike of one neuron in the OB that may

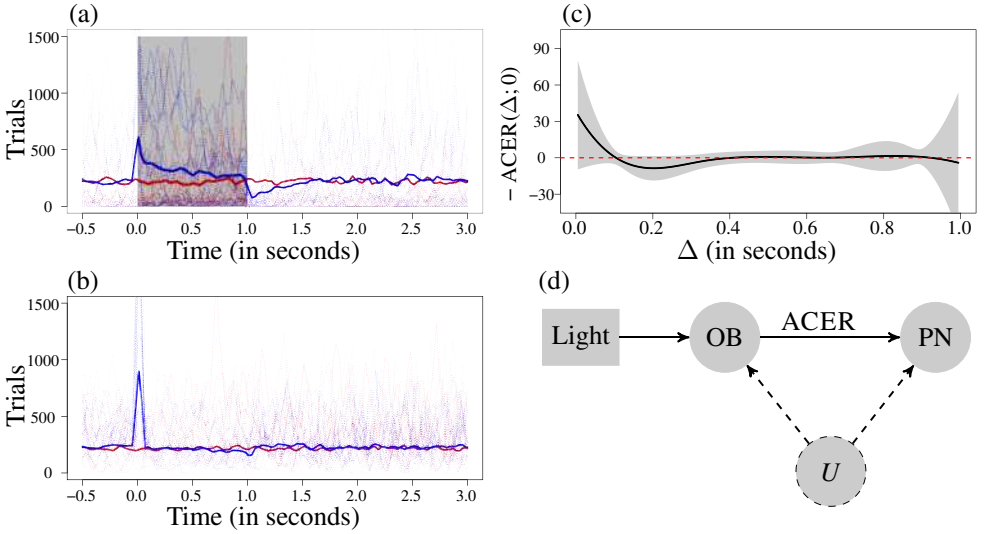

Fig. 3. Empirical intensities and the fitted ACER based on data from Bolding & Franks (2018b). Panels (a) and (b) show the empirical intensities of the neural activities of (a) the OB and (b) the PN in the stimulated (blue) and unstimulated (red) groups. The solid curves represent the average intensity over 80 trials, and the dashed curves the empirical intensities from 20 randomly selected trials in each group. The shaded area in panel (a) represents the duration of the light pulse. Based on the interpretation in (9), panel (c) shows the estimated $-\text{ACER}(\Delta; 0)$ from the full dataset; the shaded area represents a 90% confidence band for visualizing the uncertainty of the estimates from 5000 bootstrap samples. Panel (d) shows the causal diagram of the relationship among the variables.

trigger a distinct group of PN in the piriform cortex. Given the number of PN in the piriform cortex, the triggered groups may have few or no overlaps, limiting the interactive effect of the treatment process. Therefore, the additivity in Assumption 7 is a plausible approximation to the true underlying mechanism. Figure 3(a) and (b) show the smoothed intensities of neural activities in the stimulated and unstimulated groups. We can see that the stimulation triggers increased activities in the OB in all trials. However, large variations are present in the neural activities across trials.

We first conduct a preliminary analysis assuming no unmeasured confounders between the treatment and outcome processes. In this case, identification of the causal effects does not require the instrumental variable. We fit a model of Y_i on N_i and directly interpret the coefficient function of N_i as the causal effect. However, the conclusion is inconsistent with the findings in Bolding & Franks (2018a), implying possible unmeasured confounders or model misspecification. For more details see the Supplementary Material.

We then apply the generalized Wald estimation procedure to estimate the causal effects of neural activities in the OB on the neural activities of PN. From the observed data, we estimate the functions h and f using differences between empirical cumulative intensities in the stimulated, $Z_i = 1$, and unstimulated, $Z_i = 0$, groups. We approximate the unknown ACER using cubic B-splines with evenly spaced knots in $[0, 1]$, where two knots are selected by five-fold cross-validation. We set the tuning parameter for the ridge penalty to $\eta = 0.01$ to handle boundary effects. A 90% confidence band is constructed using the bootstrap; for details see the Supplementary Material. We use the bootstrap confidence band to approximate the uncertainty in the estimates.

Figure 3(c) displays the estimated ACER. The curve shows that an event in the OB elicits high activity in the PN immediately after the event, i.e., within 0.1 seconds, but the effect quickly turns negative for an extended duration, from 0.1 to 0.4 seconds, before dying down.

This is consistent with the findings of Bolding & Franks (2018a) that a temporal mechanism is in place to stabilize the neural activities of PN after the initial detection of odours. Additional analysis of the neural dataset can be found in the Supplementary Material. The confidence band shows that the proposed generalized Wald estimation procedure yields high uncertainty near the boundaries, despite a large number of events and inclusion of the ridge penalty. In this particular case, it appears that the ACER vanishes after 0.5 seconds, but the boundary effect causes spurious estimates in the bootstrap samples. In practice, we recommend that practitioners use the generalized Wald estimation procedure, and then apply a suitable parametric form or shape constraint to the ACER.

Acknowledgement

We thank two reviewers for helpful comments. Chen and Ding were partially supported by the U.S. National Science Foundation. Jiang and Chen contributed equally to this paper.

Supplementary material

The Supplementary Material includes proofs and further details of the empirical analysis.

References

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical Statistics and Probability Theory*, W. Klonecki, A. Kozek & J. Rosiński, eds. Berlin: Springer, pp. 1–25.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–55.

Bojinov, I. & Shephard, N. (2019). Time series experiments and causal estimands: Exact randomization tests and trading. *J. Am. Statist. Assoc.* **114**, 1665–82.

Bolding, K. A. & Franks, K. M. (2018a). Recurrent cortical circuits implement concentration-invariant odor coding. *Science* **361**, eaat6904.

Bolding, K. A. & Franks, K. M. (2018b). Simultaneous extracellular recordings from mice olfactory bulb (OB) and piriform cortex (PCx) and respiration data in response to odor stimuli and optogenetic stimulation of OB. Available at http://crcns.org/data-sets/pcx/pcx-1.

Brillinger, D. R. (1988). Maximum likelihood analysis of spike trains of interacting nerve cells. *Biol. Cybernet.* **59**, 189–200.

Carrillo-Reid, L., Han, S., Yang, W., Akrouh, A. & Yuste, R. (2019). Controlling visually guided behavior by holographic recalling of cortical ensembles. *Cell* **178**, 447–57.

Chen, H., Geng, Z. & Zhou, X.-H. (2009). Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data (rejoinder). *Biometrics* **65**, 689–91.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc.* B **34**, 187–202.

Cox, D. R. & Isham, V. (1980). *Point Processes*. Boca Raton, Florida: CRC Press.

Daley, D. J. & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*. New York: Springer.

Dattner, I., Goldenshluger, A. & Juditsky, A. (2011). On deconvolution of distribution functions. *Ann. Statist.* **39**, 2477–501.

Dattner, I., Rei, M. & Trabs, M. (2016). Adaptive quantile estimation in deconvolution with unknown error distribution. *Bernoulli* **22**, 143–92.

Diggle, P. J. & Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *J. R. Statist. Soc.* B **55**, 523–31.

Ermentrout, G. B. & Terman, D. H. (2010). *Mathematical Foundations of Neuroscience*. New York: Springer.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**, 1257–72.

Gao, Y., Archer, E. W., Paninski, L. & Cunningham, J. P. (2016). Linear dynamical neural population models through nonlinear embeddings. In *Proc. 30th Int. Conf. Neural Information Processing Systems*. Red Hook, New York: Curran Associates, pp. 163–71.

Gao, Y., Busing, L., Shenoy, K. V. & Cunningham, J. P. (2015). High-dimensional neural spike train analysis with generalized count linear dynamical systems. In *Proc. 28th Int. Conf. Neural Information Processing Systems*. Cambridge, Massachusetts: MIT Press, pp. 2044–52.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58**, 83–90.
Heckman, J. J. (1996). Identification of causal effects using instrumental variables: Comment. *J. Am. Statist. Assoc.* **91**, 459–62.
Jiang, B., Li, J. & Fine, J. (2018). On two-step residual inclusion estimator for instrument variable additive hazards model. *Biostatist. Epidemiol.* **2**, 47–60.
Johannes, J. (2009). Deconvolution with unknown error distribution. *Ann. Statist.* **37**, 2301–23.
Kulkarni, J. E. & Paninski, L. (2007). Common-input models for multiple neural spike-train data. *Network: Comp. Neural Syst.* **18**, 375–407.
Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In *Proc. 16th Int. Conf. Neural Information Processing Systems*. Cambridge, Massachusetts: MIT Press, pp. 329–36.
Lax, P. D. (2002). *Functional Analysis*. New York: Wiley.
Li, J., Fine, J. & Brookhart, A. (2015). Instrumental variable additive hazards models. *Biometrics* **71**, 122–30.
Lok, J. (2008). Statistical modeling of causal effects in continuous time. *Ann. Statist.* **36**, 1464–507.
Macke, J. H., Buesing, L. & Sahani, M. (2015). Estimating state and parameters in state space models of spike trains. In *Advanced State Space Methods for Neural and Clinical Data*, Z. Chen, ed. Cambridge: Cambridge University Press, pp. 137–59.
MacKenzie, T. A., Tosteson, T. D., Morden, N. E., Stukel, T. A. & O'Malley, A. J. (2014). Using instrumental variables to estimate a Cox's proportional hazards regression subject to additive confounding. *Health Serv. Outcomes Res. Methodol.* **14**, 54–68.
Mardinly, A. R., Oldenburg, I. A., Pégard, N. C., Sridharan, S., Lyall, E. H., Chesnov, K., Brohawn, S. G., Waller, L. & Adesnik, H. (2018). Precise multimodal optical control of neural ensemble activity. *Nature Neurosci.* **21**, 881–93.
Martinussen, T., Vansteelandt, S., Tchetgen Tchetgen, E. J. & Zucker, D. M. (2017). Instrumental variables estimation of exposure effects on a time-to-event endpoint using structural cumulative survival models. *Biometrics* **73**, 1140–9.
Newey, W. K. & Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–78.
Noble, B. (1959). *Methods Based on the Wiener-Hopf Technique for the Solution of Partial Differential Equations*. New York: Pergamon Press.
Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., Trautmann, E. M., Kaufman, M. T., Ryu, S. I., Hochberg, L. R. et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Meth.* **15**, 805–15.
Pensky, M. & Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.* **27**, 2033–53.
Ricciardi, F., Mattei, A. & Mealli, F. (2020). Bayesian inference for sequential treatments under latent sequential ignorability. *J. Am. Statist. Assoc.* **115**, 1498–517.
Richardson, A., Hudgens, M. G., Fine, J. P. & Brookhart, M. A. (2017). Nonparametric binary instrumental variable analysis of competing risks data. *Biostatistics* **18**, 48–61.
Ridder, G. & Moffitt, R. (2007). The econometrics of data combination. In *Handbook of Econometrics*, vol. 6. Amsterdam: Elsevier, pp. 5469–547.
Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test Comment. *J. Am. Statist. Assoc.* **75**, 591–3.
Tchetgen Tchetgen, E. J., Walter, S., Vansteelandt, S., Martinussen, T. & Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology* **26**, 402–10.
Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.* **11**, 284–300.
Wang, L. & Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *J. R. Statist. Soc.* B **80**, 531–50.
Wu, A., Roy, N. A., Keeley, S. & Pillow, J. W. (2017). Gaussian process based nonlinear latent structure discovery in multivariate spike train data. In *Proc. 31st Int. Conf. Neural Information Processing Systems*. Red Hook, New York: Curran Associates, pp. 3496–505.
Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V. & Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Proc. 21st Int. Conf. Neural Information Processing Systems*. Red Hook, New York: Curran Associates, pp. 1881–8.
Zhao, Y. & Park, I. M. (2017). Variational latent Gaussian process for recovering single-trial dynamics from population spike trains. *Neural Comp.* **29**, 1293–316.

[*Received on* 25 *November* 2021. *Editorial decision on* 5 *January* 2023]