한국경영과학회지
제39권 제4호
2014년 11월
http://dx.doi.org/10.7737/JKORMS.2014.39.4.051

다사건 시계열 자료 분석을 위한 베이지안 기반의 통계적 접근의 응용

석준희[1] · 강영선[2†]
[1]고려대학교 공과대학 전기전자공학부, [2]서울시립대학교 경영대학 경영학부

A Bayesian Approach for the Analysis of Times to Multiple Events : An Application on Healthcare Data

Junhee Seok[1] · Yeong Seon Kang[2]
[1]School of Electrical Engineering, Korea University
[2]Department of Business Administration, University of Seoul

■ Abstract ■

Times to multiple events (TMEs) are a major data type in large-scale business and medical data. Despite its importance, the analysis of TME data has not been well studied because of the analysis difficulty from censoring of observation. To address this difficulty, we have developed a Bayesian-based multivariate survival analysis method, which can successfully estimate the joint probability density of survival times. In this work, we extended this method for the analysis of precedence, dependency and causality among multiple events. We applied this method to the electronic health records of 2,111 patients in a children's hospital in the US and the proposed analysis successfully shows the relation between times to two types of hospital visits for different medical issues. The overall result implies the usefulness of the multivariate survival analysis method in large-scale big data in a variety of areas including marketing, human resources, and e-commerce. Lastly, we suggest our future research directions based multivariate survival analysis method.

Keywords : Multivariate survival analysis, Times to multiple events, Bayesian method, Big data

논문접수일 : 2014년 08월 22일 논문게재확정일 : 2014년 10월 18일
논문수정일(1차 : 2014년 10월 14일)
† 교신저자 yskang2014@uos.ac.kr

1. 서 론

의학이나 사회과학에서는 연구대상의 시간에 따른 상태의 변화를 관찰한 시계열 자료를 분석하는 연구가 종종 있다. 이러한 시계열 자료는 관찰 기간 내에 모든 연구대상에 대한 결과가 존재하지 않고, 연구대상의 이탈이나 사건의 발생 시점 관측 불가 등의 이유로 불완전한 중도절단된 자료(censored data)로 남게 되는 경우가 많다. 생존분석(survival analysis)은 이러한 중도절단된 불완전한 자료의 분석을 위해 개발된 방법으로, 특정 사건(event)이 발생할 때까지의 시간을 분석하는데 사용되는 통계분석 방법론이다.

연구대상을 일정시간 관찰하여 특정 사건이 발생하기까지의 시간을 분석하는 생존분석은 오늘날 의학 분야에서 가장 활발하게 활용되고 있다. 예를 들어, 질병의 과정은 노출, 발병, 치유, 재발, 사망 등의 일련의 사건이 발생하는데, 이러한 사건이 발생할 때까지의 시간을 연구하는데 생존분석 기법이 활용된다. 그 외에도 인구통계학, 사회학, 경제학, 제조업 등에서 사용되고 있다. 사회학에서는 사건사분석(event history analysis), 경제학에서는 지속기간분석(duration analysis), 산업계에서는 신뢰도분석(reliability analysis)으로 부르기도 한다. 좀 더 포괄적인 의미로 어떠한 사건이 발생할 때까지의 시간을 분석하는 의미로 “사건까지의 시간분석(time-to-event analysis)”으로 부르기도 한다[2].

경영학에서도 생존분석은 다양하게 활용되고 있다. 고객의 가입, 구매, 유지, 재구매, 이탈과 같은 시간에 따른 다양한 사건을 분석하는 것은 기업의 전략 수립에 유용하게 사용될 수 있다. 특히, 통신, 항공, 금융서비스, 인터넷 상거래 등과 같은 산업에서는 고객의 가입이 매우 중요한 사건으로 분류된다. 또한 이러한 기업의 입장에서는 신규 고객의 가입을 유도해야 하며, 기존 고객의 이탈을 막아야 한다. 오늘날 기업들은 고객정보데이터베이스를 통해 방대한 양의 고객 정보를 보유하고 있다. 생존분석을 통해 고객의 가입, 유지, 이탈을 분석할 경우, 우리는 중도절단된 자료로 인해 분석의 어려움을 겪게 된다. 고객의 가입 시점이 나타나지 않았거나, 이탈 시점을 알 수 없거나, 중도에 연구대상이 사라져서 미완성으로 남는 경우 등을 포함한다. 이러한 고객 정보를 통계적 방법론을 바탕으로 한 빅데이터 분석을 적용한다면, 고객유지프로그램 및 마케팅 전략에 활용할 수 있을 것이다. 생존분석은 고객유지프로그램뿐만 아니라 다양한 경영학 분야에서 활용될 수 있는데, 신제품의 확산, 중소기업의 생존과 부도예측분석, 인사관리에서 직원의 이직율 예측 등에서도 이용될 수 있다. 국내외 여러 경영학 연구 분야에서 생존분석 기법은 활용되어 왔다[1, 3, 8, 11, 14, 17]. 예를 들어, Lariviere and Van den Poel[14]는 금융서비스 자료를 이용하여 왜 고객이 이탈하는지, 어떻게 회사가 고객의 이탈을 막을 수 있는지에 대한 해답을 찾고자 하였다. 금융상품의 특성이 고객의 이탈과 유지에 미치는 영향을 분석하고자 생존분석을 이용하여 고객의 이탈 시점을 추정하였다. DuWors, Jr. and Haines, Jr.[8]은 식료품 판매 스캔 자료를 통해 브랜드 충성도를 측정하였으며, 중도 절단된 자료를 이용하기 위하여 습관적 구매행동 패턴을 생존분석을 이용하여 추정하였다. 그들은 연구에서 브랜드충성도가 고정적이고 불변의 것이 아니라, 시간에 따라 다르게 나타날 수 있음을 보였다. 예를 들면, 한 가구에서 일정 기간 동안 한정된 상표 혹은 한정된 제품을 습관적으로 구매하고, 다음 시점에서는 다른 제품들을 시도해보면서 브랜드와 제품들에 대해 내적으로 실험하며 학습하는 기간(inter-experimentation time period)을 가지게 되며, 그 다음 시점에는 최종적으로 결정된 한정된 상표나 한정된 제품을 습관적으로 구매하게 되는 패턴을 밝혀냈다.

현재까지 대부분의 이러한 연구는 단일 특정 사건이 발생할 때까지의 시간(TSE : time to a single event)에 집중되었으나, 최근 급증하는 데이터의 종류와 양에 따라 여러 사건이 동시다발적으로 발생하는 경우(TME : times to multiple events)에 대한 분석과 활용에 대한 관심이 높아지고 있다.

특히, 하나의 사건(예를 들어, 인터넷 쇼핑몰에서 한 제품군의 구매)이 다른 하나의 사건(다른 제품군의 구매)에 미치는 영향 및 두 사건 사이에 걸리는 시간 등은 비즈니스의 영역에서 흥미로운 문제이다. 하지만 널리 사용되고 있는 TSE 분석법과는 달리 TME 데이터에 대한 일반적인 분석은 매우 어렵다[22]. 이러한 방법론의 부재는 다양한 사건 간의 상관관계를 이해하는데 큰 걸림돌이 되어왔다. 향후 빅데이터 분석에 있어서, 새롭게 대두된 TME 데이터에 대한 효과적인 분석방법을 제시하고 유효성을 보이는 것은 중요한 과제이다.

TME 데이터를 효과적으로 분석하기 위한 첫 번째 단계로, 본 연구진은 기존 연구[26]에서 베이지안(Bayesian) 기반의 다변량 생존분석(multivariate survival analysis) 방법을 개발하고 제안하였다. 이 방법은 전체 표본 공간을 균일한 확률분포를 갖는 작은 공간들의 트리(tree) 형태로 분할하고, 주어진 다변량 중도 절단 자료로부터 서로 다른 분할 방식에 대한 사후확률을 계산함으로써, TME 간의 확률 결합 분포를 예측한다. 본 논문에서는 위의 다변량 생존분석 방법을 응용하여 실제 보건 자료의 분석에 적용하였다. 서로 다른 사건들 간의 연관성, 특히 선행성(precedence), 의존성(dependency), 인과성(causality)에 대한 분석법을 실제 자료에 적용하여 그 유효성을 입증하였다. 미국의 한 소아병원을 방문한 2,111명의 유아 환자를 대상으로, 서로 다른 두 가지 질환으로 병원을 방문하게 되는 시점에 대해 분석하였다. 한 질환으로 인한 방문이 다른 질환으로 인한 방문에 선행할 확률 및 두 방문 간의 상관관계와 인과관계에 대해 분석하였다.

이러한 분석은 환자의 병원 방문 형태를 이해하고, 더 나아가 병원 방문을 예측할 수 있는 가능성을 제시하였다. 환자의 입장에서는 발생 가능성이 높은 질병에 대한 사전 예방에 주의를 기울일 수 있으며, 병원 운영자의 입장에서는 각 환자의 질병 및 건강 상태에 맞춘 사전 계획된 체계적인 의료 서비스를 제공할 수 있는 기반이 될 수 있을 것이다. 또한 국가의 보건 정책을 수립하는 데도 도움이 될 것이다.

이러한 중요성으로 인해, 최근 보건 의료 분야[12, 30]뿐만 아니라 헬스 마케팅 분야[18, 19]에서도 질병 사이의 상관관계를 찾는 것은 중요한 주제로 다루어지고 있다. 예를 들어, Hidalgo et al.[12]은 미국 메디케어 의료보험의 청구 자료에 나타난 환자의 질병 코드(ICD9 code)를 분석하여, 질병 사이의 상관관계를 나타내는 네트워크를 구성하였다. Hidalgo et al.은 이 질병 네트워크를 통해 질병에 대한 새로운 치료법을 찾아내고 합병증을 예방할 수 있을 것으로 예상하였다. 하지만 그들의 연구는 질병이 발생한 시점에 대한 정보를 포함하지 않았고, 그렇기 때문에 질병 사이의 인과성을 찾기에는 적합하지 않다. 그에 비해 본 논문의 연구는 질병으로 인한 병원 방문 시점을 중심으로 자료를 분석하였다. 다변량 생존분석을 통해 질병의 발생 시점에 대한 정보, 곧 질병으로 인한 병원 방문 시점을 분석함으로써, 질병 사이의 선행성, 의존성 및 인과성에 대한 분석을 수행하였다.

본 연구는 다음과 같이 구성된다. 제 2장에서는 생존분석에 대한 개념을 소개한다. 제 3장에서는 본 연구에 적용된 연구모형을 설명한다. 제 4장에서는 실제 자료에 적용시킨 실증분석에 대한 결과를 소개한다. 마지막으로 제 5장에서는 본 연구에서의 의의와 연구의 한계점을 짚어보고, 향후 연구방향을 제시한다.

2. 이론적 배경

2.1 일변량 생존분석(Univariate Survival Analysis)

TME 데이터 분석 방법을 설명하기에 앞서, 대표적인 TSE 분석 방법인 일변량 생존분석(univariate survival analysis)에 대해 간략히 소개한다. 대표적인 일변량 생존분석모형으로는 준모수적접근법(semi-parametric)인 Cox의 비례위험모형(proportional hazard model)과 비모수적접근법(non-

parametric)인 Kaplan-Meier 모형[13]을 들 수 있는데, 두 모형 모두 관심 있는 단일 특정 사건(one event)이 발생할 때까지의 기간을 추정하는 것이 가능하다. Cox의 비례위험모형은 생존에 영향을 미칠 것으로 예측되는 독립변수들의 모수를 추정하여 생존 기간을 예측할 수 있다. 이와 달리, 본 연구에서 사용하는 분석방법은 다른 설명 변수들이 존재하지 않고, 여러 사건들의 발생시점 데이터만으로 각각의 지속 기간에 대한 확률 분포를 측정한다는 점에서 Kaplan-Meier 모형(KM법)에 더 근접하다고 할 수 있다.

기본적인 원리를 설명하기 위해, 표본의 크기가 n개인 관찰 자료를 가정하자. 어떤 특정 사건에 대하여 관찰이 완벽히 이루어지는 경우, 사건 발생까지의 시간 $T^{(1)}, T^{(2)}, \cdots, T^{(n)}$이 수집될 수 있다. 하지만 중도절단 자료에서는 관찰이 $C^{(1)}, C^{(2)}, \cdots, C^{(n)}$의 시점에서 중단될 수 있다. 그래서 표본 i에 대하여 $T^{(i)}$대신 중도절단 혹은 사건 발생 시점 $X^{(i)} = \min(T^{(i)}, C^{(i)})$와 사건의 발생여부 $\Delta^{(i)} = I(T^{(i)} \le C^{(i)})$를 관찰하게 된다. 여기서 $I(\ )$는 표시함수(indicator function)이다. 결국, 불완전 자료인 X와 Δ로부터 완전한 T에 대한 확률 분포를 얻어내는 것이 생존분석의 목표이다. 특히, 관심 사건이 어떠한 시간 t 이후에 발생할 확률, 즉 생존함수 $S(t) = \Pr[T > t]$를 구하는 것을 목표로 한다. 생존함수란 특정의 시간을 넘겨 생존하는 확률 혹은 특정의 시간까지 사건이 발생하지 않는 확률을 일컬으며, 이러한 확률을 생존확률이라고도 부른다. 관찰을 시작하는 시점($t = 0$)에서는 아직 사건이 발생하지 않았으므로 생존함수의 값은 1이고, 시간이 무한대로 흐르면 모든 개체에 사건이 발생함으로 생존함수는 0으로 수렴한다. 그러므로 생존함수는 시간에 따라 지속적으로 감소하는 형태를 나타낸다.

KM법은 조건부 생존확률을 곱해가며 누적생존확률을 계산하고, 그에 따라 생존함수를 계산한다. 간단한 경우를 살펴보기 위해, 먼저 모든 자료가 완벽히 관찰되었고 어떤 두 사건도 동시에 일어나지 않는다고 가정하자. 이 경우 모든 표본에 대해 $X^{(i)} = T^{(i)}$이고 $\Delta^{(i)} = 1$이다. 일반성을 잃지 않고, 편의상 $X^{(1)} < X^{(2)} < \cdots < X^{(n)}$으로 가정하자. 이 경우, 시간 $X^{(i)}$를 넘어서 생존하고 있는 (혹은 사건이 아직 발생하지 않은) 표본은 모두 $n-i$개이고, 시간 $X^{(i)}$를 넘겨 생존할 생존확률 $S(X^{(i)}) = (n-i)/n$으로 추정될 수 있다. $X^{(i-1)}$의 시점까지 생존했다는 가정하에 $X^{(i)}$를 넘어서까지 생존할 조건부 생존확률을 $P^{(i)}$로 표시하면, $P^{(i)} = \frac{n-i}{n-i+1} = 1 - \frac{1}{n-i+1}$로 주어진다. 조건부 생존확률의 $\frac{1}{n-i+1}$항에서 분자 1은 $X^{(i)}$시점에서 사건이 발생한 표본의 수, $n-i+1$은 $X^{(i)}$를 포함한 시점에서 아직 사건이 발생하지 않은 표본의 수로 해석할 수 있다. $P^{(i)}$는 $X^{(i-1)}$에서 $X^{(i)}$까지의 구간에 대한 조건부 생존확률이므로, 시점 $X^{(i)}$를 넘겨 생존하는 확률은 $S(X^{(i)}) = \Pi_{j=1}^{i} P^{(j)} = (n-i)/n$으로 표현할 수 있다.

이 모형을 확장하여, 중도절단된 자료를 포함하고 한 시점에 복수의 사건이 발생할 수 있음을 가정하자. 시점 $X^{(i)}$직전에 생존하고 있는 총 개체 수를 $n^{(i)} = \Sigma_j I(X^{(j)} \ge X^{(i)})$라고 하고, $X^{(i)}$에서 발생한 사건의 수를 $d^{(i)} = \Sigma_j I(X^{(j)} = X^{(i)}) I(\Delta^{(j)} = 1)$라고 하자. $n^{(i)}$에 해당하는 개체의 집합을 위험군(risk set)으로 부른다. 이러한 상황에서, i번째 구간의 조건부 생존확률 $P^{(i)} = 1 - d^{(i)}/n^{(i)}$로 표현되고, 생존확률은 $S(X^{(i)}) = \Pi_{j=1}^{i} (1 - d^{(i)}/n^{(i)})$로 표시할 수 있다. 중도절단이 없는 경우에는 그 이전 시점에 사망이 발생하였을 때 살아남은 수가 위험군이 되고, 중도절단이 있는 경우에는 이전 시점에서 사망이 발생하였을 때 살아남은 수에서 구간에서 일어난 중도절단 건수를 뺀 것이 위험군이 된다. 결국, 임의의 시점 t에 대하여 KM법은 생존함수 $S(t) = \Pi_{i : X^{(i)} < t, \Delta^{(i)} = 1} (1 - d^{(i)}/n^{(i)})$로 계산한다.

이러한 비모수형 생존분석은 생존함수의 분포에 대한 아무런 가정 없이 주어진 자료에 대하여 관찰한대로 생존확률을 계산할 수 있다. KM법은 계산의 단순성과 비모수적 특성으로 인해, 생존분석이 필요한 각종 연구에 표준 분석 방법(gold-standard)으로 널리 사용되고 있다.

2.2 다변량 생존분석(Multivariate Survival Analysis)

단일 사건에 대한 TSE 데이터 분석을 위한 일변량 생존분석과는 달리, 두 개 이상의 사건에 대한 TME 데이터 분석을 위해서는 다변량 생존분석(multivariate survival analysis)이 필요하다. 다변량 생존분석에서는, 일반적으로 m개의 사건의 발생 시간 $\mathbf{T} = (T_1, T_2, \cdots, T_m)$에 대한 관측이 $\mathbf{C} = (C_1, C_2, \cdots, C_m)$의 시점에서 중단되게 된다. 그래서 사건 p에 대한 완벽한 발생 시간 T_p대신, 중도절단된 $X_p = \min(T_p, C_p)$와 $\Delta_p = I(T_p \le C_p)$를 관찰하게 된다. 결국, T대신에 $\mathbf{X} = (X_1, \cdots, X_m)$과 $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_m)$을 관찰하게 된다. 이들 중도절단 자료로부터 m차원 상의 생존함수,

$$S(t_1, t_2, \cdots, t_m) = \Pr[T_1 > t_1, T_2 > t_2, \cdots, T_m > t_m]$$

혹은 벡터 형태로 $S(\mathbf{t}) = \Pr[\mathbf{T} > \mathbf{t}]$를 구하는 것을 목표로 한다.

일변량 생존분석에 성공적으로 적용된 KM법을 다변량 생존분석에 확대 적용하기 위한 몇몇 시도가 있었다[6, 7, 23]. 이러한 시도는 주로 일반적인 다변량 생존분석이 아닌, 이변량(bivariate) 생존분석에 적용되었다. Campbell and Földes[6]는 먼저 이차원 상의 결합확률로 표현되는 생존확률을 연쇄법칙(chain rule)을 이용하여 조건부 일변량 확률로 분리하였다. 그리고 각각의 일변량 생존확률을 KM법을 이용하여 계산하고, 이를 다시 결합하여 이차원 상의 생존확률을 계산하였다. 그 외에도, Dabrowska [7]는 누적고장률함수(cumulated hazard function)를 연속적으로 결합하는 방법을 제시하였고, Prentice and Cai[23]는 두 개의 생존시간 사이의 관계를 먼저 예측하고 그를 바탕으로 생존확률을 계산하는 방법을 사용하였다.

이러한 기존의 다변량 생존분석 방법에서는 몇 가지 중요한 문제점이 발견되었다. 첫째, 이론적으로는 계산방식에 상관없이 동일한 최종결과가 나와야 함에도 불구하고, 계산방식과 순서에 따라서 결과가 달라진다. 둘째, 계산된 생존함수가 시간의 진행에 따라 지속적으로 감소하지 않는다. 그로 인해 때때로 확률이 음(negative)의 값으로 계산되기도 한다. 셋째, 이변량 생존분석에 집중되어 일반적인 다변량 생존분석으로의 확장이 어렵다. 결국, 비모수적인 KM법을 다변량 생존분석으로 확장하는 일은 매우 어렵고[22], 그렇기 때문에 새로운 방법의 개발이 필요하다. 이러한 어려움으로 인해 실제 자료에 대한 다변량 생존분석은 단순한 이변량의 경우조차 거의 이루어지지 않았다.

3. 연구모형

기존에 제시된 다변량 생존분석 방법들은 일반적으로 다차원 상의 생존함수를 직접 계산하고자 하였다. 하지만, 본 논문에서 사용되는 다변량 생존분석 방법은 생존확률 밀도 함수(survival probability density function)를 먼저 계산하고, 그 함수를 적절히 적분하여 생존함수를 계산한다. 생존확률 밀도 함수 $\delta(t_1, \cdots, t_m)$와 생존함수$S(t_1, \cdots, t_m)$와의 관계는 다음과 같이 주어진다.

$$\delta(t_1, \cdots, t_m) = \frac{\Pr[t_1 \le T_1 < t_1 + \Delta t_1, \cdots, t_m \le T_m < t_m + \Delta t_m]}{\Delta t_1 \Delta t_2 \cdots \Delta t_m}$$

$$S(t_1, \cdots, t_m) = \int_{t_1}^{\infty} \int_{t_2}^{\infty} \cdots \int_{t_m}^{\infty} \delta(\tau_1, \cdots, \tau_m) d\tau_1 d\tau_2 \cdots d\tau_m$$

본 논문에서 사용된 방법[26]은 다차원 공간에서 생존확률의 밀도 함수를 계산하기 위해, 베이지안 기반의 Optional Polya Tree(OPT)[15 16, 20, 21, 29]를 사용한다. 본 장에서는 OPT를 이용한 일반적인 결합 확률 밀도의 분포 도출 방법과 그를 이용한 다변량 생존분석 방법을 설명한다. 또한, 다변량 생존분석으로부터 사건 간의 선행성, 독립성, 인과성을 분석하는 방법에 대해 설명한다.

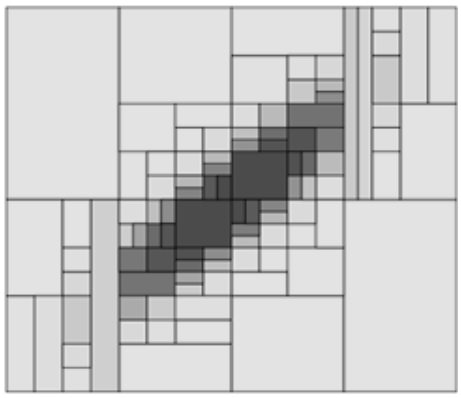
3.1 Optional Polya Tree를 이용한 확률 분포 도출

OPT(Optional Polya Tree)는 다변량 확률결합 분포의 확률밀도에 대한 측도를 계산하는 방법으로 제안되었다[29]. OPT는 기본적으로 전체의 표본공간을 균일한 밀도를 갖는 여러 개의 작은 공간으로 나눔으로써 확률 밀도의 분포를 계산한다. 공간의 분할 여부와 분할 방식은 확률에 의해 임의로 주어진다. 확률 밀도의 변화가 큰 영역일수록 높은 분할 확률을 갖게 되어 좀 더 세밀하고 정확하게 확률 밀도를 계산할 수 있다. [그림 1]은 OPT를 이용한 확률 밀도 예측의 예를 보여준다. 표본들은 서로 상관관계에 있는 결합 정규 분포에서 임의로 생성되었다. [그림 1]에서 나타나듯이 확률 밀도의 변화가 낮은 곳은 덜 분할 되었고, 변화가 큰 부분은 세밀하게 분할 되었다. OPT는 각 분할 방식과 분할 여부에 대한 확률로 정의되기 때문에, 확률 밀도에 대한 확률 분포로 해석된다.

[그림 2]는 구체적으로 이차원 상에서 OPT가 어떻게 적용되는지 보여준다. 표본 공간 상의 어떠한 영역 A는 데이터의 분포에 따라 A_{11}과 A_{12}, 혹은 A_{21}과 A_{22}로 분할될 수 있다. OPT는 먼저 A에 대하여 주어진 데이터 D에 대한 우도(likelihood) Φ(A|D)를 다음과 같이 계산한다.

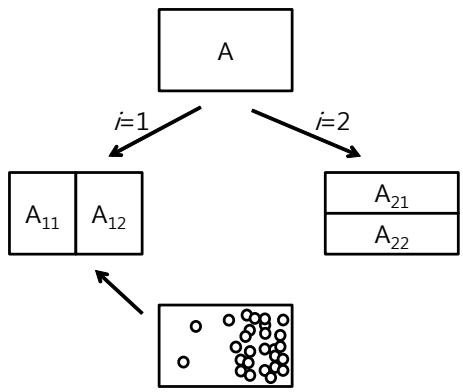
$$\Phi(A|D) = \rho\Phi_0(A|D) + (1-\rho)\sum_{i=1}^{2}\lambda_i \frac{B(n_{i1}+\alpha,\ n_{i2}+\alpha)}{B(\alpha,\ \alpha)} \Phi(A_{i1}|D)\Phi(A_{i2}|D)$$

위의 식에서 Φ_0(A|D)는 영역 A가 더 이상 분할하지 않았을 경우, 즉 A가 전체적으로 균일한 밀도를 갖고 있다고 가정했을 경우의 우도를 나타낸다. 두 번째 항의 경우는 A가 분할된 경우의 우도를 나타낸다. 세로($i=1$)와 가로($i=2$) 두 가지 방식의 분할이 가능하기 때문에, 각각의 경우에 대한 우도가 가중치 $\lambda_i \frac{B(n_{i1}+\alpha,\ n_{i2}+\alpha)}{B(\alpha,\ \alpha)}$로 합해진다. 여기에서 λ_i는 임의로 정할 수 있는 변수로 $\Sigma_i\lambda_i = 1$을 만족해야 하며, 보통 모든 분할 방식에 대해 동일하게 주어진다. $B(\ )$는 베타함수(beta function)이다. n_{ij}는 A_{ij}에 속해있는 표본의 수이고, α는 베이지안 분석에서 일반적으로 사용되는 허위 표본의 수(pseudo count)이다. A를 A_{i1}과 A_{i2}의 두 영역으로 나누는 경우, n_{i1}과 n_{i2}가 서로 비슷하다면(혹은 표본이 균일하게 분포되어 있다면), 이 가중치는 작게 계산된다. 반대로 n_{i1}과 n_{i2}가 크게 다르다면(혹은 표본이 균일하게 분포되어 있지 않다면), 이 가중치는 큰 값으로 계산된다.



각 분할된 영역은 균일한 확률 밀도를 갖는다. 짙은 색은 높은 확률 밀도를 연한 색은 낮은 확률 밀도를 나타낸다.

[그림 1] OPT를 이용한 확률 분포 측정의 예



영역 A는 A_{11}과 A_{12}, 혹은 A_{21}과 A_{22}로 분할 될 수 있다. 실제로 표본이 A에서 제일 아래의 그림과 같이 분포한 경우(작은 원으로 표시), A는 높은 확률로 A_{11}과 A_{12}로 분할된다.

[그림 2] OPT 계산을 위한 영역 분할의 예

Φ(A|D)는 A를 더 이상 분할하지 않았을 경우의 우도 Φ_0(A|D)와 분할했을 경우의 분할된 영역의 우도 Φ(A_{ij}|D)로부터 가중치 변수 ρ를 결합하여 계산된다. ρ는 0과 1사이에서 임의로 정할 수 있는 변수로, 보통은 동등한 가중치를 두기 위하여 0.5로 주어진다. 분할된 영역에 대한 우도 Φ(A_{ij}|D)는 비슷한 방식으로, 더 이상 분할이 의미 없을 때까지 회귀적(recursive)으로 계산된다. 분할이 의미 없다는 것은 특정 공간 내에 샘플이 하나만 존재하거나 없을 때를 의미한다. 이러한 방식으로 표본 공간 상의 모든 가능한 영역들에 대하여 우도를 계산할 수 있다.

OPT는 이렇게 계산된 우도로부터 역으로 확률 결합분포를 예측한다. 먼저 주어진 영역 A에 대하여 이 영역을 더 분할할지 말지를 ρ(A|D) = $\rho\Phi_0$(A|D)/Φ(A|D)로 주어지는 확률에 따라 결정한다. 분할 방식은 $\lambda_i(A|D) = \lambda_i \frac{B(n_{i1}+\alpha,\, n_{i2}+\alpha)}{B(\alpha,\, \alpha)}$에 비례하는 확률로 정해진다. 각 분할된 영역에 할당되는 확률은 $B(n_{i1}+\alpha,\, n_{i2}+\alpha)$에 따라 주어진다. 예를 들어, [그림 2]에서 주어진 영역에서 관찰된 표본은 균일하게 분포되어 있지 않기 때문에, 작은 ρ(A|D)를 갖게 된다. 또한, 가로로 분할(i = 2)하는 것보다 세로로 분할(i = 1)했을 때 분할된 영역들이 균일한 분포를 가질 가능성이 높기 때문에 λ_1(A|D) < λ_2(A|D)로 계산되어, 세로로 분할될 확률이 커진다. 세로로 분할하였을 경우, 각 분할된 영역에 할당되는 확률은 그 영역에 속해있는 표본의 수에 따라 주어진다.

[그림 2]는 이차원 공간에서 각 영역을 같은 크기의 두 개의 영역으로 분할하는 간단한 경우에 대한 예제이다. 하지만 일반적으로 OPT는 m차원의 공간에서 적용 가능하고, 각 영역을 임의의 방식으로 분할할 수 있다. 일반적으로 M개의 분할 방식이 있고, 각 분할방식을 통해 J_1, …, J_M개의 영역으로 나뉜다고 할 때, 우도 Φ(A|D)를 다음과 같이 계산된다.

$$\Phi(A|D) = \rho\Phi_0(A|D) + (1-\rho)\sum_{i=1}^{M}\lambda_i \frac{D(n_{i1}+\alpha, \cdots, n_{iJ_i}+\alpha)}{D(\alpha, \cdots, \alpha)} \prod_{j=1}^{J_i}\Phi(A_{ij}|D)$$

D( )는 베타함수의 일반적인 형태인 Dirichlet 함수로, $D(x_1, \cdots, x_n) = \Pi_i \Gamma(x_i)/\Gamma(\Sigma_i x_i)$로 주어진다.

OPT에서 분할을 결정하는 모든 확률은 분할된 영역에 속해있는 표본의 수, 즉 [그림 2]의 예제의 경우 n_{11}, n_{12}, n_{21}, n_{22}에 따라 계산된다. 어떤 특정 영역 A에 속해있는 표본의 수를 n(A)로 표시하자. 모든 분할 가능한 영역에 대한 n(A)를 모아놓은 벡터를 $\mathbf{n}$으로 표시하자. 이 때, $\boldsymbol{\rho}(\mathbf{n}|D)$와 $\boldsymbol{\lambda}(\mathbf{n}|D)$는 모든 분할 가능한 영역에 대한 ρ(A|D)와 λ_i(A|D)를 모아놓은 벡터로 표시할 수 있다. 결국, OPT는 주어진 데이터 D에 따른 사후 확률들의 변수 $\boldsymbol{\theta}(\mathbf{n}|D)$ = ($\boldsymbol{\rho}(\mathbf{n}|D)$, $\boldsymbol{\lambda}(\mathbf{n}|D)$)에 의해 정해진다. 이후 본 논문에서는 이러한 OPT를 따라 주어지는 확률 분포에 대한 확률 측도를 OPT($\boldsymbol{\theta}(\mathbf{n}|D)$)로 표시한다.

3.2 TME간의 확률 결합 분포 도출

TME간의 확률 결합 분포는 Seok et al.[26]이 제시한 방법에 따라 OPT를 적용하여 계산할 수 있다. 비중도절단자료의 경우 사건의 확정발생시간인 $\mathbf{T}$가 주어지는데 비해, 중도절단자료는 사건의 관찰중단시간 $\mathbf{X}$와 발생여부 Δ가 주어진다. OPT를 중도절단자료에 적용하기 위해서 먼저 표본 공간이 전사(projection)된 가상의 관심 표본 공간(ROI : region of interest)을 상정한다. 비중도절단자료의 경우 모든 표본의 자료가 확정적이기 때문에, 표본 공간을 한정 지을 수 있는 경계를 설정하는 것이 가능하다. 그 경계 내에서만 공간을 분할하여 OPT를 적용하고 경계 바깥은 확률 밀도를 0으로 가정할 수 있다. 하지만 중도절단자료의 경우 가장 마지막으로 관측된 지점 바깥에서는 총 확률의 합만 계산할 수 있을 뿐 그 분포를 계산할 수 없기 때문에, 표본 공간을 한정 짓는 것이 불가능하다. 그렇기 때문에 우리는 먼저 가장 바깥 지점에서 관측된 표본을 바탕으로 가상의 ROI를 지정하고, ROI 바깥의 모든 확률은 ROI의 경계에 전사되었다고 가정한다. 그럼으로써 ROI 바깥의 확률 밀도를 0으로 가정하여, OPT를 ROI에 대해서만 적용할 수 있다.

앞서 설명 되었듯이 OPT는 ROI 내의 특정 영역 A에 속해있는 표본의 수, 즉 n(A)를 기반으로 하여 계산된다. 비중도절단자료와 달리 중도절단자료에서는 어떤 사건이 관측된 시점보다 나중에 일어난다는 것만을 알 수 있기 때문에 n(A)를 확정적으로 계산할 수 없다. 주어진 자료 D = ($\mathbf{X}$, $\boldsymbol{\Delta}$)에 대해 만일 $\mathbf{T}$가 어떠한 확률분포 Q|D를 따른다면, n(A)의 예상값은 $\hat{n}(A|Q, D) = N\int_A Q|D dA$로 계산될 수 있다. 여기서 N은 총 표본의 수이다. Q|D를 가정하였을 경우, 모든 분할 가능한 영역에 대한 $\hat{n}(A|Q, D)$를 포함하는 벡터 $\hat{\boldsymbol{n}}$로부터, 앞서 설명된 방식에 따라 OPT($\boldsymbol{\theta}(\hat{\mathbf{n}}|$Q, D))가 계산될 수 있다. OPT($\boldsymbol{\theta}(\hat{\boldsymbol{n}}|Q, D)$)는 $\mathbf{T}$에 대한 확률결합분포의 확률분포를 나타내고, Q는 $\mathbf{T}$에 대한 확률결합분포를 나타낸다. 중도절단자료의 확률분포는 자기일관성(self-consistency)[9]을 가질 것으로 예상되기 때문에, 결국 다음과 같이 표현될 수 있다.

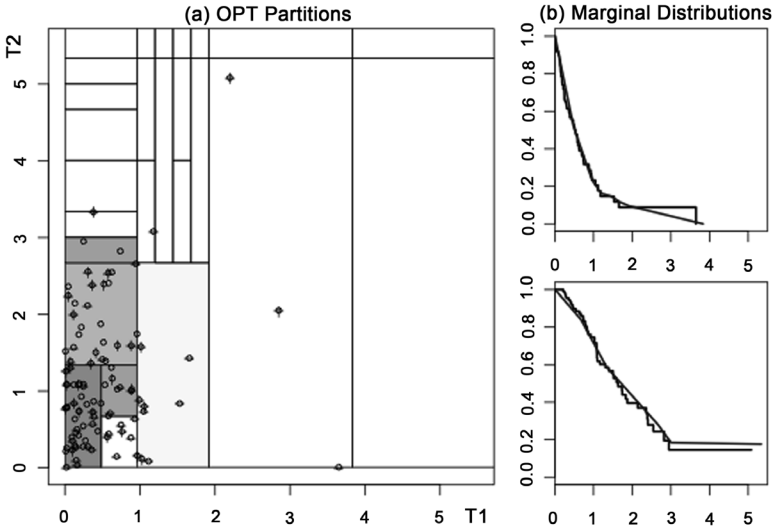
$$Q = E[OPT(\boldsymbol{\theta}(\hat{\mathbf{n}}|Q, D))]$$

이 방정식을 Q에 관하여 풀어, 최종적인 $\mathbf{T}$에 대한 확률결합분포 Q를 얻어낸다. 이 방정식은 해석적 해(analytic solution)가 존재하지 않기 때문에, 반복계산(iteration)을 통하여 수치적 해(numerical solution)를 찾는다.

위의 방정식에 대한 수치적 해를 찾기 위한 알고리즘은 Seok et al.[26]에 제시되었다. 간단히 정리하여, 먼저 어떤 확률 분포 $Q^{(k)}$를 가정하고 주어진 데이터 D를 고려하여 이에 해당하는 $\hat{\boldsymbol{n}}^{(k)}$를 예측한다. $\hat{\boldsymbol{n}}^{(k)}$는 각각의 분할된 영역에 속한 표본의 수에 대한 예측값이다. 이로부터 앞서 설명한 방식에 따라 OPT($\boldsymbol{\theta}(\hat{\mathbf{n}}^{(k)}|Q^{(k)}, D)$)를 구축한다. OPT는 표본 공간의 확률결합분포에 대한 확률분포를 나타내기 때문에, 이에 대한 기대값(Expectation)을 몬테카를로(Monte-Carlo) 방식을 이용하여 수치적으로 계산할 수 있다. 혹은 계산의 편리성을 위하여 MAP (Maximum a Posterior) 방식을 이용하여 근사적으로 계산할 수도 있다. 이렇게 계산된 기대값은 $Q^{(k+1)}$이 되어, 같은 방식으로 $Q^{(k+2)}$를 계산하기 위해 사용된다. Q가 수렴할 때까지 위의 연산을 반복함으로써, 위의 주어진 방정식을 푼다. 초기값인 $Q^{(0)}$를 위해서는 균일분포(uniform distribution)가 사용될 수 있지만, 계산의 효율성을 높이기 위해 일변량 분석을 통해 얻어진 근사적 분포도 초기값으로 사용될 수 있다. 이 알고리즘은 초기값이 상관없이 최종적으로 같은 분포로 수렴한다는 것이 시뮬레이션 연구를 통해 확인되었다.

일변량 TSE 분석에서 많이 사용되는 KM법 또한 자기일관성을 기초로 한다[28]. OPT는 KM법과 같이 $\mathbf{T}$의 확률분포에 대해 어떠한 가정도 하지 않기 때문에 비모수적 방법이다. 그렇기 때문에 OPT의 결합확률로부터 계산된 단일 변량에 대한 주변 확률분포는 KM법으로 계산된 것과 거의 일치한다[26]. 이러한 이유로 인해 이 방식은 KM법의 일변량 분석을 다변량 분석으로 확장시킨 것으로 간주될 수 있다. Seok et al.[26]은 다양한 시뮬레이션 분석을 통하여 TME간의 확률 결합분포가 OPT를 이용하여 정확하고 안정적으로 계산될 수 있음을 보였다.

[그림 3]은 두 가지 사건의 발생 시간에 대한 결합 분포를 OPT를 이용하여 계산한 예를 보여준다. U_1과 U_2는 각각 Exp(1)과 Exp(1/2)의 분포에 따라 임의로 생성되었고, $T_1 = U_1$, $T_2 = T_1+U_2$로 주어졌다. 여기서 Exp(λ)은 기대값 $1/\lambda$을 갖는 지수분포를 나타낸다. C_1과 C_2는 각각 독립적으로 Exp(1)에 따라 임의로 생성되었다. 이 경우 확정 사건 발생 시간 T_1과 T_2 대신, 관측 중단 시간 $X_1 = \min(T_1, C_1)$과 $X_2 = \min(T_2, C_2)$, 그리고 사건 발생 여부 $\Delta_1 = I(T_1 \le C_1)$와 $\Delta_2 = I(T_2 \le C_2)$가 주어진다. 여기서부터 T_1과 T_2 간의 확률 결합 분포가 계산되었다. [그림 3]은 OPT에서 얻어진 최대 사후 확률(MAP : maximum a posteriori)을 갖는 확률분포를 보여준다. 각각의 분할된 영역 내에서는 표본이 균일하게 분포된다고 생각되고, 높은 확률 밀도를 갖는 영역일수록 진하게 표시되어 있다. 예상과 같이 T_2가 T_1보다 큰 영역에서 높은 확률 밀도를 갖는다. 결합분포로부터 계산된 T_1과 T_2에 대한



(a) T_1과 T_2에 대한 예측된 확률 결합 분포. 각 표본의 관찰된 자료는 원으로 표시되었다. x축 방향으로 중도 절단 되었을 경우 세로 선으로 표시되고, y축 방향으로 중도절단 되었을 경우 가로 선으로 겹쳐 표시되었다. 각 분할된 공간은 균일한 확률 분포를 갖고, 높은 확률 밀도일수록 진한 색으로 표시되었다. (b) T_1과 T_2에 대한 주변 확률 분포. 직선으로 표시된 것은 결합 확률 분포로부터 계산된 주변 확률 분포이고, 계단형으로 표시된 것은 KM법에 의해 계산된 확률 분포이다.

[그림 3] OPT를 이용한 TME 분석의 예

주변 확률(marginal probability) 분포(직선으로 표시)는 KM법으로 계산된 확률 분포(계단형으로 표시)와 비교하여 거의 근접하다는 것을 알 수 있다.

3.3 두 사건 간의 선행성, 의존성, 인과성 분석

일반적으로 다수의 사건에 적용가능하나, 본 논문에서는 간단한 경우인 두 사건에 대해서만 분석을 실행하고, 응용 사례를 보였다. 사건 A와 B에 대하여 T_A와 T_B를 각 사건의 발생 시간이라고 할 때, 우리는 앞서 설명된 방법을 통하여 자료로부터 결합 확률 밀도 분포(joint probability density distribution) $\Pr[T_A, T_B]$를 계산할 수 있다. 얻어진 결합 확률 밀도 분포를 바탕으로, 두 사건간의 선행성, 의존성, 인과성에 대한 분석 방법을 설명한다.

3.3.1 선행성(Precedence)

한 사건이 다른 사건에 선행하는지를 정량화하여 제시한다. 이는 선행 확률, 즉 $\Pr[T_A > T_B]$ 혹은 $\Pr[T_A < T_B]$로 나타낼 수 있다. 전자는 사건 B가 사건 A에 선행할 확률을 나타내고, 후자는 그 반대의 경우이다. 이 확률은 결합 확률 밀도 분포를 적절한 영역 상에서 적분함으로써 쉽게 구할 수 있다. 앞서 설명한 바와 같이, OPT를 이용한 결합 확률 밀도 분포는 ROI 바깥에서는 정의되지 않는다. 그렇기 때문에, T_A와 T_B가 모두 ROI 밖에 위치하는 경우를 제외하고 선행 확률을 계산한다. 선행 확률은 결합확률분포로부터 구해지기 때문에, 이를 예측하는 다변량 생존분석이 선행성 분석에는 필수적이다. 만일 다변량 생존분석을 사용하지 않는다면, 명백히 존재하는 선행성을 놓치거나 혹은 과다하게 예상할 우려가 있다.

3.3.2 의존성(Dependency)

두 사건이 서로 독립적인지 아니면 상호 연관이 있는지는 정량화하여 제시한다. 이는 사건 발생 시간의 분포, 즉 T_A와 T_B가 독립적인지를 측정함으

로써 알 수 있다. 두 확률 변수 간의 독립성은 상호정보량(mutual information)으로 측정될 수 있다. X와 Y간의 상호정보량은 $\mathrm{MI}(X, Y) = \sum p(x, y)\log(p(x, y)/p(x)p(y))$로 계산된다. 상호정보량은 비모수적방법으로 두 확률 변수 사이의 관계를 측정하는 일반적인 방법이다. 완벽히 독립적인 두 변수에 대해서는 상호정보량이 0으로 주어지고, 서로 의존적일수록 큰 상호정보량을 갖는다.

본 논문에서 $\mathrm{MI}(T_\mathrm{A}, T_\mathrm{B})$는 TME 결합 확률 분포 계산에서 얻어진 분할된 영역을 바탕으로 계산된다. 각 분할된 영역은 균일한 확률 분포를 갖고 있기 때문에, 더 분할한다고 하여도 상호정보량은 변하지 않는다. 이 점을 이용하여 손쉽게 상호정보량을 계산할 수 있다. 각 분할로부터 얻어진 상호정보량은 해당 분할에 대한 확률값만큼 가중치가 주어져 최종 상호정보량으로 합산된다. 이렇게 얻어진 $\mathrm{MI}(T_\mathrm{A}, T_\mathrm{B})$는 임의 교차를 통해 얻어진 상호정보량과 비교된다. 하나의 표본 i에 대하여 $(X_A^{(i)}, \Delta_A^{(i)})$와 $(X_B^{(i)}, \Delta_B^{(i)})$의 쌍이 관찰된다. 여기에서 사건 A에 대한 자료와 B에 대한 자료를 임의로 교차시켜 $(X_A^{(i)}, \Delta_A^{(i)})$와 $(X_B^{(j)}, \Delta_B^{(j)})$의 쌍을 만든다. 이렇게 생성된 임의의 자료는 가상의 T_A와 T_B를 나타내는 자료로 생각될 수 있다. 이 경우 임의로 교차시켰기 때문에 가상의 T_A와 T_B는 서로 독립이다. 결국, 임의 교차 자료로부터 계산된 상호정보량은 T_A와 T_B가 독립일 경우에 대한 귀무분포(null distribution)를 실증적(empirical)으로 얻어내기 위해 사용될 수 있다. 이러한 임의 교차를 반복하여, 귀무분포에 대한 평균과 표준편차를 계산하고 이를 표준분포에 근사시킴으로써, 실제 자료로부터 얻어진 상호정보량이 귀무분포에서 얻어졌을 확률을 계산할 수 있다. 이 확률은 실측된 상호정보량이 귀무분포에서 얻어진 것이 아니라는 확률적 검증에 대한 p값으로 사용된다.

결합확률분포를 예측하는 다변량 생존분석을 사용하지 않고, 직접적으로 두 생존시간 사이의 의존성 혹은 독립성을 측정하는 몇몇 연구가 존재한다[24, 25]. 하지만 이들 방법은 대부분 두 생존시간 사이의 선형 관계를 가정하여 상관계수를 구하는 등의 방식으로 의존성을 측정하였다. 하지만 본 논문에서는 일반적인 경우에 의존성을 측정하는 상호정보량을 사용하였다. 상호정보량은 선형관계에 대한 가정을 필요로 하지 않기 때문에, 더 많은 형태의 의존성을 검출해 낼 수 있다는 장점이 있다. 그렇기 때문에 본 논문에서 제시하는 의존성 검출은 방법은 다양한 영역의 연구에 적용이 가능할 것으로 예상된다. 상호정보량의 계산에는 두 생존시간 사이의 결합확률분포가 필요하고, 따라서 다변량 생존분석이 필수적이다.

3.3.3 인과성(Causality)

인과성은 다양한 방법으로 정의될 수 있지만, 여기에서는 Granger 인과성을 기초로 정의한다[10]. Granger 인과성은 한 변수가 다른 변수를 예측하는데 도움이 된다면 인과성이 존재한다고 정의한다. 본 논문에서는 이를 좀 더 엄격한 의미로 사용하여, (1)한 사건이 다른 사건에 선행하고, (2)선행 사건과 후행 사건 발생 시간의 차이가 선행사건 발생 시간에 독립적이라면, 선행 사건이 후행 사건에 대해 인과성을 갖는다고 정의한다. 첫 번째 조건은 앞서 설명한 선행성 분석을 통해 확인한다. 두 번째 조건은 후행 사건의 발생 시간을 선행 사건의 발생 시간에 따른 선형 함수로 모델링함으로서 확인할 수 있다. 두 사건 A와 B에 대해 A가 B에 선행하는 경우, 즉 $T_\mathrm{A} < T_\mathrm{B}$인 경우, 우리는 $T_B = \alpha T_A + M$으로 모델링한다. 이 선형 모델에서 T_B는 T_A에 비례하는 부분 αT_A와, T_A에 독립적인 확률 변수 M으로 분리된다. 이 모델은 $T_B - T_A = (\alpha - 1)T_A + M$와 동치이다. M은 T_A와 독립적으로 구해지기 때문에, α가 1에 가까울수록 T_A와 T_B-T_A가 서로 독립적이라고 해석할 수 있다. 결국, 인과성 계수 α를 통해 두 사건간의 인과성 정도를 나타낼 수 있다. 위의 선형 관계에서 T_B를 완벽하게 독립 관계에 있는 두 확률 변수로 나누는 것은 쉽지 않다. 대신 완화된 독립의 조건인 공분산(covariance) 0을 만족하는 두 변수로 나눌 수 있다. 이 경우 $\alpha = \mathrm{Cov}(T_\mathrm{A}, T_\mathrm{B})/\mathrm{Var}(T_\mathrm{A})$로 쉽게 계산되고, $\mathrm{Cov}(T_\mathrm{A}, M) = 0$을 만족한다.

완벽한 인과성을 나타내는 1과 실제 계산된 α의

차이는 숨겨진 사건 H의 사건 B에 대한 영향력을 나타낸다고 해석할 수 있다. 사건 A에 대해 완벽한 인과성을 갖지만 실제로는 관측되지 않는 숨겨진 사건 H를 가정하였을 경우, $T_A = T_H + A$로 표현 가능하다. 또한 일반적으로 $T_B = (1+b)T_H + A + B$로 표현 가능하다. 여기서 b는 T_H가 T_B에 추가적으로 미치는 영향력을 나타내고, B는 A와 T_H에 독립적인 확률변수로 사건 B의 발생에 대한 추가적인 시간을 나타낸다. 이러한 모델의 경우 $\alpha = 1 + bVar(T_H)/Var(T_A)$로 주어진다. 결국 1과 α의 차이는 H의 B에 대한 추가적 영향력을 나타내는 b에 따라 정해진다. 만일 b가 0이라면 H가 B에 미치는 영향은 오로지 A를 통해서만 나타나기 때문에, 실제로 관측되는 A가 B에 대해 완벽한 인과성을 갖는다고 해석할 수 있다. 하지만 b가 0이 아니라면 숨겨진 사건 H는 실제로 관측되는 A를 통해서뿐만 아니라 관측되지 않는 다른 방식으로도 B에 영향을 미치기 때문에, A가 B에 대해 완벽한 인과성을 갖는다고 말할 수 없고, 대신 약한 인과성을 갖는다고 말할 수 있다. 특히, α가 0인 경우는 T_A와 T_B가 서로 독립이기 때문에 완벽한 비인과성을 갖는다.

사건 사이의 인과 관계를 분석하기 위해 베이지안 네트워크(Bayesian Network)가 많이 사용되어 왔다. 베이지안 네크워크는 기본적으로 사건의 발생 여부에 따라 사건들 간의 의존성을 분석하여 인과 관계로 해석될 수 있는 네트워크를 구성한다. 하지만 베이지안 네트워크의 특성상 시간 정보를 포함하기가 힘들고, 관측되지 않은 사건(중도절단자료)을 다루기가 쉽지 않다. 최근 중도절단자료를 이용한 베이지안 네크워크에 대한 연구가 진행 중에 있으나[5, 27], 대부분 일변량분석(하나의 사건만 중도절단 됨)에 국한되어 있어 중도절단 된 사건 간의 경우(두 사건 모두 중도절단 됨) 인과성을 분석할 수 없다. 본 논문에서 제시하는 인과성 분석은 시간정보를 포함시킴으로써, 중도절단된 사건 간의 인과성에 대한 분석을 가능하게 한다. 이는 또한 두 사건 발생 시간 사이의 결합확률분포를 기초로 하기 때문에, 다변량 생존분석이 필수적이다.

4. 실증 분석

4.1 자료설명

본 논문에서는 2,111명의 유아에 대하여 두 가지 서로 다른 호흡기 관련 질환으로 병원을 방문하는 사건에 대하여 분석하였다. 이들은 유아에 대한 자료는 미국 캘리포니아 주 한 어린이 병원으로부터 얻어졌다. 출생 후의 모든 진찰과 진료 및 수술 기록은 의무전자기록(EHR : Electronic Health Record)으로 보관되어 있는데, 이 EHR을 바탕으로 우리는 다음의 조건을 만족하는 2,111명의 유아를 선별하였다. 이 조건은 각 유아에 대한 최대한 완벽한 의무기록을 확보하기 위하여 제시되었다.

(1) 2006년에서 2008년 사이 병원에서 태어났다.
(2) 출생 후 최소 1년간의 의무기록이 남아있다.
(3) 최소 5번 이상의 진료 기록을 갖고 있다.
(4) 추적 가능한 시간까지 사망하지 않았다.

이들 유아에 대한 자료로부터 호흡기 관련 질환의 종류와 병원 방문 시점이 추출되었다. 호흡기 관련 질환은 의무기록의 ICD9(International Classification of Diseases, World Health Organization's 9th Revision) 코드를 이용하여 구별되었다. ICD9 코드는 5자리로 구성되어 있는데, 여기서는 3자리의 대분류를 이용하여 2,111명의 유아에게서 가장 많이 나타나는 두 가지 호흡기 관련 질환에 대해 분석하였다.

(1) Symptoms : ICD9 코드 780-789
(2) Upper Respiratory Tract Diseases(URTD) : ICD9 코드 470-478

이러한 질환으로 인한 병원의 최초 방문 시점은 출생 후 날짜(Days Since Birth)로 조사되었다. 이들 사건의 발생은 유아의 부모가 더 이상 이 병원을 이용하지 않았거나 혹은 4년간의 조사기간 동안 해당 질병으로 인한 방문이 없었기 때문에 중도절

단자료로 남게 되었다. Symptoms의 경우 27%의 표본에서 중도 절단이 발생하였고, URTD의 경우 82%의 표본에서 중도 절단이 발생하였다. 사건의 발생이 관측된 표본에 한정하였을 때, Symptoms으로 병원을 최초 방문한 시점의 중간값은 생후 240일째이고, URTD의 경우 중간값 641일째 최초 방문이 이루어졌다. 연구 대상 표본에 대한 기본적인 통계는 <표 1>에 주어져 있다.

〈표 1〉 분석 대상 환자군 분포

성별	남성(1,127, 53.3%)	여성(984, 46.7%)
인종	Asian(151, 7.2%) Black(59, 2.8%) White(712, 33.7%) Others(205, 9.7%)	Asian(131, 6.2%) Black(45, 2.1%) White(655, 31.0%) Others(205, 9.7%)
민족	Hispanic(608, 28.8%) Non-Hispanic(519, 24.6%)	Hispanic(557, 26.4%) Non-Hispanic(427, 20.2%)

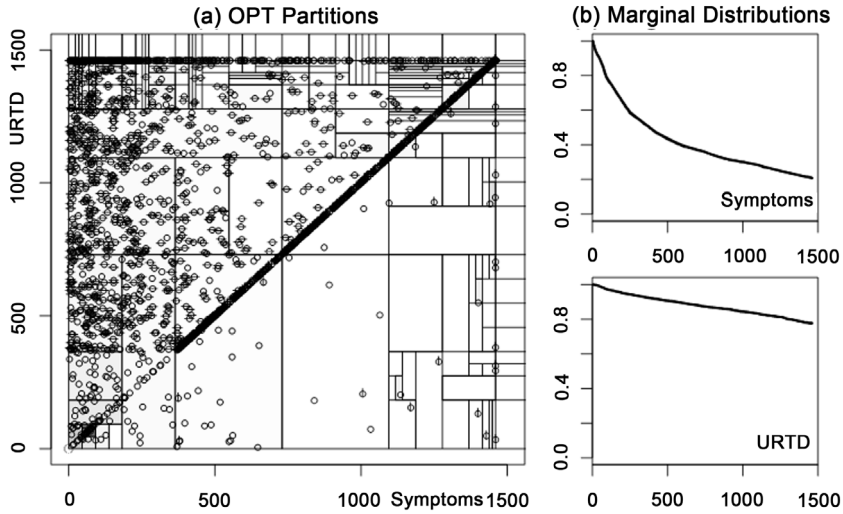
4.2 분석 결과

4.2.1 두 병원 방문 시점 간의 확률 결합 분포

앞서 설명한 분석 방법을 적용하여 [그림 4]와 같이 Symptoms으로 인한 최초의 병원 방문 시점 (T_S)과 URTD로 인한 병원 방문 시점 (T_U) 간의 확률 결합 분포가 계산되었다. [그림 4]의 표본 공간 분할은 OPT로부터 얻어진 최대사후확률에 대한 확률 결합 분포를 나타낸다. 확률 결합 분포로부터 각각의 방문 시점에 대한 주변 확률 분포 또한 계산되었고, 이 분포는 KM법을 이용한 결과와 거의 유사하다. 시간대 별로 각 구간에서 병원 방문이 이루어질 경우에 대한 확률 예측은 <표 2>에 주어져 있다.

〈표 2〉 Symptoms과 URTD로 인한 병원 방문 시점에 대해 예측된 확률 결합 분포

$T_S \backslash T_U$	0~1	1~2	2~3	3~4	>4 yr
0~1	0.0601	0.0327	0.0250	0.0221	0.3484
1~2	0.0110	0.0110	0.0072	0.0121	0.1059
2~3	0.0017	0.0017	0.0054	0.0049	0.0645
3~4	0.0010	0.0011	0.0028	0.0132	0.0601
> 4 yr	0.0034	0.0017	0.0013	0.0046	0.1969



각 표본의 관찰된 자료는 원으로 표시되었다. x축 방향으로 중도 절단 되었을 경우 세로 선으로 표시되고, y축 방향으로 중도절단 되었을 경우 가로 선으로 겹쳐 표시되었다.

[그림 4] (a) Symptoms과 URTD로 인한 병원 방문 시점 간의 확률결합분포와 (b) 각각의 방문 시점에 대한 생존함수

T_S와 T_U간의 확률결합분포로부터 다양한 결과를 도출할 수 있다. 생후 4년간 Symptoms이나 URTD로 병원을 방문하지 않을 확률은 $\Pr[T_S > 1460, T_U > 1460] = 0.1969$로 계산된다. 이는 LPCH에서 태어나고 진단받는 유아 중 19.7%의 유아는 생후 4년간 이 두 가지 질환으로 병원을 방문하지 않는다는 것을 의미한다. 관찰 기간 내에 Symptoms이나 URTD로 병원을 방문하지 않은 유아의 수는 544명으로 전체 조사 대상의 25.8%에 해당한다. 이러한 단순 계산은 중도절단자료를 고려하지 않았기 때문에 실제 확률인 19.7%보다 높게 계산된다. 비슷한 방식으로 생후 4년 이내에 Symptoms으로만 병원을 방문할 확률은 $\Pr[T_S \le 1460, T_U > 1460] = 0.5789(57.9\%)$, URTD로만 병원을 방문할 확률은 $\Pr[T_S > 1460, T_U \le 1460] = 0.0110(1.1\%)$, 두 질환 모두로 방문할 확률은 $\Pr[T_S \le 1460, T_U \le 1460] = 0.2132(21.3\%)$로 계산된다. 이를 단순히 관찰된 병원 방문만으로 계산했을 경우 각각의 확률은 56.4%, 1.2%, 16.6%의 부정확한 값으로 계산된다. 그러므로 정확한 확률 값을 계산하기 위해서는 본 논문에서 설명하고 있는 TME 분석 방법[26]이 필수적이다.

4.2.2 두 사건 간의 선행성 분석

두 질환으로 인한 병원 방문 중 어느 질환으로 먼저 병원을 방문하는지에 대한 선행성을 확률결합분포로부터 조사할 수 있다. 전체 유아 중 19.7%는 4년간 이 두 가지 질환으로 병원을 방문하지 않기 때문에 어느 질환으로 인한 방문이 선행하는지 판단할 수 없다. 하지만 나머지 80.3%의 유아에 대해서는 선행성의 판별이 가능하다. 명백히 4년 이내에 Symptoms으로만 병원을 방문한 57.9%의 유아에 대해서는 Symptoms으로 인한 방문이 URTD로 인한 방문을 선행한다. 반대로 1.1%의 URTD로만 병원을 방문한 유아에 대해서는 UTRD로 인한 방문이 선행한다. 두 질환 모두로 4년 이내에 병원을 방문한 21.3%의 유아에 대해서는 $\Pr[T_S < T_U \mid T_S \le 1460, T_U \le 1460]$를 계산함으로써 선행성을 판단할 수 있다. 그 결과 21.3%의 유아 중 74.3%(전체 유아에서 15.8%)는 Symptoms으로 병원을 먼저 방문하였고, 25.7%(전체 유아의 5.5%)는 UTRD로 먼저 방문하였다. 전체적으로 종합해 보았을 때, 선행성이 판단 가능한 80.3%의 유아 중 91.8%의 확률로 Symptoms으로 병원 방문이 URTD로 인한 방문을 선행하였다. 그 반대의 경우, URTD로 인한 방문이 Symptoms으로 인한 방문을 선행하는 경우는 8.2%에 그친다. 이와 같은 분석으로부터 두 질환으로 인한 병원 방문 간의 명백한 선행성을 찾을 수 있다.

정확한 선행성 분석은 본 논문에서 설명하고 있는 TME 분석 방법[26]을 통해 가능하다. 예를 들어 관찰된 방문, 즉 비중도절단 자료만으로 선행성을 판별할 경우, 65.7%의 확률로 Symptoms으로 인한 방문이 선행한다는 결과를 얻게 되어, 실제로 존재하는 명확한 선행성을 놓칠 우려가 있다.

4.2.3 두 사건 간의 의존성 분석

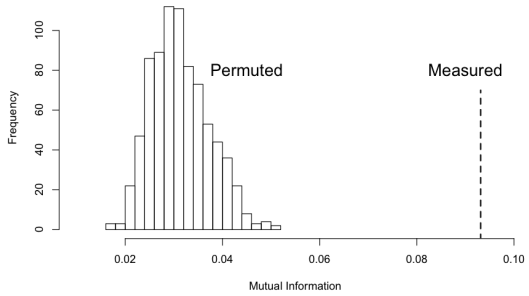
두 질환으로 인한 병원 방문 시점 간의 의존성 혹은 독립성을 측정하기 위해 상호정보량(mutual information)을 이용했다. 상호정보량은 두 확률 변수 간의 의존성을 나타내는 값으로, 변수 간의 관계에 대한 특별한 가정없이 비모수적인 방법으로 구해진다. 두 개의 독립적인 변수에 대해서는 상호정보량은 0이고, 서로에 대한 의존성이 높을수록 큰 값을 갖게 된다. 상호정보량은 두 변수 간의 결합확률 분포로부터 쉽게 구해질 수 있기 때문에, 본 논문에서 설명하고 있는 TME 분석방법[26]을 통해 계산된 결합 확률 분포에 바로 적용 가능하다.

실제로 측정된 자료를 바탕으로 두 변수 간의 상호정보량을 계산하고 해석하는 것은 쉽지 않다. 첫째, 연속변수의 경우 상호정보량의 계산은 각 변수를 여러 구간을 나누어서 하게 되는데, 구간을 나누는 방식에 따라 계산된 상호정보량의 값이 달라진다. 둘째, 상호정보량의 값은 각 변수의 확률분포에 따라 달라지기 때문에, 특정 값의 상호정보량이 얼마나 큰 독립성이나 의존성을 나타내는지 불분명

하다. 예를 들면, 두 독립적인 변수간의 상호정보량은 이론상 0이지만, 제한된 실측 자료로부터 계산된 상호정보량은 오차로 인해 항상 0보다 크다. 이때 측정된 상호정보량이 0에 얼마나 가까운지를 판단하는 것은 쉽지 않다.

이러한 근본적인 문제는 Seok et al.[26]에서 제안하는 OPT를 이용한 TME 분석 방법으로 해결될 수 있다. 첫째, OPT를 이용한 TME 분석 방법은 표본 공간을 균일한 확률 분포를 갖는 여러 개의 작은 영역으로 나눈다. 균일한 확률 분포를 갖는 공간 내에서는 구간을 나누는 방식에 관계없이 상호정보량의 값이 동일하다. 그렇기 때문에, 본 논문에서 설명하고 있는 TME 분석 방법은 상호정보량을 계산하기 위한 적정한 구간을 제공한다. 둘째, 계산된 상호정보량이 얼마나 큰 의존성을 나타내는지 해석하기 위해, 우리는 임의로 교차된(randomly permuted) 자료로부터 계산된 상호정보량을 이용한다. 두 변수를 임의로 교차시킴으로써, 가상의 독립적인 변수를 만들 수 있다. 이 가상의 변수에 대하여 같은 방식으로 상호정보량을 계산함으로써, 측정된 상호정보량이 상대적으로 얼마나 큰지 혹은 작은지를 판단할 수 있다(자세한 내용은 분석 방법 참조).

위와 같은 방식으로 Symptoms과 URTD로 인한 병원 방문 시점 간의 의존성에 대해 조사하였다. 먼저 앞서 설명된 방식으로 T_S와 T_U에 대한 확률 결합분포가 계산되었고, 그에 따라 두 변수 간의 상호정보량이 0.0931로 계산되었다. 하나의 표본에 대한 T_S와 T_U의 쌍을 임의로 교차시켜, 가상의 T_S와 T_U를 생성한다. 이 가상의 자료에 같은 방식을 적용하여 상호정보량을 계산할 수 있다. [그림 5]는 이와 같이 임의 교차를 800회 반복하여 가상의 독립 변수가 갖는 상호정보량의 분포를 계산한 결과를 보여준다. 이 분포는 평균값 0.0313과 표준편차 0.0059를 갖는다. 이 상호정보량의 분포를 정규 분포로 근사 시켰을 경우, 측정된 상호정보량 0.0931에 대한 z값은 (0.0931-0.0313)/0.0059 = 10.41로 계산되고, 이것은 10^{-16}보다 작은 p값에 해당한다. 이러한 분석은 Symptoms과 URTD로 인한 병원 방문 시점이 독립적이지 않고, 서로 연관되어 있음을 통계적 검증을 통해 보여준다.



임의로 교차된 자료로부터 계산된 상호정보량의 분포(Permuted)와 실제 데이터에서 측정된(Measured) 상호정보량이 표시되었다.

[그림 5] Symptoms과 URTD로 인한 병원 방문 시점 간의 상호정보량

4.3.4 두 사건 간의 인과성 분석

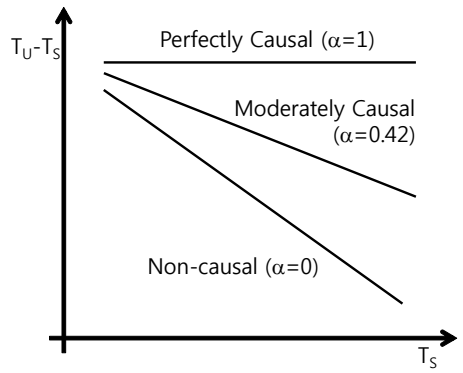
본 논문에서는 두 사건 간의 인과성을 분석하기 위해 두 사건의 발생 시점 간의 차이를 이용하였다. 두 사건 중 어느 한 사건이 높은 확률로 다른 사건보다 먼저 일어나고, 선행한 사건이 일어난 시점부터 일정 시간이 지난 후 다른 사건이 일어나면 이를 인과성이 존재한다고 정의하였다. 이는 선행 사건이 일어난 시점을 아는 것이 다른 사건이 일어나는 시점을 예측하는데 도움이 된다는 것을 의미한다. 그렇게 때문에, 본 논문의 인과성의 정의는 Granger 인과성[10]의 일종으로 생각될 수 있다. 이러한 인과성의 정의는 기본적으로 선행 사건이 발생한 시간과 두 사건 발생 시간 차이 간의 독립성을 의미한다. 두 변수 간의 독립성 혹은 의존성은 앞서 제시한 분석 방법과 같이 상호정보량과 임의 교차를 통하여 측정될 수 있다.

하지만 이러한 인과성에 대한 분석을 실제 자료에 적용하기 위해서는 다음의 문제에 대한 해결이 필요하다. 첫째, 모든 표본 공간에서 두 사건 발생 시간 차이에 대한 확률 분포를 구하는 것은 불가능하다. 절단자료에 대한 분석의 경우 필연적으로 어떠한 영역 밖에서는 확률 분포를 계산할 수 없다. 그렇기 때문에 본 논문에서 사용하는 TME 분석

방법의 경우 ROI를 먼저 설정하고 그 안에서만 확률 분포를 계산하였다. 사건 발생 시간 차이의 경우 ROI 근처에서도 다양한 값들이 발생할 수 있기 때문에 이를 정확히 계산하는 것은 불가능하다. 둘째, 선행사건 발생 시간과 사건 발생 시간 차이 간의 독립성을 명확히 계산하는 것은 어렵다. 이는 먼저 설명한 바와 같이 사건 발생 시간 차이에 대한 분포를 구하기가 어렵기 때문이기도 하지만, 상호정보량의 계산이 어렵기 때문이다. 상호정보량의 계산은 변수의 구간을 어떻게 나누냐에 따라 달라진다. 앞서 의존성 분석의 경우 본 논문에서 설명하는 TME 분석 방법에 의해 적절한 구간이 제공되는데 비해, 사건 발생 시간 차이의 경우 적절한 구간이 제공되지 않는다. 이는 상호정보량의 계산에 필요한 계산량을 증가시킨다. 특히 임의 교차를 통해 가상의 독립적인 자료를 생성하는 경우, 이와 같은 상호정보량의 계산을 수백 회 반복해야 하는 어려움이 있다.

이와 같은 어려움을 감안하여 본 논문에서는 인과성 계수를 통해 두 사건 간의 인과성을 분석한다. 인과성 계수는 후행 사건의 발생 시간을 선행 사건의 발생 시간의 선형 관계로 나타낼 때, 선행사건 발생 시간의 계수로 계산된다. 이 계수가 1에 가까울수록 두 사건 발생 시간의 차이는 선행 사건의 발생 시간에 덜 의존하게 된다. 인과성 계수가 1이라면 이는 선행 사건의 발생 시간과 두 사건 발생 시간의 차이 간의 공분산이 0임을 의미한다. 이는 넓은 의미의 독립성으로 해석될 수 있다. 엄격한 의미의 독립성을 측정하는 대신, 완화된 의미의 독립성인 공분산을 측정하여 필요한 계산량을 줄일 수 있다. 또한, 두 사건 발생 시간 차이의 확률 분포를 계산하는데 있어서 표본 공간 전체에 대해서 계산하는 대신에 적절한 범위 내에서 계산함으로써, 확률 분포 계산에서 오는 어려움을 해소하였다 (자세한 내용은 분석 방법 참조).

이러한 인과성 분석을 Symptoms과 URTD로 인한 병원 방문 시점 자료에 적용하였다. 앞서 분석에 따라 두 병원 방문 시점 사이에는 명확한 선행성이 존재한다. 그러므로 여기에서는 선행 사건의 발생 시간 T_S와 두 사건의 발생 시간 차이 T_U-T_S간의 독립성에 대해 주로 분석한다. T_U-T_S의 확률 분포의 계산을 위해서는 4년(1460일)까지의 전체 표본 공간 대신에 1100일까지의 부분 공간이 사용되었다. 이 분석에서 인과성 계수는 0.42로 계산되었다. 이는 T_S와 T_U-T_S사이의 공분산이 0이 아님을 보여준다. 인과성 계수가 1과는 차이가 있음에도 불구하고, T_S를 아는 것은 T_U-T_S를 예측하고, 나아가 T_U를 예측하는데 도움을 준다. [그림 6]은 인과성 계수에 따른 T_S와 T_U-T_S간의 관계를 보여준다. 인과성 계수가 1인 경우는 URTD로 인한 병원 방문 시점이 전적으로 Symptoms으로 인한 병원 방문 시점에 의존하기 때문에, Symptoms으로 인한 병원 방문이 URTD으로 인한 병원 방문에 대해 완벽한 Granger 인과성을 갖는다. 이는 의학적인 측면에서 Symptoms이 URTD를 유발하거나 혹은 URTD의 전조 증상으로 해석될 수 있다. 인과성 계수가 0인 경우는 URTD로 인한 병원 방문과 Symptoms으로 인한 병원 방문이 별개로 이루어진다. 그래서 전혀 Granger 인과성을 갖지 않는다. 실제 자료에서 계산된 인과성 계수 0.42는 두 질환으로 인한 병원 방문 간의 약한 Granger 인과성을 나타낸다고 해석될 수 있다.



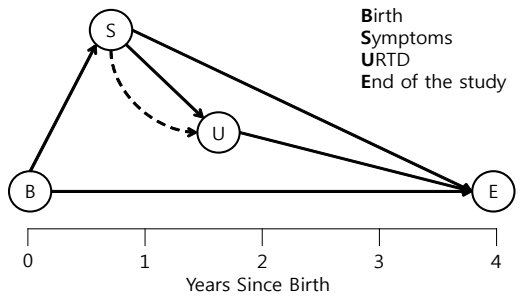
[그림 6] 인과성 계수(α)에 따른 Symptoms으로 인한 방문 시점(T_S)과 Symptom으로 인한 방문에서 URTD로 인한 방문까지의 경과 시간(T_U-T_S)과의 관계

4.3.5 전체 분석 요약

주어진 자료의 분석 결과는 [그림 7]과 같이 요약할 수 있다. 병원에서 태어난(사건 B) 유아는 57.9%의 확률로 Symptoms으로 병원을 다시 방문(사건 S)하고, 본 연구에서 설정된 관측 범위에서 벗어나게 된다(사건 E). 전체 유아의 19.7%는 Symptoms이나 URTD로 인한 방문 없이 관찰에서 제외된다. 15.8%의 유아는 Symptoms으로 병원을 방문하고 그 후 다시 URTD로 병원을 방문(사건 U)한다. 이 세 가지 경로는 LPCH에서 태어난 유아가 겪게 되는 호흡기 관련 질환으로 인한 병원 방문이 가장 빈번한 경우를 나타낸다. Symptoms으로 인한 병원 방문은 91.8%의 확률로 URTD로 인한 병원 방문에 선행하고, 두 방문 시점은 통계적으로 매우 유의미한 (p값 $< 10^{-16}$) 명백한 의존성을 갖는다. 선행하는 Symptoms으로 인한 방문은 후행하는 URTD로 인한 방문에 대해 약한 인과성(α = 0.42)을 갖고 있다.

본 자료의 분석결과를 통해 의학적인 측면에서는 URTD를 유발하거나 혹은 URTD에 선행하는 주요 증상으로 Symptoms에 속한 질병이나 증상이 연구될 수 있다. 이는 URTD를 치료하거나 예방하는데 주요한 정보로 작용한다. 병원 경영적 측면에서는 현재 Symptoms으로 방문한 환자의 자료를 바탕으로 얼마나 많은 환자가 얼마 후에 URTD로 병원을 찾을지를 예측할 수 있다. 이러한 정보는 병원의 서비스를 향상시키고, 운영 비용을 줄이는 주요한 정보로 활용될 수 있다.

위의 모든 분석은 Seok et al.[26]이 제안한 다변량 생존분석 방법을 기초로 한다. 이 방법은 기존의 다변량 생존분석의 주요한 문제점을 해결하여, 실제 분포에 더 가깝게 확률분포를 예측한다. 예를 들어, 기존의 다변량 분석방법 중의 하나인 Pathwise 방법[6]을 적용하여 논문의 데이터를 분석하는 경우, 생후 4년 이내에 URTD로만 병원을 방문할 확률, 즉 $\Pr[T_S > 1460,\ T_U \le 1460]$는 음의 값(-3.4%)으로 계산된다. 음의 확률은 기존 다변량 생존분석의 주요한 문제점 중 하나이다. 확률의 전체 합은 1이기 때문에 음의 확률의 영향으로 다른 부분의 확률이 과대 예측된다. 그 결과 Symptoms으로 인한 병원 방문이 URTD로 인한 병원방문보다 선행할 확률($\Pr[T_S < T_U]$)은 108.3%로 계산되어, 그 해석이 불가능해진다. 본 논문에서는 새로운 분석방법을 적용하여, 이변량분석이 실제적으로 유용하다는 것을 보였다.



표본의 탄생(B : Birth)로 관찰이 시작되어 생후 만 4년의 시점에서 관찰이 종료(E : End of the study)된다. Symptoms과 URTD로 인한 병원 방문은 각각 S와 U로 표현된다. 각 사건은 발생 시점의 관찰된 중간값 지점에 표시되어 있다. 실선으로 표시된 화살표는 사건을 발생 순서를 나타내고, 점선으로 표시된 화살표는 약한 인과성을 나타낸다.

[그림 7] 사건 관계 요약

5. 결론 및 향후 연구

본 연구는 기존 연구[26]에서 제안된 중도절단자료를 포함한 다사건 시계열 자료를 분석하는 베이지안 기반의 다변량 생존분석 방법론을 설명하고, 유효성을 검증하기 위한 응용 사례로 소아환자의 진료기록 자료를 통해 실증 분석하였다. 본 연구에서 사용하는 OPT를 이용한 TME 분석 방법은 전체 표본 공간을 균일한 확률분포를 갖는 작은 공간들의 트리(tree) 형태로 분할하고, 주어진 다변량 중도 절단 자료로부터 서로 다른 분할 방식에 대한 사후확률을 계산함으로써, 다사건 시계열 자료의 사건 간의 확률 결합 분포를 예측하였다. 본 논문에서는 이러한 방법론을 실제 데이터에 적용하여, 서로 다른 사건들 간의 연관성, 특히 선행성, 의존성,

인과성에 대한 분석법을 설명하고 있다. 의료데이터에서 한 질환으로 인한 방문이 다른 질환으로 인한 방문에 선행할 확률 및 두 방문 사이의 상관관계와 인과관계에 대해 분석하였다. 본 연구를 첫걸음으로 하여 다사건 시계열 분석방법론을 경영과학 분야에서 보다 폭넓게 활용할 수 있도록 비즈니스 영역 자료에 적용한 향후 추가 연구가 기대된다.

가장 먼저 고객관계관리(CRM) 시스템의 향상에 기여할 수 있을 것이다. 오늘날에는 고객의 가입, 방문, 구매 이력 데이터들이 데이터베이스 기반이 갖춰진 환경 하에 축적되고 있으며, 고객 정보를 통해 기업의 이익증대에 기여할 수 있는 활용 방안을 찾고자 하는 움직임이 활발하다. 본 연구에서 설명하고 있는 다사건 시계열 분석 방법은 고객 데이터베이스를 바탕으로 한 다양한 마케팅 전략 수립 연구에 큰 기여를 할 것으로 기대된다. 신규고객 획득, 기존 고객의 이탈을 방지하고 보다 깊은 관계의 고객으로 발전시켜 나가서, 궁극적으로 고객 수익성 증대를 위하여, 기업은 고객과 지속적인 커뮤니케이션을 해야 한다.

둘째, 고객생애가치(customer lifetime value, CLV)를 추정하기 위한 연구에 적용한다면 흥미로운 결과를 얻을 수 있을 것으로 기대된다. 지금까지는 고객생애가치의 초점이 종합금융서비스(은행, 카드, 증권), 통신서비스(초고속인터넷, 이동통신, 유선전화, 또는 세대 간 교체구매) 등의 업종과 같이 고객과의 계약적 상황이 존재할 경우에 한 개인의 가입 시점부터 이탈 시점까지의 한 고객으로부터 발행되는 모든 수익의 가치였다면, 이제 일반적으로 고객의 반복 구매나 다른 제품으로의 교차 구매까지 고려해야 한다. 다사건 시계열 분석 방법론을 이용한다면 미리 우량고객과 그렇지 않은 고객을 분류하여 보다 효율적으로 소비자와의 깊은 관계를 구축할 수 있을 것이다. 특히, 제품군 간의 상관관계를 통해 하나의 서비스에 가입한 고객의 다른 서비스 가입 여부와 이탈 등을 사전에 예측하여 보다 효과적으로 고객 관계에 활용할 수 있을 것으로 기대된다.

셋째, 본 연구에서 설명하고 있는 다사건 시계열 분석 방법론을 활용한다면, 고객정보데이터베이스를 통한 보다 효과적인 일대일 타켓 마케팅도 가능할 것이다. 고객 구매 이력 데이터를 통해, 구매 확률이 높은 고객에게 적합한 상품을 추천해주는 개인화 알고리즘 개발에 기여할 수 있을 것으로 기대된다. 제품 간 상관성 분석을 통해 인터넷 쇼핑몰에서 서로 다른 제품군의 구매에 걸리는 시간을 분석하여, 소비자들의 구매에 대한 예측 분석에 활용할 수 있을 것이다. 이외에도 가격전략에 대한 개별 소비자 반응을 예측하는데 활용될 수 있을 것이다. 가격 인상이나 가격 하락에 따른 소비자의 구매 시점 변화, 가격의 일시적 할인에 따른 구매와 재구매 기간 분석 등에서 의미 있는 결과를 얻을 수 있을 것이다. 이를 바탕으로 맞춤형 할인 쿠폰의 발급이나 할인 정보 제공 등의 일대일 타켓 마케팅 기법으로 확장할 수 있을 것으로 기대된다. 이와 같이 다사건 시계열 분석 방법은 실제 고객 데이터에 적용하여 고객관계관리, 고객가치를 중요시하는 여러 산업에 걸쳐서 다양하게 활용할 수 있을 것으로 기대된다.

참 고 문 헌

[1] 남재우, 이회경, “생존분석 기법을 이용한 기업 도산 예측 모형”, 「한국경영과학회 추계학술대회 논문집」, (2000), pp.40-43.

[2] 박재빈, 「생존분석 : 이론과 실제」, 신광출판사, 2007.

[3] 하성호, 양정원, 민지홍, “코호넨네트워크와 생존분석을 활용한 신용 예측”, 「한국경영과학회지」, 제34권, 제2호(2009), pp.35-54.

[4] Ascarza, E. and B.G. Hardie, “A Joint model of usage and churn in contractual setting,” *Marketing Science*, Vol.32(2013), pp.570-590.

[5] Bandyopadhyay, S., J. Wolfson, D. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. Johnson, and P. O'Connor, “Data mining for censored time-to-event data : A Bayesian network model for predicting cardio-

vascular risk from electronic health record data," *working paper*, arXiv : 1404.2189 (2014).

[6] Campbell, G. and A. Földes, "Large sample properties of nonparametric bivariate estimators with censored data," Proceedings, International Colloquium on Nonparametric Statistical Inference, (1982), pp.23-28.

[7] Dabrowska, D., "Kaplan-Meier estimate on the plane," *Annals of Statistics*, Vol.16(1988), pp.1475-1489.

[8] DuWors, Jr., R.E. and G.H. Haines, Jr, "Event history analysis measures of brand loyalty," *Journal of Marketing Research*, Vol.27(1990), pp.485-493.

[9] Efron, B., "The two-sample problem with censored data," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol.4(1967), pp.831-853.

[10] Granger, C.W.J., "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, Vol.37(1969), pp.424-438.

[11] Helsen, K. and D.C. Schmittlein, "Analyzing duration times in marketing : evidence for the effectiveness of hazard rate models," *Marketing Science*, Vol.11, No.4(1993), pp. 395-414.

[12] Hidalgo, C.A., N. Blumm, and A-L. Barabasi, and N.A. Christakis, "A Dynamic network approach for the study of human phenotypes," *PLoS Computational Biology*, Vol.5, No.4(2009), e1000353.

[13] Kaplan, E.L. and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of American Statistical Association*, Vol.53(1958), pp.457-481.

[14] Lariviere, B. and D. Van den Poel, "Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling : the case of financial services," *Expert Systems with Applications*, Vol.27(2004), pp.277-285.

[15] Lavine, M., "Some aspects of Polya tree distributions for statistical modeling," *Annals of Statistics*, Vol.20(1992), pp.1222-1235.

[16] Lavine, M., "More aspects of Polya tree distributions for statistical modeling," *Annals of Statistics*, Vol.22(1994), pp.1161-1176.

[17] Li, S., "Survival Analysis," *Marketing Research*, Vol.7, No.4(1995), pp.17-23.

[18] Milovic, B. and M. Milovic, "Prediction and decision making in health care using data mining," *International Journal of Public Health Science*, Vol.1, No.2(2012), pp.69-78.

[19] Mukherjee, A. and J. McGinnis, "E-healthcare : an analysis of key themes in research," *International Journal of Pharmaceutical and Healthcare Marketing*, Vol.1, No.4(2007), pp. 349-363.

[20] Muliere, P. and S. Walker, "A Bayesian nonparametric approach to survival analysis using polya trees," *Scandinavian Journal of Statistics*, Vol.24(1997), pp.331-340.

[21] Neath, A.A., "Polya tree distributions for statistical modeling of censored data," *Journal of Applied Mathematics and Decision Sciences*, Vol.7, No.3(2003), pp.175-186.

[22] Oakes, D., "Biometrika Centenary : Survival analysis," *Biometrika*, Vol.88, No.1(2001), pp. 99-142.

[23] Prentice, R. and Cai, J., "Covariance and survival function estimation using censored multivariate failure time data," *Biometrika*, Vol. 79(1992), pp.495-512.

[24] Rigobon, R. and T. Stoker, "Estimation with censored regressors : basic issues," *Interna-*

tional Economic Review, Vol.48, No.4(2007), pp.1441-1467.

[25] Schemper, M., A. Kaider, S. Wakounig, and G. Heinze, "Estimating the correlation of bivariate failure times under censoring," *Statistics in Medicine*, Vol.32, No.27(2012), pp. 4781-4790.

[26] Seok, J., L. Tian, and W.H. Wong, "Density estimation on multivariate censored data with optional Pólya tree", *Biostatistics*, Vol.15, No.1(2014), pp.182-95.

[27] Stajduhar, I. and B. Dalbelo-Basic, "Learning Bayesian networks from survival data using weighting censored instances," *Journal of Biomedical Informatics*, Vol.43, No.4(2010), pp.613-622.

[28] Turnbull, B., "The empirical distribution function with arbitrary grouped censored and truncated data," *Journal of the Royal Statistical Society Series B*, Vol.38(1976), pp.290-295.

[29] Wong, W.H. and L. Ma, "Optional Polya Tree and Bayesian Inference," *Annals of Statistics*, Vol.38(2010), pp.1433-1459.

[30] Zhou, X., J. Menche, A. Barabasi, and A. Sharma, "Human symptoms-disease network," *Nature Communications*, Vol.5, No. 4212(2014).