The Cost-Sharing, Shadow Price and Cluster in Medical Care Utilization: A Self-Exciting Perspective *

Yuhao LI
yuli@eco.uc3m.es

Departamento de Economia, Universidad Carlos III de Madrid

Abstract

In this paper, a self-exciting counting process modelling method is proposed to study the frequency of medical care service utilization under a non-linear budget constraint health insurance policy. This modelling strategy enables researchers to investigate individual's dynamic behavior in a more detailed way. Specifically, for each individual, every doctor visiting record is represented as a point in a self-exciting counting process. Cost associated with such visiting is included in this counting process as a mark. A minimum distance method is employed to find the estimators. Using the Rand Health Insurance Experiment data, we find that individuals respond to a change of shadow price. In addition, we use a matured cluster analysis algorithm to investigate the cluster patterns and discover that compared to free plan, cost-sharing insurance plan with out-of-pocket fees suppress the use of medical services by limiting the number of clusters as well as follow-up visiting within each cluster.

JEL. C41, C13, C51, I13, I12

Keywords. Health insurance, non-linear budget constraint, medical service utilization, self-exciting process, history-dependent dynamic, minimum distance estimation

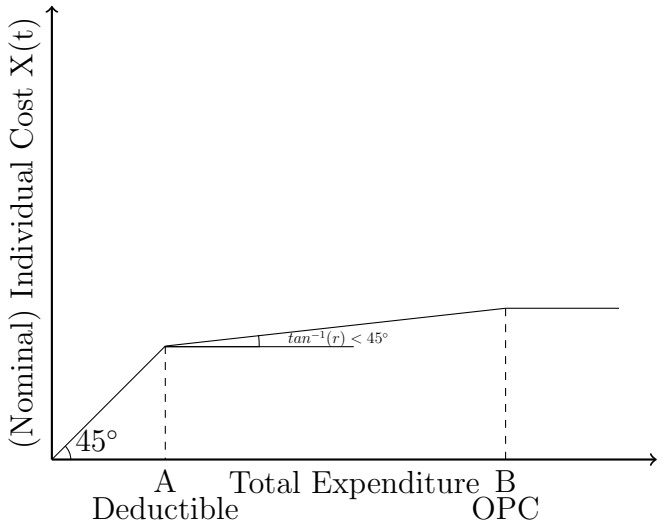
1 Introduction

In this paper, we aim to model medical utilization (outpatient only) under a non-linear cost environment using a self-exciting process. Recent studies on health

*I am grateful to Prof.Miguel.A.Delgado for supports and guidance throughout this project and to Prof.Winfried Stute for his inspiration and valuable comments. I would also like to thank conference and seminar participants at the EEA-ESEM Lison, the IAAE Montreal, the UC3M Econometrics workshop.

insurance (e.g., Aron-Dine et al. (2012),Einav et al. (2015)) deviate from the classical assumption that individuals only respond to a single linear spot cost[1] and find strong evidence that individuals respond to the dynamic incentives associated with the non-linear nature of a typical health insurance contract. These conclusions suggest that 'it is unlikely that a single elasticity estimate can summarize the spending response to changes in health insurance' and 'such an estimate is not conceptually well defined.' (Aron-Dine et al., 2012).

The driving forces of such a non-linear nature are cost-sharing policies implemented in a health insurance contract. The most common ones are the deductible, the co-insurance rate and the out-of-pocket fee cap (OPC). In a typical setup, individuals need to cover all their medical expenditures below the deductible. Once the threshold is passed, co-insurance is applied, where individuals pay part of the expenditures based on the co-insurance rate. Finally, if the total expenditure paid by the individual passes the OPC, no cost (or very little cost) would be paid by this individual. Figure 1 illustrates such a typical non-linear budget constraint.



The total expenditure is the sum of individual costs and costs paid by the insurance. Points A and B are the deductible threshold and OPC, respectively. When the total expenditure is below A, the co-insurance is 100% (individuals pay all cost) and the slope is 1. Between A and B, a co-insurance rate (the slope) $0 < r < 1$ is applied. Whenever the total expenditure is beyond B, there is no cost for individuals (the slope is 0).

Figure 1: Non-linear Individual Cost (Medical Price)

At the heart of this non-linearity is the stochastic cumulative individual cost $X(t)$[2]. Keeler, Newhouse, and Phelps (1977) is the first theoretical paper that studies the consumer's optimal choice under such a non-linear medical price schedule. Using a dynamic programming model, they show that the shadow price of j^{th} episode is

[1]That is, a linear budget constraint.

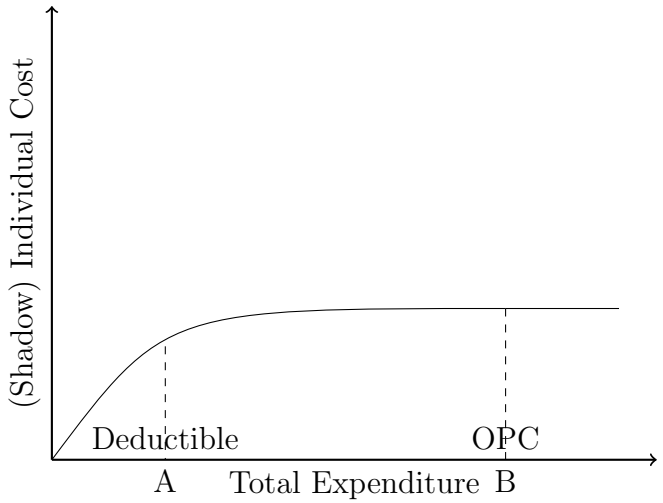
[2]Or equivalently, the total expenditure. Because there is a clear insurance regulation on individual's out-of-pocket cost, cumulative individual cost and total expenditure are one-to-one mapped.

a function of demand prior to this episode (hence the cumulative individual cost). One may construct the shadow price (co-insurance rate) as:

$$p^s(t) = 1 - V(X(t))$$

where $0 \leq V(X(t)) \leq 1$ is a bonus that is related to the cumulative individual cost with $V^{'} > 0$. The intuition behind this equation is simple: under the range of deductibles, although individuals need to fully bear the medical cost, each time this person consumes, the remaining deductible is reduced and the next instance consumption is more easily to exceed the deductible. As a result, the shadow price for the next purchase is cheaper than the price of the current one (hence the name 'bonus'). Moreover, as the cumulative individual cost gets closer to the deductible, individuals have greater incentive to consume[3]. That is, there should be a positive (negative) relationship between cumulative individual cost $X(t)$ (remaining deductibles) and the probability of medical utilization. For the purpose of self-contain, we review this theory in detail in Appendix A.

The shadow price theory has profound implications on estimating medical demand. First, it suggests one should not use the nominal price. Since the difference between the nominal price and shadow price is not randomly generated, an incorrectly chosen nominal price would lead to a biased estimation. Second, because the shadow co-insurance rate is a function of cumulative individual cost, it implies that individuals will make medical service utilization decisions in a sequential and contingent way. Figure 2 illustrates the situation.



Points A and B are the deductible threshold and OPC, respectively. When the total expenditure is below B, the co-insurance rate (the slope) $0 < r(X) < 1$ is a function of cumulative individual cost with $r^{'} < 0$. Whenever the total expenditure is beyond B, there is no cost for individuals.

Figure 2: Non-linear Individual Shadow Cost (Medical Price)

Different sources of stochastic disturbances should be distinguished in order to properly model $X(t)$. The first source of randomness is that at any given time $\bar{t}$

[3]We assume that medical service is a normal good

within the insurance year, $X(\bar{t})$ is a random variable satisfying $X(\bar{t}) \geq X(s), \forall s \leq \bar{t}$. This non-decreasing random process is difficult to model directly. However, notice that $X(t)$ is a piece-wise constant step function. We may then decompose $X(t)$ as 1) the occurrence time of i^{th} illness episode t_i (the position of i^{th} jump in this step function) and 2) conditional on the occurrence of i^{th} illness, the individual cost $x(t_i)$ for such illness (the size of i^{th} jump). Thus, we could represent the cumulative individual cost as a compound counting process: $X(t) = \sum_{i=1}^{\infty} x(t_i)\mathbb{I}\{t_i \leq t\}$. This structure suggests that we could model the time t_i and the cost $x(t_i)$ separately. Medical costs are convenient to assume to be i.i.d and their distribution is well approximated by a log-normal. This is known among literatures (Handel et al., 2015; Keeler and Rolph, 1988). Thus the key to model $X(t)$ is to model its occurrence times $\{t_i\}_{i \in \mathcal{N}^+}$.

The second source of randomness comes from other contributions to the individual cost. For example, in this paper, we mainly focus on the outpatient medical utilization. But the costs associated with doctor visits are not the only source of individual cost; other sources could be inpatient expenditures and drug purchase costs. These random costs serve as external shocks to our interested outpatient costs.

The primary goal of this paper is to model $X(t)$, especially the occurrence times of illness episodes. Since the stochastic cumulative individual cost can be represented as a compound counting process, we advocate to use the counting process as our analysis workhorse. In particular, a marked self-exciting process, whose filtration is generated by the counting process itself, is employed to contain the historical information. This information includes both the occurrence time and the non-covered individual charge of each illness episode.

We also aim to model the dependence structure among episodes and to study the effects of cost-sharing policies on episode cluster structures. Episodes are often in a form of a cluster, this is known among medical literatures. A typical example is chronic diseases where patients need to receive treatment periodically.

As will be shown later, the self-exciting process can 1) fully account for different sources of stochastic disturbances of cumulative individual cost $X(t)$, 2) capture the essence of the shadow price: the negative relationship between shadow price and medical utilization (or equivalently, the positive relationship between $X(t)$ and medical utilization) and 3) model the dependence structure among episodes. In comparison, conventional methods such as count data regression and duration models are inadequate to deal with the randomness of individual cost $X(t)$ and the episode cluster structure. As both methods assume events to be i.i.d, which excludes the non-linear price system that we aim to address.

We use the Rand Health Insurance Experiment dataset. Besides it is widely used in the health insurance literature, one advantage of this dataset is that it includes a detailed episode-level claim-by-claim data. We can then update $X(t)$ whenever an event or external shock occurs.

We use a minimum distance method to obtain the estimators. This method is first introduced by Kopperschmidt and Stute (2013) and has the advantage to incorporate external shocks.

This paper contributes three strands of literatures. First, we enrich the ever-

expanding literatures that aim to study individual response to a non-linear budget constraint. Second, we introduce a new econometric tool that can be applied beyond health insurance studies. Potential applications include but not limited to labour economics (studies of multiple unemployment, work absences), industry organization (sequential entry games) and criminology etc. Last, we provide a simulation study to exam the performance of this new minimum distance estimation method.

The paper is constructed as follow. Section 2 provids a brief literature review on non-linear budget constraint problem in health insurance. Section 3 introduces some notations and basic concepts about the self-exciting process. Section 4 discusses the minimum distance estimation method. In addition, a simulation study is performed to study the performance of this method. Section 5 introduces the dataset and provides a simple test for state dependence. Section 6 presents our model, in which the stochastic property of cumulative individual cost, the effect of cost-sharing policy and the dependence structure of episodes are fully considered. Section 7 presents estimation results of the model. We also use a mature machine learning algorithm to analyse the cluster structure of episodes in this section. In section 8 we discuss the advantages of our modeling strategy over other conventional reduced form or structural form methods. Section 9 concludes the paper.

2 Literature Review

We briefly review some literatures that try to include the non-linear budget constraint in their models.

Aron-Dine et al. (2012) construct a future price $p^f = 1 - Pr(X(T) \geq \bar{X})$ in order to reject the null hypothesis that individuals only respond to a single spot price. Here $X(T)$ is the cumulative individual cost on the last day of an insurance contract year, and $\bar{X}$ is the deductible. They find a negative relationship between the future price and the initial medical use. Notice that in their construction of future price, only $X(T)$ is used, the rest of $X(t), \forall t < T$ is ignored. In principle, one could construct future price as a function of time using the same method: $p^f(t) = 1 - Pr(X(t) \geq \bar{X})$. But in practice this would lead to a complicated procedure as one needs to use simulated future price to instrument the future price to correct the estimation bias, see Aron-Dine et al. (2012) for details.

Brot-Goldberg et al. (2017) define their shadow expected marginal end-of-year price at month m as a conditional expectation: $p^e_m = \mathbb{E}(r_{EOY}|X(m), Z, H)$ where r_{EOY} is the end-of-year co-insurance rate, Z is a vector of covariates and H is a measurement of health stock. They non-parametrically estimate the probability density function on cells of equivalent consumers using triple $(X(m), Z, H)$. In practice, they only use age as their sole explanatory variable. Their results suggest that shadow price have a limited impact on spending reduction.

Einav et al. (2015) construct and estimate a dynamic economic model to study individual's drug purchase behavior. In each period, the cumulative individual cost is updated by: $X(t) = X(t-1) + x(t)$, where $x(t)$ is the aggregate individual cost in the current period. Thus $X(t)$ here is not 'totally' stochastic: the occurrence time of illness is ignored and $x(t) = \sum_{i:t_i \in \text{current period}} x(t_i)$ is an aggregate random variable.

Moreover, one may find difficulties to model the shadow price in a structural model, since the shadow budget constraint (as illustrated in figure 2) is actually unobserved by researchers.

To the best of our knowledge, no literature has ever explicitly taken the sources of stochastic disturbances mentioned before into consideration. In addition, no literature has ever measured the individuals' responds to the shadow price on a episode level. The main contribution of this paper is to fill this gap from a self-exciting process perspective.

3 Some Notations and Basic Concepts about Self-Exciting Process

Recall that our interested individual cost is a multiplicative form of occurrence times and non-covered charges: $X(t) = \sum_{i=1}^{\infty} x(t_i)\mathbb{I}\{t_i \leq t\}$. In the study of counting process, it is conventionally called the compound process. But this process itself is difficult to analysis. Luckily, one can actually identify it with the marked counting process (see Karr (1991)). One advantage to do so is that the intensity of a marked counting process can be separated as a ground intensity and a conditional mark density. This separation allows us to concentrate on the occurrence times of events. For the moment, we postpone the discussion of this separation.

We begin with an un-marked counting process:

$$N(t) = \sum_{i=1}^{\infty} \mathbb{I}\{t_i \leq t\} \tag{1}$$

where $t_i, i \in \mathcal{N}^+$ are occurrence times of realized events, $\mathbb{I}\{\cdot\}$ is the indicator function. An example of such a counting process is illustrated in figure 3

Next, we may include marks in the counting process by extending the definition:

$$N(t, R^+) = \sum_{i=1}^{\infty} \mathbb{I}\{t_i \leq t, x_i \in R^+\} \tag{2}$$

where x_i is the mark associated with i^{th} event. In our application, an event is a medical utilization (e.g., a visit to a doctor) and a mark is the associated expenditure $x(t_i)$.

As mentioned earlier, compound counting process can be identified with its corresponding marked counting process. This is because, using the language of random measure, the compound counting process is purely atomic on R^+:

$$X = \textstyle\sum_{i=1}^{K} U_i \delta_{T_i}$$

where U_i are non-negative random variables, T_i are measurable mappings from $(\Omega, \mathcal{F})$ into $(R^+, \mathcal{B}^+)$ and δ is the point mass at t. While the marked counting process, represented as a random measure is:

$$N = \textstyle\sum_{i=1}^{K} \delta_{(T_i, U_i)}$$

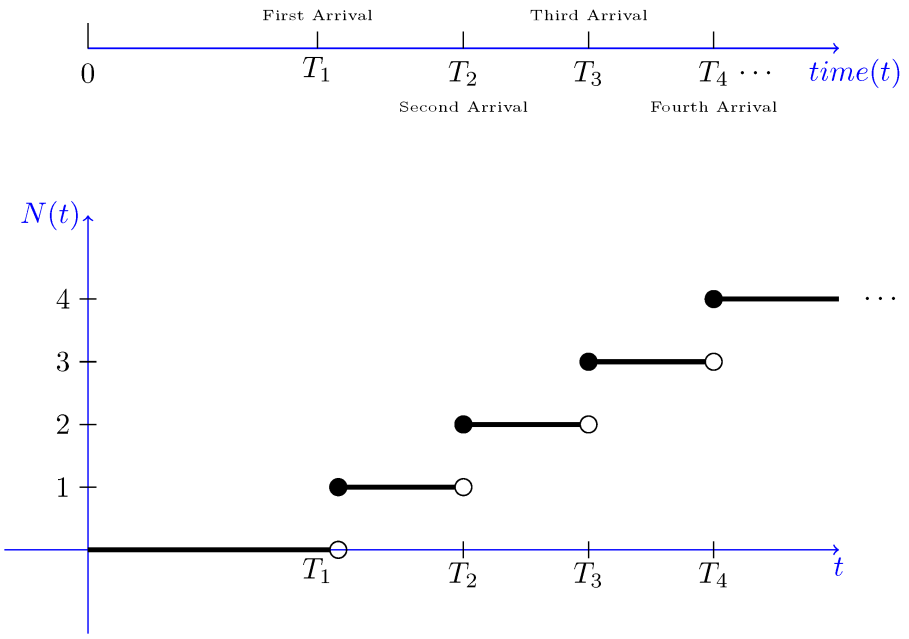


Figure 3: A possible realization of a counting process

That is, we may view a purely atomic random measure on R^+ as the marked counting process on $R^+ \times R^{+4}$. Intuitively, compound counting process can be identified as marked counting process because each of these two processes contains information sufficient to reconstruct the other.

3.1 Intensity

The intensity λ of a counting process is a measure of the rate of change of its predictable part. Conditional on a time dependent filtration $\mathcal{F}_{t-}$, the intensity is defined as:

$$\lambda(t|\mathcal{F}_{t-}) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(N([t, t+\Delta t])|\mathcal{F}_{t-})}{\Delta t} \tag{3}$$

In a similar way, we define the intensity for a marked counting process as:

$$\lambda(t, x|\mathcal{F}_{t-}) = \lim_{|\Delta t \times \Delta x| \to 0} \frac{\mathbb{E}(N([t, t+\Delta t], [x, x+\Delta x])|\mathcal{F}_{t-})}{\Delta t \Delta x} \tag{4}$$

By construction, the counting process $N(t)$ is a sub-martingale, therefore the intensity is non-negative.

The cumulative intensity, as its name suggests, is defined as the integral of an intensity over a period of time, say $[0, t]$:

$$\Lambda(t|\mathcal{F}_{t-}) = \int_0^t \lambda(s|\mathcal{F}_{s-})ds \tag{5}$$

The self-exciting process is characterized by its filtration: if the filtration is generated by the counting process itself: $\mathcal{F}_t = \sigma(N(s) : s \leq t)$, we call this counting

[4]More technical details can be found on (Karr, 1991)

process a self-exciting process. We dedicate the next subsection to discuss the issue of filtration. Here, to streamline the illustration, we suppress reference $\mathcal{F}_{t-}$ in λ and Λ .

The cumulative intensity and the counting process are connected by the well-known Doob-Meyer decomposition theorem:

$$N(t) = \Lambda(t) + M(t) \tag{6}$$

This theorem states that any counting process can be decomposed as the sum of a cumulative intensity (also known as compensator in some literature) and a martingale $M(t)$. Moreover, the cumulative intensity is predictable and unique. By the martingale property we have,

$$\mathbb{E}(N(t)) = \mathbb{E}\Lambda(t) \tag{7}$$

We can interpret the cumulative intensity as the mean of the underlying counting process at $t, \forall t$. From an economic perspective, we may say that the cumulative intensity summarizes all the systematic parts of a counting process, while the martingale accounts for the stochastic part. Notice that the cumulative intensity $\Lambda(t)$ may not be absolutely continuous. One common assumption regarding this (e.g.,Kopperschmidt and Stute (2013)) is to let the process $t \to \Lambda(t)$ be almost surely continuous, while allows unexpected jumps in the intensity function $\lambda(t)$.

The intensity also connects to the probability density of the underlying counting process. Let $U_{n+1} = T_{n+1} - T_n$ be the duration between n^{th} and $n + 1^{th}$ arrivals, for each arrival n, let $F_n(du) = Pr\{U_{n+1} \in du\}$ then

$$\Lambda(t) = \Lambda(T_n) + \int_0^{t-T_n} \frac{F_n(dx)}{F_n[x, \infty)}, t \in (T_n, T_{n+1}]$$

where T_i is stopping time [5]. The proof can be found in Karr (1991).

There are some similarities between the intensity and the hazard rate, which is the workhorse in duration models. However, it is important to note that these two concepts are fundamentally different. Consider a system that begins in time 0 and fails at some random time $T > 0$, that is, the length of the spell is T. The hazard rate (or hazard function) $h(t)$ is defined as:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{Pr\{T \in (t, t + \Delta t)\}}{Pr\{T > t\}\Delta t} \\ &= \frac{f_T(t)}{1 - F_T(t)} \end{aligned} \tag{8}$$

This hazard rate tells us the conditional probability of the system failing in the interval $(t, t + \Delta t]$ conditional on the system working at time t.

In general, intensity deals with reoccurring arrivals with focus on the timing per se, while the hazard rate deals with the duration or the length of one spell. If the first arrival occurs at time T, for the intensity, the analysis of the next arrival is based on $\lambda(t), \forall t > T$, whereas for duration analysis there are no new arrivals, and the failed system is immediately replaced by a new one that begins at time T. The hazard rate for this new system is $h(t - T)$.

[5] Appendix B contains an introduction to the concept of stopping time.

3.2 Filtration and Marks

As mentioned before, in the case of self-exciting, the filtration is generated by the marked counting process itself $\mathcal{F}_t = \sigma(N(s, x_s) : s \leq t, x_s \in R^+)$. Thus, the history, such as the timings of occurrences and their marks, is contained in this filtration. However, the order of timings and marks are different. This difference is crucial to properly understand the filtration $\mathcal{F}_{t-}$, usually referred as the strict history. Intuitively, we will never know the values of marks without the occurrences of events. This suggests that marks are adjuncts to the occurrences. Formally, we can show that:

$$\mathcal{F}_{T_n-} = \sigma((T_1, X_1), (T_2, X_2), \cdots, (T_{n-1}, X_{n-1}), T_n) \tag{9}$$

where T_i are again the stopping times and X_i are the associated marks. It shows that the accumulating information changes only at the arrival times T_i, whereas $T_i \in \mathcal{F}_{T_i-}$, X_i does not belong to this filtration. We provide details of the concept of filtration, strict history and a proof of equation 9 in Appendix B.

In many economic studies it is of interest to introduce some external information. In the example of health insurance, we may investigate the impact of income or education on the usage of medical services. We can, in fact, enrich this filtration to include these covariates. Let $\mathcal{H}_{t-} = \mathcal{H}_0 \vee \mathcal{F}_{t-}$ be the conditioned filtration, where $\mathcal{H}_0$ is the $\sigma - algebra$ generated by some external covariates, such as age, sex, race, income, etc. We interpret this filtration as the 'whole history'. Notice that $\mathcal{H}_0$ can also be time-dependent, i.e., $\mathcal{H}_0 = \mathcal{H}_0(t-)$.

The fact that marks are adjuncts to occurrence times inspires us to separate timings and marks. Intuitively we may re-write the intensity for the marked counting process as:

$$\begin{aligned}
\lambda(t, x|\mathcal{H}_{t-}) &= \lim_{|\Delta t \Delta x| \to 0} \frac{Pr\Big(t < T \leq t + \Delta t, x < \mathbf{X} \leq x + \Delta x | \mathcal{H}_{t-}\Big)}{\Delta t \Delta x} \\
&= \lim_{\Delta t \to 0} \frac{Pr\Big(t < T \leq t + \Delta t | \mathcal{H}_{t-}\Big)}{\Delta t} f(x|\mathcal{H}_{t-}, t) \\
&= \lambda_g(t|\mathcal{H}_{t-}) f(x|t, \mathcal{H}_{t-})
\end{aligned} \tag{10}$$

where we call $\lambda_g(t|\mathcal{H}_{t-})$ the ground intensity, and it happens to be the intensity of the original marked counting process if marks are ignored, or the ground counting process:

$$N_g(t) = \sum_i \mathbb{I}\{T_i \leq t\} \tag{11}$$

The second part $f(x|t, \mathcal{H}_{t-})$ is called the conditioned mark density since it is not only conditioned on the filtration, but also conditioned on the occurrence time t. In appendix C, we use the Janossy measurement language to discuss this separation in a more rigorous way.

3.3 Heterogeneity in a Self-Exciting Process

In a variety of contexts, it is often noticed that individuals who have experienced an event in the past are more likely to experience the event again in the future than are individuals who have not experienced the event (Heckman, 1981). One explanation, best known as the unobserved heterogeneity, is that in addition to the observed variables, there are other relevant variables that are unobserved but correlated with the observed ones. Unobserved heterogeneity (UH) is an important issue in health insurance literature. This is because prices are endogenous: they are lower on average for those who tend to have more episodes. Sickly individuals tend to consume more care services and hence are likely to exceed their OPCs. Therefore, it is crucial to separate the sickliness effect from the shadow price effect. In models like count data regression and duration analysis, the most common way to characterize UH is through the random effect: integrate out the UH term to obtain a marginal distribution.

This strategy faces two problems: 1) A lack of economic theory supporting the choice of the unobserved heterogeneity distribution G. In most duration analysis literature, G is chosen to be Gamma, it is more mathematical convenience than anything else, since by doing so, one can have a closed form of the marginal distribution. 2) It is well known among literatures that there is severe damage of misspecification of G. Van den Berg (2001) provides a theoretical example, and Heckman and Singer (1984) use real application data and various G distributions (Normal, Log Normal and Gamma) to demonstrate that the estimation results tend to be unstable.

Another explanation of heterogeneity is related to state dependence (SD). This concept says that past experience has a genuine effect on future events in a sense that an otherwise identical individual who did not experience the event would behave differently in the future. The definition of the self-exciting process naturally includes the idea of state dependence. Thus in this study, we advocate to assume the state dependence as the source of individual heterogeneity. Specific to our application, individuals are assumed to have no knowledge about their health status at first but gradually update their awareness as episodes and medical utilization are experienced. Different past experiences will generate different behaviors even if the underlying intensity is the same. In the next section, we will demonstrate this fact by a simulation.

SD shares many common properties with random effect. For example, in the mixed proportional hazard model, it is well known that unobserved heterogeneity leads to a 'weeding out' or 'sorting' phenomenon (Van den Berg, 2001). That is, individuals with the highest values of UH term leave the default state quickest on average, and the individuals who are still in the state tend to have lower values of UH term. In our medical utilization application, the state of interest is the duration of keeping healthy. A high value of UH term means a bad health condition (or high level of sickness), and on average it means the duration of being health is short. If only the outpatient is interested, this means on average, we would observe more doctor visiting histories among these individuals on a defined time period. Thus, simply by counting the number of past events, we can capture this 'weeding out' phenomenon.

The SD term can be flexibly modelled. In this study, we suggest the form

$$\sum_{i:t_i<t} K(t_i, t) \tag{12}$$

and include it in the intensity function.

There are three advantages to do so. First, the SD term is time dependent, which means that it can be updated. Compared to the conventional time-invariant unobserved heterogeneity term, we believe such a modelling method is more realistic in our empirical study.

Second, the choice of $K(t_i, t)$ is flexible and can be consistent with economic theory. For example, we may capture the seasonality effect by setting $K(t_i, t) = \alpha sin(\beta(t - t_i) + \gamma) + \delta$, or in our application, as explained later, we may study the cluster phenomenon of medical care utilization by letting $K(t_i, t) = \mu exp(-\mu(t - t_i)), \mu > 0$.

Lastly, since the SD term has a similar numerical structure to the non-parametric kernel density estimation, it can be used as a smoother and numerically it is possible that by choosing different but similar kernels, the value of two SD terms would be close, i.e.,

$$\sum_{i:t_i<t} K^1(t_i, t) \approx \sum_{i:t_i<t} K^2(t_i, t)$$

such that the rest of the estimators in the intensity model are stable.

4 Estimation and Simulation

Since one can separate the marked intensity into a multiplicative form of ground intensity and conditional mark density, it would be straightforward to estimate these two parts separately. In this study, we mainly focus on the estimation of the ground intensity, later in this section, we will briefly discuss the estimation of conditional mark density.

In the counting process literature, likelihood based methods are the most commonly used estimation tools, (e.g.,Ogata and Katsura (1988),Zhuang et al. (2002),Aït-Sahalia et al. (2015),Bacry and Muzy (2014) and Mohler et al. (2012)). One requirement of using them is the predictability of the cumulative intensity Λ with respect to the filtration $\sigma(N_g(s) : s \leq t)$. That is, conditional on the filtration, the values of all the explanatory variables at time t should be known and observed just before t. However, as pointed out by Kopperschmidt and Stute (2013), in many complicated economic situations, there is little reason to maintain such an assumption. Instead, the cumulative intensity should respect external shocks or impulses. In that case, the model is most likely not dominated and the likelihood methods are difficult to apply.

In our application, a core task is to update the cumulative individual cost whenever an event occurs. Two sources of cost are considered, the first one comes from the main counting process $N^1(t)$ in which an event is a doctor visit and a mark is the associated individual cost. Another one is the drug purchase cost, represented by a mark linked to a drug purchase counting process $N^2(t)$. As a result, the individual cost coming from the drug purchase serves as an external shock to the main counting process.

More precisely, the conditional filtration $\mathcal{H}_{t-}$ in our model is generated not only by the main counting process, but also by the external drug purchase counting process, i.e., $\mathcal{H}_{t-} = \mathcal{H}_0 \vee \mathcal{F}_{t-} \vee \mathcal{G}_{t-}$, where $\mathcal{F}_t = \sigma(N^1(s) : s \leq t), \mathcal{G}_{t-} = \sigma(N^2(s) : s \leq t)$.

Figure 4 helps to understand.

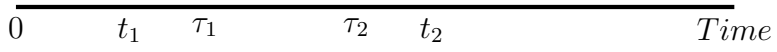


Figure 4: A possible realization of illness episodes and drug purchases

Here t_i are occurrence times of illness episodes and τ_i are drug purchase times. The interested intensity λ is not predictable with respect to the filtration $\mathcal{F}$ generated only by N^1 since the cumulative individual cost is updated due to drug purchase events. But λ is predictable with respect to $\mathcal{H}$.

4.1 Minimum Distance Estimation

To overcome this problem,Kopperschmidt and Stute (2013) develop a parametric minimum distance estimation method. The basic idea consists of using the Doob-Meyer decomposition to minimize the distance between the counting process and its cumulative intensity. This method only requires the observations to be i.i.d. It does not assume the differentiability of the cumulative intensity and allows unexpected jumps in the intensity function.

Formally, let $v_0 \in \Theta \subset \mathbb{R}^d$ be the true parameters, and let $N_{g,1}, ..., N_{g,n}$ be i.i.d copies of n observed ground counting process. For each $1 \leq i \leq n$, let $\mathcal{H}_i(t)$ be an increasing filtration comprising the relevant information about the marked counting process N_i as well as some other external information. Let $\Lambda_{g,v,i}$ with $v \in \Theta \subset \mathbb{R}^d$ be a given class of parametric cumulative ground intensities. Let the true one be $\Lambda_{g,i} = \Lambda_{g,v_0,i}$.

Let,

$$\bar{N}_{g,n} = \frac{1}{n}\sum_{i=1}^{n} N_{g,i}; \bar{\Lambda}_{g,v,n} = \frac{1}{n}\sum_{i=1}^{n} \Lambda_{g,v,i} \tag{13}$$

We call the former averaged (ground) point process and the latter averaged cumulative (ground) intensity. Naturally the associated averaged innovation martingale is,

$$d\bar{M}_{g,n} = d\bar{N}_{g,n} - d\bar{\Lambda}_{g,v_0,n} \tag{14}$$

The optimization object is:

$$||\bar{N}_{g,n} - \bar{\Lambda}_{g,v,n}||_{\bar{N}_{g,n}} \tag{15}$$

Where

$$||f||_\mu = [\int_0^T f^2 d\mu]^{1/2}$$

T is a terminating time. This statistic 15 is an overall measurement of fitness of $\bar{\Lambda}_{g,v,n}$ to $\bar{N}_{g,n}$. The estimator v_n is computed as,

$$v_n = arg \inf_{v \in \Theta} ||\bar{N}_{g,n} - \bar{\Lambda}_{g,v,n}||_{\bar{N}_{g,n}} \tag{16}$$

Kopperschmidt and Stute (2013) show this estimator is consistency and its asymptotic behaviour is

$$\sqrt{n}\Phi_0(v_0)(v_n - v_0) \to \mathcal{N}_d(0, C(v_0)) \tag{17}$$

where

$$\Phi_0(v) = \frac{\partial}{\partial v}\int_E (\mathbb{E}\Lambda_{g,v}(t) - \mathbb{E}\Lambda_{g,v_0}(t))\mathbb{E}\frac{\partial}{\partial v}\Lambda_{g,v}(t)^T\mathbb{E}\Lambda_{g,v_0}(dt) \tag{18}$$

$C(v_0)$ is a $d \times d$ matrix with entries

$$C_{ij}(v_0) = \int_E \phi_i(x)\phi_j(x)\mathbb{E}\Lambda_{g,v_0}(dx) \tag{19}$$

and

$$\phi_i(x) = \int_{[x,\bar{t}]} \mathbb{E}\frac{\partial}{\partial v_i}\Lambda_{g,v}(t)\mathbb{E}\Lambda_{g,v_0}(dt) \mid_{v=v_0}, \underline{t} \leq x \leq \bar{t} \tag{20}$$

Remark Let Φ_n be the empirical analog of Φ_0,

$$\Phi_n(v) = \frac{\partial}{\partial v}\int_E (\bar{\Lambda}_{g,v,n}(t) - \bar{\Lambda}_{g,v_0,n}(t))\frac{\partial}{\partial v}\bar{\Lambda}_{g,v,n}(t)^T\bar{\Lambda}_{g,v_0,n}(dt) \tag{21}$$

Since all $\bar{\Lambda}_{g,v,n}$ are sample means of i.i.d non-decreasing processes, a Glivenko-Cantelli argument yields, with probability one, uniform convergence of $\bar{\Lambda}_{g,v,n} \to \mathbb{E}\Lambda_{g,v}(t)$ in each t on compact subsets of Θ, we have the expansion,

$$\Phi_n(v) = \Phi_0(v) + op(1) \tag{22}$$

Such expansion guarantees that in a finite sample situation, we can replace the unknown matrix $\Phi_0(v_0)$ by $\Phi_n(v_n)$ and $C(v_0)$ by $C^n(v_n)$ without destroying the distributional approximation through $\mathcal{N}_d(0, C(v_0))$, where C^n is the sample analog of C. In practice, one needs to plug and replace the true ones with estimators and replace $\mathbb{E}\Lambda_{g,v_0}(dt)$ with its empirical counterpart $\bar{N}_g(dt)$.

As for the estimation of the conditional mark density, if researchers are able to write down the parametric form of the conditioned mark density, a straightforward maximum likelihood estimation is in order. One may also use non-parametric techniques to estimate the conditional density function. However, the whole filtration $\mathcal{H}_{t-}$ may complicate the situation. In practice, we suggest using only part of the filtration information: $f(x|t, \mathcal{H}_{t-}) = f(x|z)$, where z is a vector of covariates and is part of the filtration. In that way, a conventional conditional KDE may apply.

4.2 Simulation Study

In the original Kopperschmidt and Stute (2013) paper, the authors do not provide a numerical simulation study. Here we complete the task by generating a self-exciting process and examining the performance of the minimum distance estimator.

The data generating process we picked is the ETAS (epidemic type aftershock sequence) model. It was first introduced by Ogata and Katsura (1988) and ever

since has been widely used in seismology (e.g. Zhuang et al. (2002)). It characterizes earthquake times and magnitudes and belongs to a marked Hawkes process family. The ETAS model has the probabilistic structure we desire: marks are part of the ground intensity and can be separated into ground intensity and conditioned mark density.

The ground intensity of a ETAS model, for its simplest form, could be:

$$\lambda_g(t|\mathcal{F}_{t-}) = \mu + \sum_{i:t_i<t} e^{\alpha x_i} \left(1 + \frac{t-t_i}{c}\right)^{-1} \tag{23}$$

where x_i is the magnitude of an earthquake occurring at time t_i, and the mark density, for simplicity, is assumed to be independent:

$$f(x|t, \mathcal{F}_{t-}) = \delta e^{-\delta x} \tag{24}$$

Hence, the completed mark intensity would be,

$$\lambda(t, x|\mathcal{F}_{t-}) = \left(\mu + \sum_{i:t_i<t} e^{\alpha x_i} \Big(1 + \frac{t-t_i}{c}\Big)^{-1}\right) \delta e^{-\delta x} \tag{25}$$

In this simulation study, we focus on the ground intensity since the mark density is the well-known exponential density whose best estimator is trivial: the inverse of sample mean.

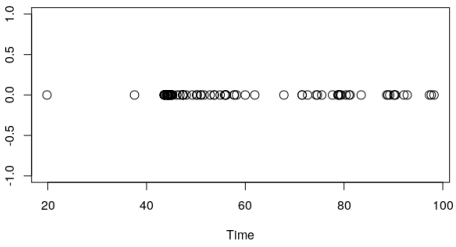
We set the true parameters as $\mu = 0.007$, $\alpha = 1.98$, $c = 0.008$ and $\delta = log(10)$. The simulation method we used is called the *thinning method*, introduced by Ogata (1981), Lewis and Shedler (1979). Briefly, this method first calculates an upper bound for the intensity function in a small time interval, simulating a value for the time to the next possible event using this upper bound, and then calculating the intensity at this simulated point. However these 'events' are known to be simulated too frequently (Lewis and Shedler, 1979). To overcome this, the method will compare the ratio of the calculated rate with the upper bound to a uniform random number to randomly determine whether the simulated time is treated as an event or not (i.e. thinning). A full description of the algorithm is provided in Appendix D.

We generate $N = 50, N = 100$ and $N = 200$ individual counting processes for each repeat of the simulation and in total we have $B = 1000$ repeats. The time-intervals are set to be $[0, 3000], [0, 500]$ and $[0, 100]$.

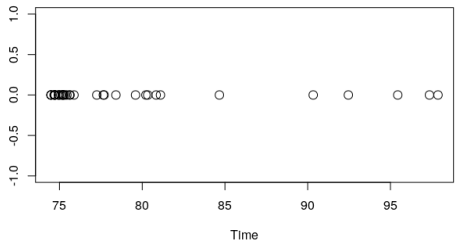
Figure 3 presents three quite different individual's event histories simulated by this ETAS DGP using the identical parameter settings as stated before.

Individual 1 has the most frequent events experience, the total number of events is 92. Individual 2 is somewhat moderate, with 37 events. Individual 3 has the least frequent events with only 2 during the time interval $[0, 100]$. Despite the hugely different behaviors, they are actually governed by the same intensity function. This example demonstrates that a self-exciting process can generate enough heterogeneity without introducing a latent variable to represent the unobserved heterogeneity.

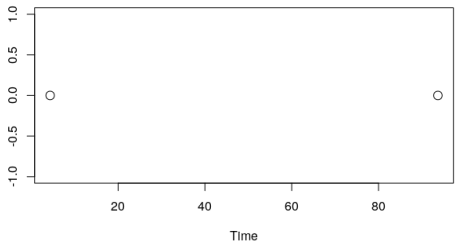
The estimation results are presented in Tables 1-3. As the number of observations N increase, the estimators become more stable and their empirical coverage rate gets



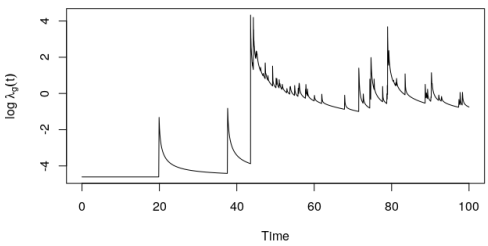
(a) Individual 1, Event Time



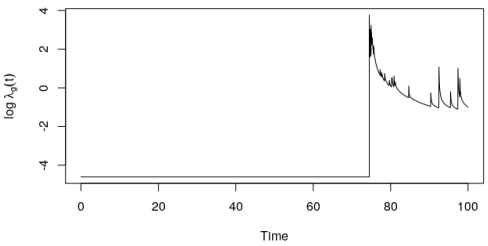
(b) Individual 2, Event Time



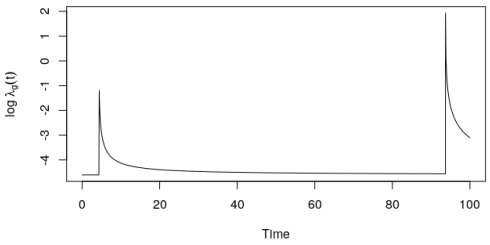
(c) Individual 3, Event Time



(d) Individual 1, log of intenstiy



(e) Individual 2, log of intenstiy



(f) Individual 3, log of intenstiy

Figure 5: Three individual's events histories

closer to the theoretical ones. It is also noticeable that the performance of estimators is closely related to the number of events per person. (We increase the length of the time horizon to increase such a number under the same true parameters.)

Table 1: Minimum Distance Estimator Results, with $T = 3000$

$N = 400$	True	MDE	sd	se	CI95	CI90
μ	0.007	0.006957441	0.0005575073	0.0006271522	94.9%	92. 5%
α	1.98	1.978269	0.04350423	0.07331051	93.5%	90.8%
c	0.008	0.008130796	0.001105742	0.001724244	93.9%	91.3%
Distance	1.48622	0.715594				
$N = 200$						
μ	0.007	0.006960397	0.0008548692	0.0008400477	93.6%	88.3%
α	1.98	1.984108	0.08547141	0.1086315	93.2%	90.9%
c	0.008	0.008042743	0.001568495	0.002413533	92.6%	90.3%
Distance	2.183783	1.226474				
$N = 100$						
μ	0.007	0.00684719	0.001007428	0.0011465	93.4%	90.9%
α	1.98	1.964071	0.06827856	0.1654297	92.1%	90.1%
c	0.008	0.008570634	0.001945219	0.003605053	92.3%	90.5%
Distance	3.169824	2.388298				
$N = 50$						
μ	0.007	0.006809876	0.001673515	0.001541488	89.1%	84.9%
α	1.98	1.974604	0.1713249	0.2765146	87.9%	83.7%
c	0.008	0.008979804	0.004142383	0.005475683	86.9%	83.1%
Distance	4.293006	3.474963				

Note: The distance is calculated using the semi-norm **??** with true parameters and the minimum distance estimators, respectively. sd is the standard deviation generated by the Monte Carlo simulation estimates. se is the mean of the standard error of each simulation. CI95(CI90) is the percentage of the 95%(90%) confidence interval generated by se that covers the true parameter.

5 The Data and State Dependence Test

5.1 The Data

The data we used come from the well-known RAND Health Insurance Experiment (RAND HIE), one of the most important health insurance studies ever conducted. It addressed two key questions in health care financing:

1. How much more medical care will people use if it is provided free of charge?
2. What are the consequences for their health?

The HIE project was started in 1971 and was funded by the Department of Health, Education, and Welfare. The company randomly assigned 5809 people to insurance

Table 2: Minimum Distance Estimator Results, with $T = 500$

$N = 400$	True	MDE	sd	se	CI95	CI90
μ	0.007	0.006818704	0.001200632	0.001282338	95.3%	93%
α	1.98	1.985548	0.05759593	0.2580961	96.3%	93.7%
c	0.008	0.008327916	0.002161951	0.005313165	96%	92%
Distance	0.2336264	0.1619424				
$N = 200$						
μ	0.007	0.007056179	0.001633427	0.001783448	92.5%	89.6%
α	1.98	1.977045	0.1709611	0.4486648	91.9%	90.6%
c	0.008	0.009058691	0.004623662	0.008174076	91.5%	89.9%
Distance	0.3579916	0.2119269				
$N = 100$						
μ	0.007	0.00660844	0.003243934	0.002295691	90.1%	86.1%
α	1.98	1.76104	0.4963242	0.85060127	86.6%	83%
c	0.008	0.01662388	0.0151135	0.0174853	86.7%	83.7%
Distance	0.477976	0.4551476				
$N = 50$						
μ	0.007	0.006672302	0.005083828	0.002964079	90.3%	88%
α	1.98	1.761366	0.5886596	2.207844	91.4%	88.9%
c	0.008	0.01808354	0.01897789	0.02508167	90.6%	87.8%
Distance	0.6452985	1.087129				

Note: The distance is calculated using the semi-norm **??** with true parameters and the minimum distance estimators respectively. sd is the standard deviation generated by the Monte Carlo simulation estimates. se is the mean of the standard error of each simulation. CI95(CI90) is the percentage of the 95%(90%) confidence interval generated by se that covers the true parameter.

Table 3: Minimum Distance Estimator Results, with $T = 100$

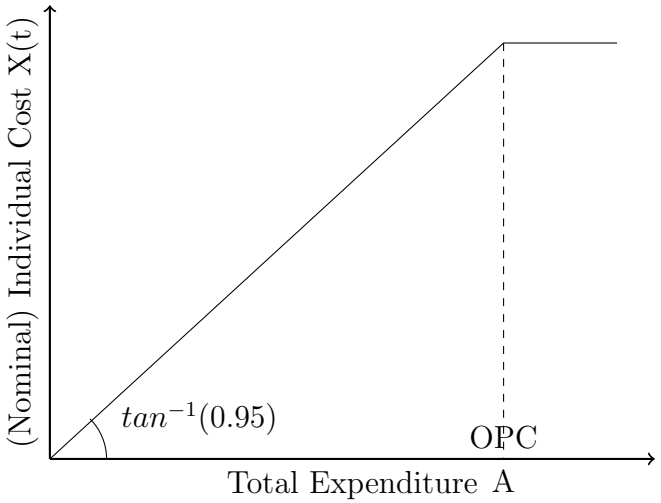
$N = 400$	True	MDE	sd	se	CI95	CI90
μ	0.007	0.0067466	0.001766587	0.002320197	95.2%	92.9%
α	1.98	1.980313	0.2536825	1.687546	95.1%	94%
c	0.008	0.01027362	0.007087134	0.01646008	95.4%	93.9%
Distance	0.0365799	0.0204637				
$N = 200$						
μ	0.007	0.006614259	0.003093314	0.002845468	93.6%	90.6%
α	1.98	1.91999	0.434588	2.273823	94.5%	93.3%
c	0.008	0.01357907	0.01251991	0.02549106	93.2%	92.1%
Distance	0.05125482	0.03664879				
$N = 100$						
μ	0.007	0.01317505	0.01093067	0.005716749	81.5%	75.7%
α	1.98	1.719879	0.7325846	2.227818	92.2%	89.6%
c	0.008	0.02089188	0.02165684	0.03664059	89%	86.9%
Distance	0.6294044	0.1808156				
$N = 50$						
μ	0.007	0.01273163	0.007051629	0.006974369	85.9%	82.9%
α	1.98	1.87436	0.8308396	3.961052	95.6%	93.5%
c	0.008	0.02130184	0.02805238	0.04548218	89.2%	87.2%
Distance	0.639077	0.1674467				

Note: The distance is calculated using the semi-norm **??** with true parameters and the minimum distance estimators respectively. sd is the standard deviation generated by the Monte Carlo simulation estimates. se is the mean of the standard error of each simulation. CI95(CI90) is the percentage of the 95%(90%) confidence interval generated by se that covers the true parameter.

plans that either had no cost-sharing, 25%, 50% or 95% coinsurance rates. The out-of-pocket cap varied among different plans. The HIE was conducted from 1974 to 1982 in six sites across the USA: Dayton, Ohio, Seattle, Washington, Fitchburg-Leominster and Franklin County, Massachusetts, and Charleston and Georgetown County, South Carolina. These sites represent four census regions (Midwest, West, Northeast, and South), as well as urban and rural areas.

Early literatures that use this data usually avoid the problem of non-linear budget constraint by assuming that individuals only respond to one price system. Typical econometric tools involved are the linear regression (after aggregating the data), the count data regression and the duration analysis. None of them is capable to fully model the stochastic structure of cumulative individual cost $X(t)$.

Because the complicated structure of our self-exciting process, to ease the burden of computation, we only use data from Seattle, which has the largest medical claim records available. We separate the data according to two different insurance plans: zero coinsurance rate plan (free plan, denoted as P0), in which the patient does not pay anything; and a cost-sharing plan (denoted as P95) in which a coinsurance rate of 95% applied and OPC is 150 USD per person or 450 USD per family[6](i.e., before exceeding the OPC, individuals need to pay 95% of the medical care cost, once the OPC is reached, all the cost is paid by the insurance.). The OPC and coinsurance rate in this plan only applied to ambulatory services; inpatient services were free. Both plans covered a wide range of services. Medical expenses included services provided by non-physicians such as chiropractors and optometrists, and prescription drugs and supplies. There is no deductible in this insurance contract. The following figure summarizes the P95 contract design.



Point A is the OPC . When the total expenditure is below A, the co-insurance (slope) $r = 0.95$ is applied. Whenever the total expenditure is beyond A, there is no cost for individuals.

Figure 6: Contract design for P95

We also include the data of drug purchase records with information such as the

[6]In 1973 dollars.

purchase dates and the values of non-covered charges. As discussed in the previous section, we may treat the drug purchase as another counting process and as an external shock to our primary one (doctor-visiting counting process). In the original dataset, one individual may have several claims in one day, and we combine all claims with an identical date into one and sum the non-covered charges.

The occurrence time stamp is defined as the annual duration between the beginning date of the insurance policy and the date this person visited a health care institution. For example, if the insurance begins on Jan-01-1977 and the date of a doctor visit is Oct-01-1977, the time stamp is then 0.748 (years). When preparing the dataset, we delete all the records with missing duration information. (Hence we exclude the cases of censoring.)

When analyzing the cost-sharing plan, we restrict our dataset within the contract year 1977-1978 since the cost-sharing policies are renewed annually. But such restriction is not needed for the free plan since there is no within-year cost sharing policy. For this plan, the time horizon ranges from 1975 to 1980. When the individual cost information is missing, we replace it with zero. In the end, we have 243 individuals in the free plan with 7638 claims over the years and 131 individuals in the cost-sharing plan and the total number of claims is 1103 within the 1977-1978 contract year.

We also include some demographic covariates: age, sex, education (in terms of schooling years) and log-income. For simplicity, we fixed all ages at the enrolment time. Thus all covariates are time-independent. More covariates can be added, but we are limited by computation capacity.

5.2 State Dependence Test

As mentioned before, we assume the source of individual heterogeneity is the state dependence. This assumption is well fitted into our self-exciting model as the filtration is generated by the past events. Here, we provide some evidence to support this assumption.

To test the state dependence, we follow Heckman's proposal (Heckman, 1981) and use the following strategy. Define an equispaced interval of time, say a week. For each period, we observe an individual's decision $d(i,t), i = 1, 2, \cdots, N; t = 1, 2, \cdots, T$ whether to visit the doctor. $d(i,t) = 1$ if individual i chooses to pay the visit in period t, and $d(i,t) = 0$ otherwise. For each individual, the history is made up of decisions $d(i,t), t = 1, 2, \cdots, T$. We may utilize this history data of sufficient length in the sample to estimate a regression of current medical utilization on previous ones. If the coefficient of past medical utilization is insignificant, we may conclude that there is no state dependence structure. Notice that we must permit each person to have his own fixed effect or intercept in the regression, otherwise one may face the danger that individual differences in absence probabilities will be correlated with past status.

Specifically, for each individual, write the regression

$$d(i,t) = \nu_i + \delta \frac{\sum_{t'<t} d(i,t')}{t} + U(i,t), t+1, \cdots, T \tag{26}$$

where $U(i,t)$ is a mean zero random variable of innovations uncorrelated with other innovations $U(i,t^{'}), t^{'} \neq t$, ν_i is an individual-specific effect and $d(i,0)$ is a fixed non-stochastic initial condition[7]. The term $\sum_{t^{'}<t} d(i,t^{'})/t$ measures the average frequency of medical utilization over a defined period.

The test is

$$H_0 : \delta = 0 \text{ VS } H_1 : \delta \neq 0$$

The following table shows the fixed effect estimation results for both the free plan and the cost-sharing plan. The results indicate that 1) The null hypothesis of no state dependence is rejected in both free and cost-sharing plans and 2) The average frequency of medical utilization is positively related to the probability of next medical utilization, even conditional on individual's fixed effect.

Table 4: State Dependence Check

	Dependent variable:	
	$d(i,t)$	
	Free Plan (P0)	*Cost-Sharing Plan (P95)*
δ	0.517584^{***}	0.126058^{**}
	(0.016532)	(0.051643)

Note: *p<0.1; **p<0.05; ***p<0.01

6 The Model

As discussed before, the focal point of the self-exciting counting process approach is to model the ground (cumulative) intensity function. We construct the intensities by explicitly taking different randomness sources of cumulative individual cost $X(t)$ and episode dependence structure into consideration.

6.1 Free Insurance Plan

Our ground intensity $\lambda_g(t)$ for each individual[8] who belongs to the free insurance plan consists of two parts:$\lambda_g(t) = \lambda_1\lambda_2(t)$. λ_1 deals with the covariates effect, while

[7]The requirement that $U(i,t)$ has a zero mean implies that the probability that $d(i,t) = 1$ conditional on $d(i,t-1), \cdots, d(i,0)$, and ν_i is

$$Pr[d(i,t) = 1 | d(i,t-1), \cdots, d(i,0), \nu_i] = \nu_i + \delta \sum_{t^{'}<t} d(i,t^{'})/t$$

[8]Therefore, we ignore the individual subscript.

λ_2 is the SD term discussed above.

Like many count data regression and duration models, the covariates effect is presented as an exponential function:

$$\lambda_1(Z) = exp(\gamma^T Z) \tag{27}$$

where Z is a vector of individual characteristics including age, sex, education and log-income, etc.

The SD term is specified as:

$$\lambda_2(t) = \sum_{i=1}^{N(t-)} \mu \cdot exp(-\mu(t - t_i)), \mu > 0 \tag{28}$$

When no event occurs before time t, we normalize the SD term to unit. That is $\lambda_2(t) = 1$ if $N(t-) = 0$. The ground intensity is then degenerated to $\lambda_g(t) = exp(\gamma^T Z)$. Implicitly, we assume that before the first doctor visit, the individuals do not understand their own health status and therefore, there is no individual heterogeneity.

The 'kernel' $\mu \cdot exp(\cdot)$ characterizes the episode dependence structure. More specifically, the propensity of a follow-up visit is governed by such a 'kernel': the intensity is high when the elapsed time is short and will gradually decrease as time goes by. We will argue such an assumption is reasonable: the individual is vulnerable when she just receives the treatment and is more likely to be sick again, but she will gradually recover as time goes by and will be less likely to experience sickness. The summation over these 'kernels' means we take all the past episodes into consideration. But the weight for each episode is different. By construction, the effects of far away past experiences will deteriorate, but the latest ones have the most important influences.

The usual method to model such phenomena in a structural form model is to assume health events arrive periodically with a probability $S^{'}$, which is drawn from $F(S^{'}|S)$ where S is the arrival probability from a previous period. Einav et al. (2015) further simplify this assumption by letting S take one of two values, S^L and S^H (with $S^L < S^H$), and that $Pr(S^{'} = S^J|S = S^J) \geq 0.5, J \in \{L, H\}$, so there is weakly positive serial correlation. This exceedingly simplified assumption is made mainly for computational reasons. And the above Markov process is most likely inadequate to model the episode cluster structure. We conclude that our cluster set up is more realistic and is quite difficult, if not impossible, to build within the conventional econometric models.

To sum up, for the free insurance plan, P0, the intensity is expressed as:

$$\lambda_{P0}(t) = \lambda_1(Z)\lambda_2(t) \tag{29}$$

6.2 Cost-Sharing Insurance Plan

As for the cost-sharing plan, λ_1 does not change. The cost sharing policy has two hypothetical effects: 1) The late year effect, that is when the contract year

is near the end, individuals, especially those who have already exceeded the OPC may use the medical service more frequently than before (cash-in effect) since the cost-sharing policy will be set to default next year and the shadow co-insurance rate would be expensive once again. 2) The shadow price effect discussed in the introduction section. We update the cumulative individual cost whenever an event occurs. To account for the cost-sharing effects, we modify λ_2 as follows:

$$\lambda_2^*(t) = \beta_1 exp(\beta_1 t) + \sum_{i=1}^{N_g(t-)} b \exp(\beta_2 X(t_i)) \mu exp(-\mu(t - t_i)) \tag{30}$$

here $X(t)$ is the cumulative individual cost at time t. It includes the non-covered charge from outpatient medical utilization as well as drug purchase:

$$X(t) = \sum_{i=1}^{N_g^1(t-)} x_i + \sum_{i=1}^{N_g^2(t-)} y_i \tag{31}$$

where x_i is the non-covered charge for i^{th} doctor visiting, $N_g^1(t)$ is the associated ground counting process. y_i is the non-covered charge for i^{th} drug purchase and $N_g^2(t)$ is the drug purchase ground counting process. The construction of $X(t)$ essentially follows the definition of cumulative individual cost mentioned in the introduction section. Recall the shadow price is defined as $1 - V(X(t))$, where $V(X(t))$ is the bonus which depends on the cumulative individual cost. If $V(X(t)) \propto \exp(\beta_2 X(t))$, then the term $b \exp(\beta_2 X(t))$ can be thought of as a measure of medical utilization bonus. We would expect $\beta_2 > 0$ to be significant if individuals do respond to shadow price.

We use the term $\beta_1 exp(\beta_1 t)$ to model the late year effect: we would observe β_1 significantly greater than zero if such an effect is true.

To summarize, the ground intensity for the cost-sharing plan is:

$$\lambda_{P95}(t) = \lambda_1(Z)\lambda_2^*(t) \tag{32}$$

There are several pieces to put together in order to estimate the parameters of the cost-sharing effects model. As Keeler and Rolph (1988), we assume that there are no interactions between within-year cost sharing effects and the effects of other explanatory variables, so that all the effects of explanatory variables other than cost sharing on frequencies of episodes are summarized in $\lambda_1(Z)$ and all episode dependence structure is captured by $\lambda_2(t)$ ($\lambda_2^{'}(t)$). We first estimate the free plan by minimizing

$$||\bar{N}_g^{P0} - \lambda_1(Z) \textstyle\int_0^T \lambda_2(t)dt||_{\bar{N}_g^{P0}}$$

thus, the individual heterogeneity and the episode dependence structure of the intensity are estimated by $\hat{\lambda}_1(Z)$ and $\hat{\lambda}_2(t)$. When estimating the cost-sharing plan, these two parts are then treated as fixed, which leaves us with only cost-sharing

effect parameters (i.e.,β_1, β_2 and b) to be estimated[9]. Thus the minimization object is:

$$||\bar{N}_g^{P95} - \hat{\lambda}_1(Z) \int_0^T \big(\beta_1 exp(\beta_1 t) + \sum_{i=1}^{N_g(t-)} b \exp(\beta_2 X(t_i)) \hat{\mu} exp(-\hat{\mu}(t - t_i))\big) dt||_{\bar{N}_g^{P95}}$$

7 Main Results

The main results are presented in Table 5. Some words on the numerical optimization are in order. Because of the non-linear nature and complexity of the cumulative intensity function, it is impossible to write down a closed form solution for this minimization problem. And some numerical algorithms are implemented to find the optimization. For the free plan, we first run a simulated annealing (SA) routine to assess reasonable ranges for all the parameters, then a down-hill (Nelder-Mead) optimization algorithm is employed to refine the results. Similar steps are used for the cost-sharing plan with individual specific parts of cumulative intensity fixed. We manually stop the SA algorithm after 24 hours but do not intervene with the down-hill algorithm until it reports success.

7.1 Interpreting the Covariates

The interpretation of coefficients is not as straightforward as in linear regression. However, we may fix a time period and treat the counting process as count data. The interpretation is then identical to that of a count data regression analysis. Formally, recall the Doob-Meyer decomposition, for a fixed time period $[0, t], \forall t \in [\underline{t}, \bar{t}]$, we have

$$\mathbb{E}(\Lambda_g(t|Z)) = \mathbb{E}(N_g(t)|Z) = \mathbb{E}(Y_t|Z)$$

The count data Y_t is the number of events occurring during this time period. Let scalar z_j denote the j^{th} covariate. Differentiating

$$\frac{\partial \mathbb{E}(Y_t|Z)}{\partial z_j} = \gamma_j \mathbb{E}(\Lambda_g(t|Z))$$

by the exponential structure of $\lambda_1(Z)$. That is, for example, if $\hat{\gamma_j} = 0.2$, $\bar{\Lambda}_n(t|Z) = 2.5$, then one-unit change in the j^{th} covariate increases the expectation of Y_t by 0.5 units.

Two remarks are in order. First, notice that the sign of the response $\partial \mathbb{E}(Y_t|Z)/\partial z_j$ is given by the sign of γ_j since the accumulated intensity Λ_t is always positive. Second, if one covariate coefficient is twice as large as another, then the effect of a one-unit change of the associated covariate is double that of the other. This result follows from

$$\frac{\partial \mathbb{E}(Y_t|Z)/\partial z_j}{\partial \mathbb{E}(Y_t|Z)/\partial z_i} = \frac{\gamma_j \mathbb{E}(\Lambda_g(t|Z))}{\gamma_i \mathbb{E}(\Lambda_g(t|Z))} = \frac{\gamma_j}{\gamma_i}$$

[9]We exploit the fact that all individuals are assigned to different plans randomly. By plugging the individual specific estimators from the free plan into the cost-sharing plan, we can still have consistent estimators.

Table 5: Basic Results

	Estimator	*Description*
μ	27.87126321***	coefficient of the episode dependent structure
	(8.39814247)	
age	-0.13394806***	
	(0.03122223)	
age2	0.15470947***	$(age)^2/100$
	(0.04225468)	
male	-0.71703944	
	(0.4738397)	
edu	-0.35495029***	
	(0.0858603)	
edu2	0.99428386***	$(edu)^2/100$
	(0.3361197)	
log income	0.59265516***	
	(0.03643453)	
b	0.65898635***	
	(0.0580308)	
β_1	0.1068388	coefficient of late year effect
	(0.27547018)	
β_2	0.00383393***	coefficient of non-covered charge
	(0.00061892885)	
Distance	0.898473	Free Plan
Distance	1.10226	Cost-Sharing Plan

Note: standard errors in brackets, *p<0.1; **p<0.05; ***p<0.01

With these in mind, we can interpret our results.

Age. The overall effect for age is as follows: at first, the intensity will decrease as age increases, after one passes the age of 43, the intensity and age are positively correlated. It is well-known that the youngsters are more risky compared to their mid-age counterparts. While as individuals begin to age, they become physically weaker and more prone to sickness.

Sex. Females seem to be more likely to go the doctor, but the result is not significant.

Education. There are two explanations about the correlation between education and frequency of doctor visits. One is that individuals with a higher level of education are positioned in more important jobs and their absence from work may damage not only their output but also that of their peers', thus the potential cost of going to hospital is much higher which leads to a negative correlation. The other explanation says that with higher education, people are more aware of the importance of good health and are willing to go to the doctor more frequently, besides, education and income are known to be positively linked, thus education should be correlated to intensity positively. Our results suggest that with education of less than 17 years (roughly equivalent to a Master's degree), the overall effect favors the first explanation. But with a higher education level (Master and above) the overall effect favors the second explanation.

Income. Income is positively related to the use of medical services, which is not surprising. A higher income gives individuals the ability to cover the opportunity cost related to absence from work (to visit a doctor).

7.2 Cost-Sharing Effects

There is weak evidence supporting the existence of the late year effect (t-ratio of 0.387842).

The shadow price effect is captured by $b\exp(\beta_2 R(t))$. The most important parameter here is β_2. If β_2 is close enough to zero, we may observe a flat, almost linear curve, which indicates that individuals only respond to one price system (the spot price system). However, if β_2 is positively away from zero, we can safely claim that individuals do understand the design of the insurance policy and take advantage of the shadow price. Our result provides strong evidence for the shadow price effect and we are confident to reject the null hypothesis: $\beta_2 = 0$. Figure 7 shows the graphical result.

Overall, the model fits the data well. To assess the goodness of fit, we generate estimated (averaged) cumulative intensity against the observed (averaged) counting process. Figure 8 presents the results. The patterns of estimated and observed are quite similar in both free plan and cost-sharing plan.

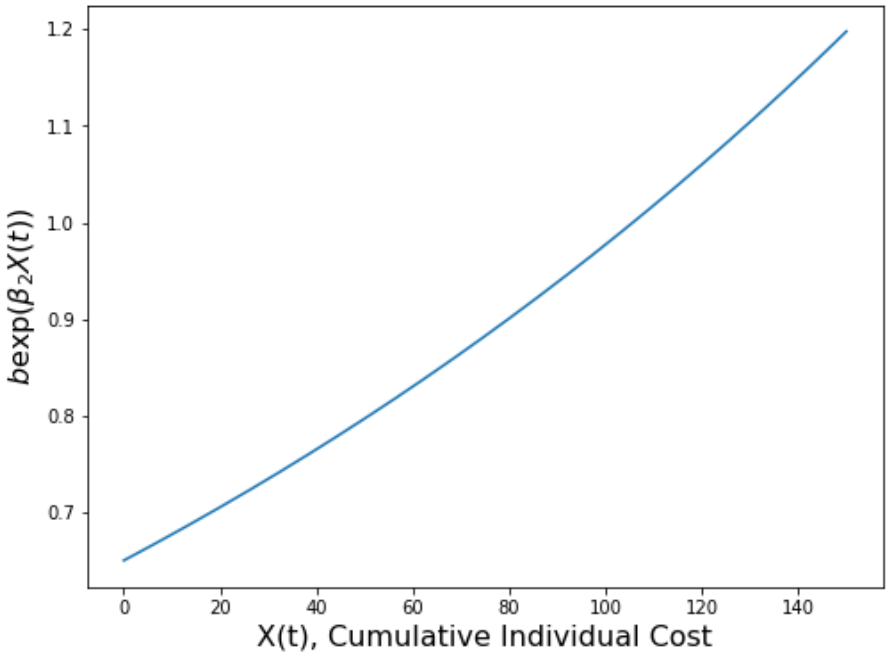
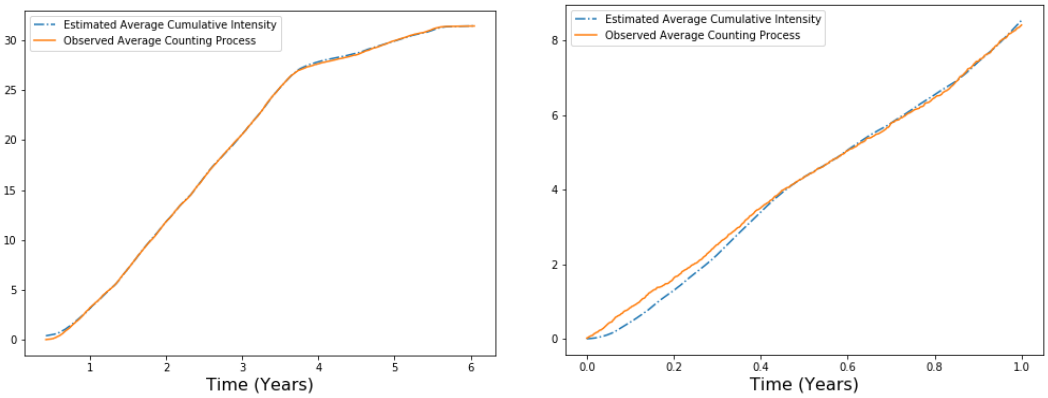


Figure 7: The shadow price effect



(a) free plan

(b) cost-sharing plan

Figure 8: Goodness of fit

7.3 Mark Density

To complete our model, we will non-parametrically estimate the conditional mark density. The interested mark in this application should be the non-covered charge. However, in the free plan, such marks are zero in most records. Because of this, and for the purpose of comparison between these two plans, we instead use total charge as our marks. Since insurance plans have clear regulation on the cost-sharing policies, once we observe the total charge, there is no ambiguity in knowing the non-covered charge.

We assume the nominal price are i.i.d distributed. Thus we have $f(x|t, \mathcal{H}_{t-}) =$

$f(x)$. A standard kernel density estimation is used to analyse this mark density. The bandwidth selection results are 3.84075 and 5.79464 for cost-sharing and free plans respectively. We plot the densities and distributions in Figure 9.

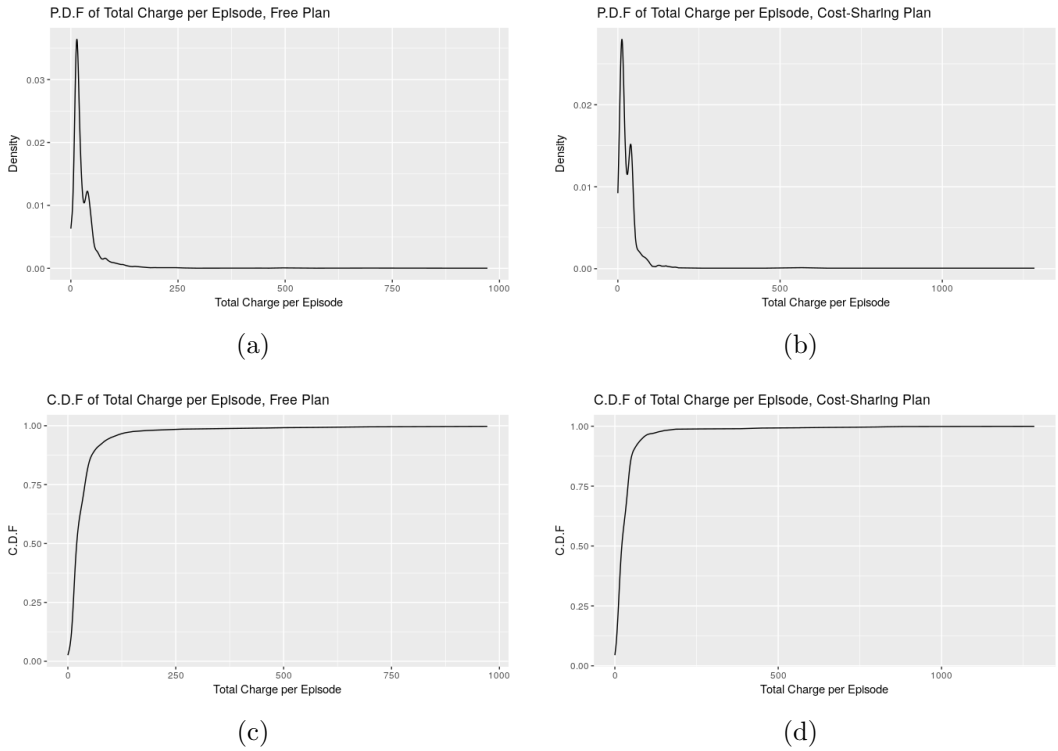


Figure 9: Plots of total charge empirical distributions with a Gaussian kernel

It is not hard to tell that the densities in the two different plans are similar, indicating the charge per episode is stable across various insurance plans. This result is consistent with previous studies (e.g. Keeler and Rolph (1988))

7.4 Cluster Analysis

The episodes tend to be clustered or grouped together (i.e., we are rejecting the assumption that episodes are independent). One reason is because of the nature of chronic diseases: regular or frequent treatments are needed to ease or eliminate the pain. Another explanation is because one disease may trigger the occurrence of another one in the short term.

As mentioned before, the dependent structure (or the cluster structure) is governed by $\exp(-\mu t)$. Figure 10 presents such a structure using our estimator $\hat{\mu}$.

The likelihood of follow-up visiting is high at the beginning and then decreases as time goes by. After roughly 3 weeks to one month, the likelihood is small enough to ignore.

Literatures have documented that cost-sharing policies reduce the frequency of medical utilization (e.g. Keeler and Rolph (1988); Aron-Dine et al. (2015)). Our question is how do these policies affect the cluster structure among doctor visiting?

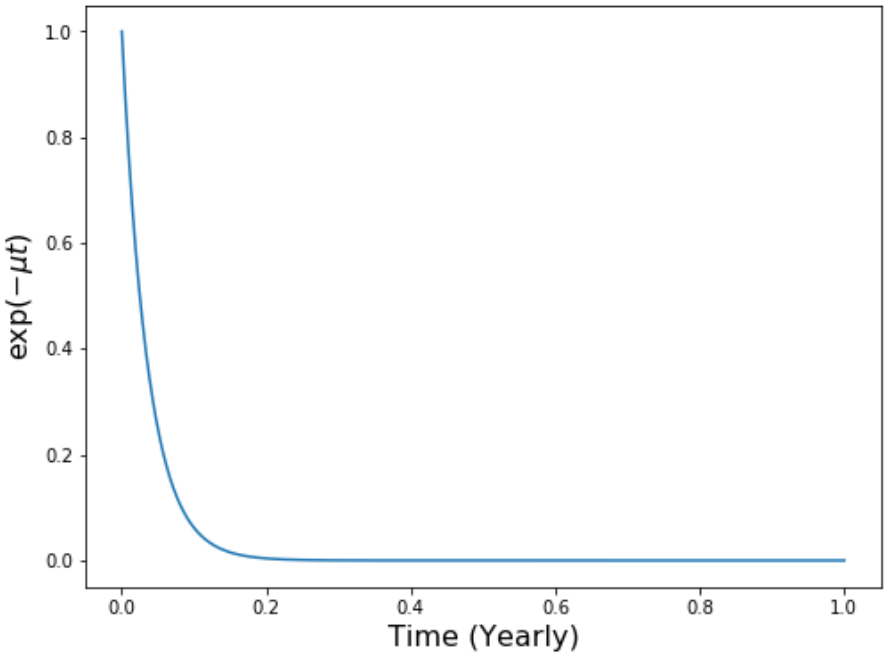


Figure 10: The cluster structure among episodes

Will they reduce the average number of clusters per person? Will they reduce the average follow-up visits inside a cluster? To the best of our knowledge, few literature has touched on these issues, since most of them use the hypothesis that episodes are independent.

Here we use a cluster analysis algorithm called DBSCAN (Density-based spatial clustering of applications with noise) that is widely used in computer science and statistical learning.

For this algorithm, there are two inputs: Eps, the radius of one density region, and minPts, the minimum number of points required to form a dense region. For the purpose of DBSCAN clustering, all points are classified as core points, border points and noise points. Core and border points form a cluster via different definitions of 'reachable'. Noise points are the points that do not belong to any cluster. We provide details of this algorithm and the definition of a cluster in Appendix E.

The ability of this algorithm to identify 'noise' points is particularly appealing to us. This is because some acute episodes are small in scale and only need one doctor visit to fully recover. They are not linked to the rest of episodes.

Based on the estimation of the SD term, we set $Eps = 21$ days and as a rule of thumb $minPts = 2^{10}$. For the purpose of comparison, we restrict the time horizon in both plans to 1977-1978 insurance year. For each individual (both free plan and cost-sharing plan), we run the DBSCAN algorithm, document the number of clusters, the average number of instances per cluster and the number of noise points. For each insurance plan, we then compute the average number of clusters per person, the average number of instances per cluster per person and the average noise points

[10]The rule of thumb is minPts = dimension +1

per person. Table 6 summarizes the results.

Table 6: Cluster Analysis

	avg cluster number	avg cluster members	avg noise points
free plan	1.2287	4.55187	1.62332
Cost-sharing plan	0.862595	3.3625	1.47328

The effects of cost-sharing policies on cluster structure are threefold. First, they reduce the average number of clusters per person. That means for the initial episode, the cost-sharing policies suppress the first doctor visiting behaviors. Second, within each cluster, they reduce the number of follow-up visits. Third, cost-sharing policies reduce the average number of noise points per person, i.e., they discourage individuals to use medical services when they have small episodes like minor injuries.

8 Discussion

Our model characterizes the individual's decision making process in both free and cost-sharing plans. The results show that individuals react to shadow price systems. The method we adopted is different than conventional reduced form or structural form econometrics tools used in non-linear budget constraint studies.

In the literature of medical utilization, how to quantify the response of medical spending with respect to its price is at the core of the debate. Cutler and Zeckhauser (2000) summarize a long list of literature all reporting an estimate of a single price elasticity of demand for medical care with respect to the out-of-pocket price. Particular to the data we used, the RAND HIE, such an estimate is -0.2(Manning et al., 1987),(Keeler and Rolph, 1988)). Most of these literatures obtain the estimate of such single elasticity by assuming individuals only respond to the spot (out-of-pocket) price. Recent literatures (e.g. Cardon and Hendel (2001), Dalton (2014), Kowalski (2015),Aron-Dine et al. (2015) and Einav et al. (2015)) deviate from this assumption and find strong evidence for such deviation.

Aron-Dine et al. (2015) proposes to use two different elasticities in a classical reduced form: one with respect to spot price and the other with respect to future price. They define the future price as the expected end-of-year price, with expectations taken over all individuals in the same insurance plan. Such future price depends on three elements: the cost-sharing features of the insurance plan, the duration of the insurance plan and the expected spending of individuals. Thus, essentially, they impose a strong assumption that individuals have no private information about their health conditions and health shocks. Although using a firm-level data, they find strong evidence rejecting the null hypothesis that individuals only respond to spot price, they fail to find similar conclusions using the Rand HIE data. The explanation they give for this result is the relatively small sample size.

Our method, on the other hand, tackles the problem of quantifying the response of medical spending from a different perspective: rather than estimating the elasticities directly, we try to measure the bonus $V(X)$ as a function of cumulative individual

cost. By doing so, we avoid unrealistic assumptions made by Aron-Dine et al. (2015) but at the same time manage to keep the non-linear property. In addition, we exploit the data in its maximum capacity. Instead of just looking $X(T)$ like Aron-Dine et al. (2015), we use all the cumulative individual cost information: $X(t) : t \leq T$. Thus, in our method, a more direct measurement of responses to the out-of-pocket spending up until now (β in our model) is estimated instead of the elasticities. We believe such features explain our success in finding evidence supporting the non-linear budget constraint in the Rand HIE data.

In a structural paper,Einav, Finkelstein, and Schrimpf (2015) build and estimate a simple dynamic model of optimizing agent's drug utilization decisions given a non-linear insurance contract design. They also find evidence supporting the hypothesis that individuals take into account the dynamic incentives by showing the discount factor of their value function is non-zero. The shortcoming of this structural form is that it fails to address the stochastic nature of the occurrences. What structural models do is to divide the time line into several equispaced cells, and within each cell, one only needs to concern whether an event is occurred or not. As for how many events occurred in that cell and more importantly the timings of occurrences, they are irrelevant. The self-exciting approach, on the other hand, is more adequate to model the stochastic properties of individual cost $X(t)$: the events' occurrence times $\{t_i\}_{i\in\mathcal{N}^+}$ are at the central of our model objectives. In addition, compared to the structural form, our method is much simpler and only requires minimal assumptions. For example, it is a common practice in structural form to use a Markov transition probability to describe the shock dependence. Einav et al. (2015) further simplify it by restricting the shocks to be either big or small. These assumptions are mainly made for computational purpose. There is littler reason to believe that the real mechanism would follow such restriction. In comparison, our approach allows all the past experiences contribute to future randomness in a form of state dependence. This is a much less restricted and more realistic and complicated assumption. Despite that, our model is still computationally easier than a typical structural form econometrics model.

9 Conclusion

In this paper, we provide a methodology to construct a behavioral model of medical care utilization. At the core of this method is the self-exciting counting process. It allows researchers to take historical information into the model. A minimum distance estimation is advocated instead of the conventional likelihood-based methods. By doing so, one may introduce external shocks to the self-exciting process. This enables researchers to use more realistic model settings. We use such a methodology to build a decision making process model of medical care utilization and find that individuals are responsive to shadow price and take into account the dynamic incentives. Furthermore, we use a matured statistical learning algorithm to analyze the cluster structure of doctor visiting behaviors. We find cost-sharing policies do affect the clusters in numerous ways.

References

Yacine Aït-Sahalia, Julio Cacho-Diaz, and Roger JA Laeven. Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606, 2015.

Aviva Aron-Dine, Liran Einav, and Amy Finkelstein. The rand health insurance experiment, three decades later. Technical report, National Bureau of Economic Research, 2012.

Aviva Aron-Dine, Liran Einav, Amy Finkelstein, and Mark Cullen. Moral hazard in health insurance: do dynamic incentives matter? *Review of Economics and Statistics*, 97(4):725–741, 2015.

Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

Zarek C Brot-Goldberg, Amitabh Chandra, Benjamin R Handel, and Jonathan T Kolstad. What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics. *The Quarterly Journal of Economics*, 132(3):1261–1318, 2017.

James H Cardon and Igal Hendel. Asymmetric information in health insurance: evidence from the national medical expenditure survey. *RAND Journal of Economics*, pages 408–427, 2001.

David M Cutler and Richard J Zeckhauser. The anatomy of health insurance. *Handbook of health economics*, 1:563–643, 2000.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*, volume 1,2. Springer Science & Business Media, 2007.

Christina M Dalton. Estimating demand elasticities using nonlinear pricing. *International Journal of Industrial Organization*, 37:178–191, 2014.

Liran Einav, Amy Finkelstein, and Paul Schrimpf. The response of drug expenditure to nonlinear contract design: Evidence from medicare part d. *The quarterly journal of economics*, 130(2):841–899, 2015.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

Benjamin R Handel, Jonathan T Kolstad, and Johannes Spinnewijn. Information frictions and adverse selection: Policy interventions in health insurance markets. Technical report, National Bureau of Economic Research, 2015.

James Heckman and Burton Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320, 1984.

James J Heckman. Heterogeneity and state dependence. In *Studies in labor markets*, pages 91–140. University of Chicago Press, 1981.

Alan Karr. *Point processes and their statistical inference*, volume 7. CRC press, 1991.

Emmett B Keeler and John E Rolph. The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics*, 7(4):337–367, 1988.

Emmett B Keeler, Joseph P Newhouse, and Charles E Phelps. Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty. *Econometrica*, pages 641–655, 1977.

Kai Kopperschmidt and Winfried Stute. The statistical analysis of self-exciting point processes. *Stast. Sinica*, 23:1273–1298, 2013.

Amanda E Kowalski. Estimating the tradeoff between risk protection and moral hazard with a nonlinear budget set model of health insurance. *International journal of industrial organization*, 43:122–135, 2015.

Peter A Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

Willard G Manning, Joseph P Newhouse, Naihua Duan, Emmett B Keeler, and Arleen Leibowitz. Health insurance and the demand for medical care: evidence from a randomized experiment. *The American economic review*, pages 251–277, 1987.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2012.

Yosihiko Ogata. On lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 27(1):23–31, 1981.

Yosihiko Ogata and Koichi Katsura. Likelihood analysis of spatial inhomogeneity for marked point patterns. *Annals of the Institute of Statistical Mathematics*, 40(1): 29–39, 1988.

Gerard J Van den Berg. Duration models: specification, identification and multiple durations. *Handbook of econometrics*, 5:3381–3460, 2001.

Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002.

A The KNP Theorem

Keeler,Newhouse and Phelps (1977) Keeler et al. (1977) considered the theory of a consumer facing a deductible health insurance contract. There are two differences from the usual consumer theory: 1) the price per unit in a specified amount of medical care services bought within a specified period is changeable and 2) illness uncertainty is present regarding the future medical service demand. When uncertainly is present, any cost below the range of deductible limit has the bonus of reducing the remaining deductible, and hence reducing the future cost. That is the greater the chance that future expenditures will exceed the deductible, the cheaper today's purchase of medical care service. They argue that when analysing the reaction to the deductible of a consumer, one needs to take account of the sequential decision problem.

Formally, they assume a rational agent whose object is to maximize a sequence of utilities $(U(e_t, H_t))_t$ under illness uncertainty, where, after ignoring the time subscription, e is the flow of other consumption goods and H, in terms of dollar, is the stock of health which is related to the medical care service h. The evolution of the perceived health stock is $H_t = H_{t-1} - l_t + g(h_t, l_t)$, where l, in terms of dollar, is the random loss of health from illness and $g(h, l)$ is the production function of the stock of health. Before exceeding the deductible, denote $o_t = p_h \times h_t$ as the out-of-pocket payments for medical care at time t where p_h is the medical service price, after reaching the deductible, the co-insurance rate C would apply to the consumers, we normalize the price of other goods to 1. Let $(d_t)_t$ be a sequence of the remaining deductibles.

The one-period demand case is shown in the figure 11

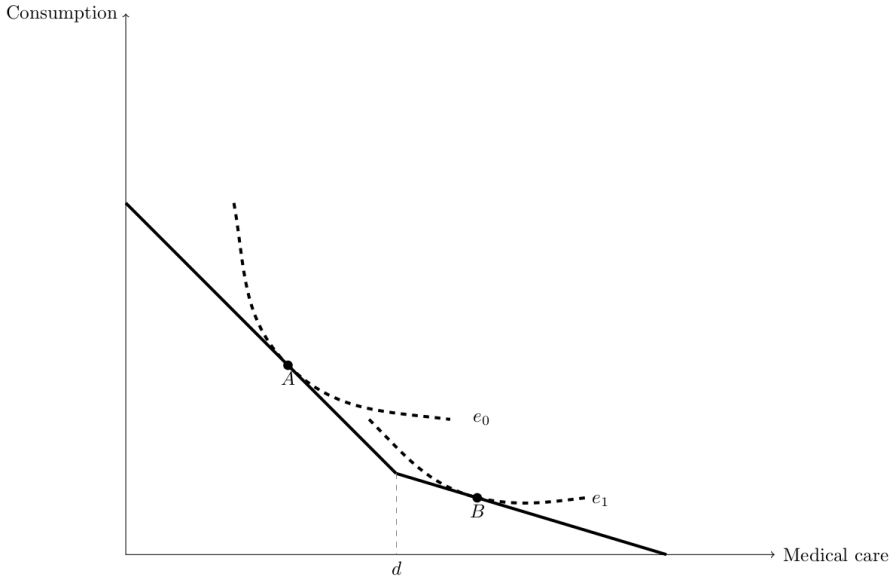


Figure 11: An insurance policy with a deductible equal to d

They have shown that in this one-decision case that consumers with smooth indifference curves would never be observed purchasing exactly d units of medical care and, in general, would not be near the kink in the budget line. Furthermore, with a normal utility function, there exists a critical level of illness l_d. All losses smaller than l_d will produce purchases of h less than d, and losses greater than l_d will lead to $h > d$.

Suppose $U(X, H)$ applies only to the stocks of other goods and health at the end of the deductible period. The multi-periods problem can be represented formally as a consumer trying to maximize the expected value of:

$$\sum U_t(e_t, H_t) + U(X, H) \tag{33}$$

subject to the budget constraint:

$$X + \sum(e_t + o_t) = Income \tag{34}$$

Notice that it is easy to see that the medical service needed to cure the illness is related the scale of illness l, hence we should let the other consumption and the medical service be functions of l: $e(l), h(l)$.

Such problem can be formulated as a dynamic program. Let $W(x_t, H_t, d_t, t)$ be the expected sum of utility from t to the end of the deductible period, for an agent who at the start of period t has wealth x_t, health H_t and remaining deductibles d_t. We suppose that the distribution of illness, $f(l, H_t)$, also depends on H_t. Then:

$$W(x_t, H_t, d_t, t) = \int \max_{e(l),h(l)} U\{e(l), H_t - l + g[h(l)]\} + W(x_t - e(l) - p_h h(l), H_{t+1}, d_{t+1}, t+1) f(l, H_t) dl \tag{35}$$

where $d_{t+1} = max(0, d_t - o_t)$. The formula for the remaining deductibles means one needs to take this optimization problem in a sequential and contingent way: all the history deductible information should be included to calculate the current remaining deductible. W can be calculated backwards from the end of the period to his present position.

Such results have important implications for estimating demand curves for medical services. For a person's j^{th} illness, the demand can be expressed as:

$$q_j = f(p_j, Z) + \epsilon_j$$

where q_j is the quantity demanded by this individual in the j^{th} episode, p_j is the shadow price for j^{th} episode, Z is vector of other variables that may affect the demand, and ϵ_j is a random error term with $\mathbb{E}(\epsilon_j) = 0$, $var(\epsilon_j) < \infty$.

Given the existence of remaining deductible, the model suggests that each p_j is a function of demands (hence out-of-pocket fees) prior to the j^{th} episode and of time

remaining in the period,

$$p_j = g\Big(\sum_{k=1}^{j} q_k, t_j\Big)$$

Suppose T is the number of episodes occurring during one year and there is no variation in T across individuals, then the average effective price is

$$p = \sum_{j=1}^{T} p_j/T \tag{36}$$

Most of the literature did not use the p as the measurement of price; instead, they use the fraction of annual expenditures paid out of pocket $\bar{p}$. However, $\bar{p}$ measures the effective price with error: $(\bar{p} - p) = \delta$. The authors argue that since the price measurement error δ is not random but is generally correlated with the true price, the estimated price coefficient could be inconsistent.

B Basic Concepts, technique version

In this section, we present some basic concepts of counting process in much more detail.

B.1 (Marked) Counting Process

In order to study the definition of the point process in a rigorous way, introduction of some concepts is in order.

To begin with, let's first introduce the 'working' space. Let **E** be a locally compact Hausdorff space whose topology has a countable base (LCCB), with Borel $\sigma-$ algebra ϵ. Denote β the algebra (ring) of bounded Borel set.

Let M_E be the space of all bounded finite measures on $\beta(E)$, the random measure maps from probability space $(\Omega, \mathcal{F}, \mathcal{P})$ to the measurable space $(M_E, \beta(M_E))$. Notice that on M_E we usually define the $\sigma-$ ring $\mathcal{M}$ generated by the coordinate mappings:

$$\mu \to \mu(f) = \int f d\mu$$

where μ is a Borel measure and f ranges over the set C_k of continuous functions whose support is compact.

Definition A point process on E is a measurable mapping N of $(\Omega, \mathcal{F})$ into $(M_p, \mathcal{M}_p)$, where $M_p = \{\mu \in M_E : \mu(A) \in \mathbb{N}^+ \text{ for all} A \in \beta\}$, $\mathcal{M}_p = \mathcal{M} \cap M_p$ and $\mathbb{N}^+$ is non-negative nature numbers.

In this paper, we focus on the case the space E is actually a time space, hence loosely speaking, we can define the point process in a counting process way:$N_t = \sum_{i=0}^{\infty} \mathbb{I}(t_i \leq t)$, moreover, throughout the whole paper unless other mentioned, we restrict the point process to be simple, i.e., $Pr\{N(\{[t, t+\Delta t]\}) = 0 \text{ or } 1 \text{ for all t}\} = 1$, that is no common jumps are allowed.

Remark With every sample point $\omega \in \Omega$, we associate a particular realization that is a boundedly finite Borel measure on E: it may denoted by $N(\cdot, \omega)$. While for each fixed set A, $N(A, \cdot)$ is a non-negative random variable. In practice, the latter means if we fix a time period $A = [0, t]$, the count data $Y_t = N_t$ is the number of events happened during this period.

Definition Given a point process N defined on $(\Omega, \mathcal{F}, \mathcal{P})$, the distribution of N is $\mathcal{P}N^{-1}$ on $(M_p, \mathcal{M})$

We can easily extend the definition to a marked point process, let $N_t = \sum_{i=0}^{\infty} \mathbb{I}(t_i \leq t)$ be a point process on E and let $E^{'}$ be a second LCCB. We define the marked point process with underlying process N and realization marks $\{x_i\}_i$ as any point process that,

$$N_t(E^{'}) = \sum_{i=0}^{\infty} \mathbb{I}(t_i \leq t, X_i \in E^{'})$$

on $E \times E^{'}$. The random element X_i of $E^{'}$ is called the mark associated with t_i.

B.2 Intensity

The idea of intensity is closely related to the compensator of point process N, which is a random measure Λ interpreted as the local conditional mean of N given its strict past, informally $\Lambda(dt) = \mathbb{E}[N(dt)|\mathcal{F}_{t-}]$. The compensator is usually smoother than the point process. When the compensator is absolutely continuous, the derivative is a random process called the intensity, i.e., $\Lambda([0, t]) = \int_0^t \lambda_u du$. The marked counterpart is expressed as $\Lambda_t(E^{'}) = \int_0^t \int_{x \in E^{'}} \lambda(u, x) dx du$.

Notice that the compensator is conditioning on a filtration $\mathcal{F}_{t-}$ interpreted as *strict past* or *strict history*. It is natural to ask 'how to define history' and 'what is filtration'? It turns out that one can use filtration to define history, hence next we are going to carefully define filtration and discuss its structure.

We will introduce the filtration under the framework of marked point process, regular point process can adopt such definition after ignoring the mark space. Let the marked point process N_t be defined as before on $\mathbb{N}^+$ with mark space $(E^{'}, \varepsilon)$, for $B \in \varepsilon$, let $N_t(B) = \sum_{i=1}^{\infty} \mathbb{I}(t_i \leq t, x_i \in B)$. Define the filtration as,

$$\mathcal{F}_t = \sigma(N_s(B) : 0 \leq s \leq t, B \in \varepsilon) \tag{37}$$

We term such filtration as the internal history of N. Notice that the filtration defined in this way is right continuous in a sense that for each t, $\mathcal{F}_t = \cap_{h>0}\mathcal{F}_{t+h}$.

To describe the structure of the filtration, we need two more concepts, stopping times and predictability. The former one is a necessary tool to describe a counting process and define history, while the latter is the key ingredient of the Doob-Meyer decomposition, upon which the minimal distance estimation method is based.

Definition A random variable T with values in $[0, \infty]$ is a stopping time if $\{T \leq t\} \in \mathcal{F}_t$ for every t. Intuitively, time T of a random event is a stopping time if at each fixed time t, one can observe whether or not the event has already occurred.

Associated with the stopping time T are the $\sigma-$ algebra,

$$\mathcal{F}_T = \sigma\{\Lambda \in \mathcal{F}_\infty : \Lambda \cap \{T \leq t\} \in \mathcal{F}_t \text{ for all t}\}$$

which sometimes be called 'the past before T' and the $\sigma-$ algebra

$$\mathcal{F}_{T-} = \sigma\{\Lambda \cap \{T > t\} : t \geq 0, \Lambda \in \mathcal{F}_t\} \vee \mathcal{F}_0$$

which comprises the strict past of $\bar{N}$

In the above two equations, $\mathcal{F}_0$ is the filtration that stores some foreknowledge of the marked point process, that is,

$$\mathcal{F}_0 = \liminf_{t>0} \mathcal{F}_t$$

And the the filtration $\mathcal{F}_\infty$ contains the common information among all the filtrations,

$$\mathcal{F}_\infty = \cap_{t>0}\mathcal{F}_t$$

Next, we will study the predictability, a key ingredient of Doob-Meyer decomposition. For simplicity, we restrict ourselves to finitely many t's, say $0 = t_0 < t_1 < t_2 < \cdots < t_k$. Put $N_i = N_{t_i}$ for short. At time t_i, only $N_0, N_1, \cdots, N_i$ are known, but not necessarily $N_{i+1}, N_{i+2}, \cdots, N_k$. We are concerned about how to predict future values of N_j given the information at $t = t_i$. For a variable Λ_{i+1} to be a predictor for N_{i+1} at t_i means this Λ_{i+1} needs to be known at t_i. Thus

Definition Let $(\mathcal{F}_i)_{0\leq i\leq k}$ be some increasing filtration, and let $(\Lambda_i)_{0\leq i\leq k}$ be a sequence of random variables. Then we call

$$(\Lambda_i)_i \textbf{ predictable w.r.t}(\mathcal{F}_i)_i \text{ iff } \Lambda_{i+1} \textbf{ is } \mathcal{F}_i \textbf{ -measurable}$$

Now, we can describe the structure of the filtrations $\mathcal{F} = (\mathcal{F}_t)$ as follow:

Proposition For each n, if T_n is a stopping time of $\mathcal{F}$ and $A_t(E')$ is a predictable process, then we have,

$$\mathcal{F}_{T_n} = \sigma((T_1, X_1), \cdots, (T_n, X_n))$$
$$\mathcal{F}_{T_n-} = \sigma((T_1, X_1), \cdots, (T_{n-1}, X_{n-1}), T_n)$$

The proof can be found in Karr (1991) Karr (1991).

Intuitively, the accumulating information changes only at the arrival times T_i, whereas $T_i \in \mathcal{F}_{T_i-}$, the mark X_i belongs to $\mathcal{F}_{T_i}$ but not $\mathcal{F}_{T_i-}$ and finally, X_i is the only information in $\mathcal{F}_{T_i}$ not contained in $\mathcal{F}_{T_i-}$.

Once we understand the term 'strict past' and 'filtration', we can re-define the compensator in a more rigorous way.

Definition Assume that $\mathbb{E}(N_t) < \infty$ for every t, then the compensator of a standard point process (absent of marks) N respect a whole history $\mathcal{H}$ such that $\mathcal{F}_t \in \mathcal{H}_t$ for each t is the unique random measure Λ on $\mathbb{R}_+$ such that

- The process (Λ_t) is $\mathcal{H}$-predictable
- For every non-negative $\mathcal{H}$-predictable process C:

$$\mathbb{E}[\int_0^\infty C dN] = \mathbb{E}[\int_0^\infty C d\Lambda]$$

A similar definition can be applied to the marked point process after adding some appropriated mark spaces.

Using the Doob-Meyer decomposition, it can be shown that $\Lambda(t) = \mathbb{E}[N(t)|\mathcal{H}_{t-}]$, if the compensator Λ is continuous, we call $\lambda_t = \Lambda(dt)$ the intensity of the counting process.

C Janossy representation of generic probabilistic structure of marked point process

Definition The Janossy measures are non-probability measures for point process N and are defined as the measures satisfying,

$$J_n(A_1\times, \cdots, \times A_n) = p_n \sum_{perm} \Pi_n(A_{i_1}\times, \cdots, \times A_{i_n}) \tag{38}$$

For marked point process $\bar{N}$ they are,

$$J_n(A_1^{'}\times, \cdots, \times A_n^{'}) = p_n \sum_{perm} \Pi_n(A_{i_1}^{'}\times, \cdots, \times A_{i_n}^{'}) \tag{39}$$

Where $A_i^{'} = A_i \times B_i$, B_i are the mark spaces.

The Janossy measure has a nice interpretation, if $E = R$ and $j_n(t_1, \cdots, t_n)$ denotes the density of $J_n(\cdot)$ with respect to a Lebesgue measure on $(R)^n$ with $t_i \neq t_j$ if $i \neq j$, then $j_n(t_1, \cdots, t_n)dt_1 \cdots dt_n = Pr\{$ there are exactly n points in the process one in each of the n distinct infinitesimal regions $(t_i, t_i + dt_i)$ $\}$. Its marked counterpart density is defined as

$$j_n(t_1, \cdots, t_n, x_1, \cdots, x_n)dt_1 \cdots dt_n dx_1 \cdots dx_n \tag{40}$$

with interpretation as $Pr\{$ points around $\{t_1, \cdots, t_n\}$ with marks around $(x_1, \cdots, x_n)$ $\}$. These interpretations, are in fact indicating that the Janossy density is nothing but the likelihood of a point process.

Definition The likelihood of a realization of a point process N on a bounded Borel set $E \subseteq R^d$, when $n = N(E)$, is the local Janossy density

$$L_E(t_1, \cdots, t_n) = j_n(t_1, \cdots, t_n | E) \tag{41}$$

In the formation of Janossy measure, the condition $\sum p_n = 1$ can take the form of

$$\sum_{n=0}^{\infty} (n!)^{-1} J_n(E^{(n)}) = 1$$

It turns out that the Janossy measures can uniquely determine the point process, one can find the proof as well as a more detailed introduction of Janossy measure in Chapter 5,7 of Daley et al (2007)Daley and Vere-Jones (2007)

We now use the Janossy measure to describe the probability of the marked point process. Let $S_n(t|\mathcal{F}_{t-}) = 1 - \int_{t_{n-1}}^{t} p_n(u|(t_1, x_1). \cdots, (t_{n-1}, x_{n-1}))du$ be the conditional survival function. The probability structure of the marked point process can be viewed as

$$\begin{aligned} J_0(T) &= S_1(T), \\ j_1(t_1, x_1|T) &= p_1(t_1) f_1(x_1|t_1) S_2(T|(t_1, x_1)) \text{ as } (0 < t_1 < T) \\ j_2(t_1, t_2, x_1, x_2|T) &= p_1(t_1) f_1(x_1|t_1) p_2(t_2|(t_1, x_1)) f_2(x_2|(t_1, x_1), t_2) S_3(T|(t_1, t_2), (x_1, x_2)) \\ &\quad \text{as } (0 < t_1 < t_2 < T) \end{aligned} \tag{42}$$

where T is the endpoint (length of the time interval), p_i are the densities, suitably conditioned, for the locations in the ground process, and the $f_i(\cdot)$ refer to the densities, again suitably conditioned, for the marks.

The interpretation is just like the one mentioned at very beginning of this subsection, take $j_1(t_1, x_1|T)$ as an example, it says the likelihood of there is one point locates in

the time interval $[0, T]$ with mark value x_1 is equal to the probability of only one event happened times the density of the mark of being x_1 conditional on such event happened at time t_1, also we need to multiple the survival function conditional on the first event information.

We now make a critical shift of view. Instead of specifying the conditional density p_n, we express them in terms of conditional hazard functions,

$$h_n(t|t_1, \cdots, t_{n-1}) = \frac{p_n(t|t_1, \cdots, t_{n-1})}{S_n(t|t_1, \cdots, t_{n-1})} \tag{43}$$

Equivalently, $p_n(t|t_1, \cdots, t_{n-1}) = h_n(t|t_1, \cdots, t_{n-1}) \exp(-\int_{t_{n-1}}^{t} h_n(u|t_1, \cdots, t_{n-1})du)$. Plug it into Equation 42, and recall from Equation **??** the likelihood of a counting process is

$$L(\cdot) = \Pi_i \lambda_{t_i}(B) \exp\Big(- \int \lambda_u(B) du \Big)$$

we have,

$$\begin{aligned} j_n(t_1, \cdots, t_n, x_1, \cdots, x_n|T) &= h_1(t_1) \cdots h_n(t_n|t_1, \cdots, t_{n-1}, x_1, \cdots, x_{n-1}) \times \\ & f_1(x_1|t_1) \cdots f_n(x_n|t_1, \cdots, t_n, x_1, \cdots, x_{n-1}) \times \\ & \exp\Big(- \int_0^{t_1} h_1(u) du \Big) \cdots \exp\Big(- \int_{t_n}^{T} h_{n+1}(T|t_1, \cdots, t_n, x_1, \cdots, x_n) \Big) \\ &= L(\cdot) \end{aligned} \tag{44}$$

Thus, the densities for the locations can be expressed in terms of corresponding hazard functions. And the conditioning in the hazard functions now can include the values of the preceding marks as well as the length of the current and preceding intervals.

$$\lambda(t, x|\mathcal{F}_{t-}) = \begin{cases} h_1(t) f_1(x|t) \\ \vdots \\ h_n(t|(t_1, x_1), \cdots, (t_{n-1}, x_{n-1}) \times f_n(x|(t_1, x_1), \cdots, (t_{n-1}, x_{n-1}, t), (t_{n-1} < t \leq t_n, n \geq 2) \\ \vdots \end{cases} \tag{45}$$

Where $\mathcal{F}_{t-}$ again is the internal history of the marked point process and $h_i(\cdot)$ are the corresponding hazard functions,suitably conditioned. Notice that an generalization from internal history to a whole history $\mathcal{H}_{t-}$ can be done quite easily, we just need to let the Janossy density to be conditional on the external history information. We can rewrite the above as

$$\lambda(t, x|\mathcal{F}_{t-}) = \lambda_g(t, x|\mathcal{F}_{t-}) f(x|t, \mathcal{F}_{t-}) \tag{46}$$

where λ_g is the ground intensity. When the mark space is continuous, we have $\lambda(t, x|\mathcal{F}_{t-})dtdx = \mathbb{E}[\bar{N}(dt \times dx)|\mathcal{F}_{t-}] = \lambda_g(t, x|\mathcal{F}_{t-}) f(x|t, \mathcal{F}_{t-})dtdx$.

D The Thinning Method for Simulation

The detailed thinning method steps can be summarised as:

1. Let τ be the start point of a small simulation interval
2. Take a small interval $(\tau, \tau + \delta)$
3. Calculate the maximum of $\lambda_g(t|\mathcal{F}_{t-})$ in the interval as

$$\lambda_{max} = \max_{t\in(\tau,\tau+\delta)} \lambda_g(t|\mathcal{F}_{t-})$$

4. Simulate an exponential random number ξ with rate λ_{max}
5. if

$$\frac{\lambda_g(\tau+\xi|\mathcal{F}_{t-})}{\lambda_{max}} < 1$$

   go to step 6.

   Else no events occurred in interval $(\tau, \tau + \delta)$, and set the start point at $\tau \leftarrow \tau + \delta$ and return to step 2
6. Simulate a uniform random number U on the interval $(0, 1)$
7. If

$$U \leq \frac{\lambda_g(\tau+\xi|\mathcal{F}_{t-})}{\lambda_{max}}$$

   then a new 'event' occurs at time $t_i = \tau + \xi$. Simulate the associated marks for this new event.
8. Increase $\tau \leftarrow \tau + \xi$ for the next event simulation
9. Return to step 2

E DBSCAN Cluster Analysis

The DBSCAN algorithm classified all points into three: core points, border points and noise points. We start by defining these points. For a set of points $X = \{x_1, x_2, \cdots, x_N\}$.

Definition ϵ neighbourhood of a point x, denoted by $N_\epsilon(x)$ is defined by $N_\epsilon(x) = \{y \in X : d(y, x) \leq \epsilon\}$. Where $d()$ is a metric.

Definition Density is defined as $\rho(x) = |N_\epsilon(x)|$, the number of points in a ϵ neighbourhood.

Definition Core point: let $x \in X$, if $\rho(x) \geq minPts$, then we call x a core point. The set of all core points is denoted as X_c, let $X_{nc} = X \setminus X_c$ be the set of all non-core points.

Definition Border point: if $x \in X_{nc}$ and $\exists y \in X$ such that $y \in N_\epsilon(x) \cap X_c$, then x is called a border point. Let X_{bd} be the set of all border points.

Definition Noise point: let $X_{noise} = X \setminus (X_c \cup X_{bd})$, if $x \in X_{noise}$, then we call x is a noise point.

To define what is a cluster under the DBSCAN setting, we need a few more definitions about 'reachable'.

Definition Directly density-reachable: if $x \in X_c$ and $y \in N_\epsilon(x)$, we may say y is directly reachable from x.

Definition Density-reachable: let $x_1, x_2, \cdots, x_m \in X, m \geq 2$. If x_{i+1} is directly density-reachable from x_i, $i = 1, 2, \cdots, m-1$. We call x_m is density-reachable from x_1.

Definition Density-connected: a point x is density connected to a point y if there exists another point $z \in X$ such that both y and x are density-reachable from z.

Definition Cluster: a non-empty subset C of X is called cluster if it satisfies:

- (Maximality) $\forall x, y$: if $x \in C$ and y is density-reachable from x, then $y \in C$.
- (Connectivity) $\forall x, y \in C$: x is density-reachable to y.

For a detailed algorithm description, we refer to the original Ester et al.(1996)Ester et al. (1996) paper.