한국ITS학회논문지 J. Korea Inst. Intell. Transp. Syst.

복합확률분포의 파라메타 추정을 위한 EM 알고리즘의 적용 연구

An Approach for the Estimation of Mixture Distribution Parameters Using EM Algorithm

심 대 영* · 김 상 구**

* 주저자 : 가톨릭관동대학교 건축학과 교수
** 교신저자 : 전남대학교 물류교통학과 교수

Daeyoung Shim* · SangGu Kim**

* Dept. of Architecture, Catholic Kwandong University
** Dept. of Logistics and Transportation, Chonnam National University

† Corresponding author : SangGu Kim, kim-sg@jnu.ac.kr

Vol. 22 No.4(2023)
August, 2023
pp.35~47

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2023.22.4.35

Received 16 May 2023
Revised 1 June 2023
Accepted 5 July 2023

요 약

그동안 차두시간분포를 나타내는 확률분포로 음지수분포, Erlang 분포, 정규분포 등 다양한 단일확률분포들이 사용되어져 왔다. 그러나, 실제 도로에서 차두시간분포의 조사결과는 단일확률분포로서 설명하기 어려운 경우가 있었다. 본 연구는 차량의 차두시간에 대해 두 개의 정규분포가 일정한 관련성을 가지고 결합된 복합확률분포의 파라메타에 대해 최우추정법 중 하나인 EM 알고리즘을 이용하여 추정하는 접근방법을 시도하였다. 이에 대한 분석결과 기존에 알려진 단일확률분포로서 잘 설명되기 어려웠던 차량도착 차두시간 분포를 EM 알고리즘을 이용하여 복합확률분포의 파라메타를 추정하여 설명하였다. χ^2 test 적합도 검정결과, 유의수준 1%에서 통계학적으로 유의성이 확보되어 EM 알고리즘을 이용한 복합확률분포의 파라메타 추정의 신뢰성이 입증되는 것으로 분석되었다.

핵심어 : 차두시간, 최우추정법, EM알고리즘, 복합확률분포

ABSTRACT

Various single probability distributions have been used to represent time headway distributions. However, it has often been difficult to explain the time headway distribution as a single probability distribution on site. This study used the EM algorithm, which is one of the maximum likelihood estimations, for the parameters of combined mixture distributions with a certain relationship between two normal distributions for the time headway of vehicles. The time headway distribution of vehicle arrival is difficult to represent well with previously known single probability distributions. But as a result of this analysis, it can be represented by estimating the parameters of the mixture probability distribution using the EM algorithm. The result of a goodness-of-fit test was statistically significant at a significance level of 1%, which proves the reliability of parameter estimation of the mixture probability distribution using the EM algorithm.

Key words : Time headway, Maximum likelihood estimation, EM algorithm, Mixture probability distribution

Ⅰ. 서 론

자료의 특성을 정확하게 파악하기 위해서는 주어진 자료의 확률분포 특성을 파악하는 것이 필요하다. 그러나 실제 조사된 자료는 모집단으로부터 표본 조사된 자료(sample)이기 때문에 표본으로부터 전체 자료의 특성을 파악하기에는 부족한 면이 많은 경우가 많으며 또 주어진 자료의 특성이 지금까지 알려진 확률분포로는 설명하기 어려운 형태를 띄는 경우에는 자료의 확률분포 특성을 파악하기 어려운 경우가 많이 있다.

교통분야에서는 확률분포가 도착 차두시간 분포를 설명할 때 주로 활용된다. 미시적으로 교통류 특성을 분석하고자 할 때 차두시간 분포 형태를 가장 잘 모사할 수 있는 수학적 확률분포를 결정하는 것으로 오랫동안 교통분석가들의 주된 관심사였다. 그동안 차두시간 분포를 나타내는 확률분포로 교통량 수준에 따라 음지수분포, Erlang 분포, 정규분포 등 다양한 단일확률분포들이 사용되어져 왔다. 그러나, 실제 도로에서 차두시간분포의 조사결과는 단일 확률분포로서 설명하기 어려운 경우가 있어서 두 가지 이상 분포의 특성을 가진 복합확률분포의 필요성이 대두되고 시도되었다(May, 1990).

복합확률분포(mixture distribution)는 자료 내에 두 가지 이상의 확률분포 특성이 포함되어 있기때문에 하나의 확률분포로 설명하기 어려운 경우에는 두 가지 이상의 확률분포 특성을 분리하여 차두시간 자료를 설명할 수 있어야 한다.

그러나 이러한 복합확률분포의 특성을 파악하기 위한 확률변수 파라메타를 도출하는 방법에 아직 신뢰성이 부족한 면이 있어 두 가지 이상 복합적 분포의 비율을 단순한 변수로 가정하여 복합분포를 설명하는 방법이 사용된 경우가 많다.

본 연구는 두 가지 이상의 교통류 특성이 포함되어있는 확률분포를 잘 설명할 수 있는 복합확률분포의 파라메타 추정을 위해 EM 알고리즘(Expectation Maximization Algorithm)을 교통분야에 적용해보는 새로운 시도를 해보는 데 목적을 가지고 있다.

EM 알고리즘은 파라메타 추정을 위한 전통적인 최우추정법을 발전시킨 파라메타 추정방법으로, 조건부 기대값의 추정과 파라메타의 우도(likelihood)를 최대화해 가는 방향의 파라메타 결정의 두 가지 단계의 반복적 시행으로 최종적인 파라메타를 탐색해 가는 방법으로, 기존의 최우추정법에 비해 수렴의 속도가 다소 느린 점은 있지만, 초기값을 가정한 반복수행 과정을 통해 결과를 추정하는 방법의 적용으로 단순 최우추정법으로 설명하기 어려운 확률분포의 파라메타 추정이 가능하고 무엇보다 자료에 불충분자료(incomplete data 또는 missing data)가 포함된 경우에도 사용할 수 있는 방법으로서 많은 개선 내용이 진행되고 있는 추정 알고리즘으로 자리잡고 있다.

Ⅱ. 관련 연구고찰 및 복합확률분포

1. 차두시간분포의 수학적 확률분포

차두시간과 관련된 수학적 확률분포에 관한 주제는 1930년대부터 연구되어왔고 이를 위해 개별 차두시간 관측자료가 여러 문헌에서 이용되어왔다. May(1990)는 여러 연구자들에 의해 연구되어왔던 차두시간분포의 수학적 확률분포를 체계적으로 잘 정리하였고 이를 토대로 관련 연구를 살펴보면 다음과 같다.

차두시간 분포의 형태는 교통량 수준이 증가함에 따라 상당히 변화하는데 이는 교통류내 차량들간 상호작용이 증가하기 때문이다. 예를 들어 낮은 교통류 상태에서는 차량들간 상호작용이 적기 때문에 차두시간

은 다소 무작위(랜덤)로 나타난다. 그리고 교통류 수준이 증가함에 따라 차량들간 상호작용도 증가하고 이로 인해 일부 차량의 차두시간은 무작위로 나타나고 다른 차량들은 차량추종 상태를 나타내기도 한다. 교통류 수준이 용량상태로 높아지게 되면 거의 모든 차량들은 차량추종 상태의 상호작용을 보인다. 따라서 차두시간 분포는 낮은 교통량 수준에서 랜덤분포, 보통의 교통량 수준에서 중간분포, 그리고 높은 교통량 수준에서 균일분포로 분류할 수 있다.

무작위 차두시간 상태에서 랜덤간격 분포를 나타내는 수학적 분포는 음지수분포가 주로 이용되고 이는 모든 차량들이 서로 독립적으로 주행하면서 매우 낮은 교통량 조건일 때 주로 적용된다. 차두시간이 모두 일정하게 나타나거나 운전자가 일정한 차두시간으로 유지하다가 운전자 실수로 일정한 차두시간이 변화하는 경우일 때 수학적 분포로서 균일분포가 이용될 수 있다. 실제 교통류 상황에서 가장 흔히 발생하는 차두시간 분포는 랜덤분포와 균일분포 경계의 사이에 놓여있는 중간 차두시간 상태로서 수학적 분포는 피어슨 Ⅲ형 분포가 주로 이용되고 이는 음지수분포, Erlang 분포, Gamma 분포, 정규분포 등 다양한 형태의 확률분포를 포함하고 있다. 이러한 중간 차두시간 상태는 교통류내 일부 차량들은 서로 독립적으로 움직이고 다른 차량들은 서로 영향을 주고 받는 상호작용하는 상태로 주행한다.

이상과 같이 교통량 수준에 따른 단일 확률분포의 적용과는 다르게 교통류 상태를 2개로 구분하여 각각에 적합한 수학적 확률분포를 조합하는 합성모형(composite model)이 또 다른 접근방법이다. May(1990)는 차량추종이나 차량군 상태의 정규분포와 상호작용이 없는 차량들을 위한 음지수분포를 조합하는 방식으로 차두시간 분포를 설명하는 합성모형을 적용하였다. 이외에도 다른 분포들의 조합으로는 변위된 음지수분포(shifted negative exponential distribution)와 음지수분포로 조합된 합성모형도 있다. 기타 기하정규분포(log-normal distribution)가 차두시간 분포를 설명하는 확률분포로 사용되기도 하였고 이항모형(binomial model), hyperlang model, semirandom model 등 많은 모형들이 그동안 연구되어왔다.

그동안 적용해오던 차두시간 분포를 위한 수학적 확률분포는 교통류내 차량들의 특성이 2개 이상으로 복잡하거나 확률분포 형태를 결정하는 파라메타 추정이 쉽지 않아 신뢰성이 저하되는 단점을 가지고 있었다. 따라서, 교통류내 다양한 교통특성을 가진 복합적인 차두시간 분포를 설명하거나 이러한 복합확률분포의 파라메타를 신뢰성있게 추정하는 새로운 접근방법에 대한 연구가 필요하다.

2. 복합확률분포

복합확률분포(mixture distribution)의 확률분포함수($f(x)$)는 다음의 일반식으로 표현할 수 있으며 , 아래에 표시된 이러한 수식에 대해서 Everitt and Hand(1981)를 참고할 수 있다.

$$f(x) = \int g(x;\Theta)\, dH(\Theta) \quad \cdots\cdots (1)$$

여기서 $g(x;\Theta)$는 파라메타 Θ를 가지는 각각의 확률밀도함수(probability density function)이며 H는 각 확률분포의 복합관계를 설명하는 누적확률분포이다. 한편 위의 완전한 일반식은 보편적인 경우에 비해서는 너무 일반적이므로 보편적으로는 위의 식을 c개의 확률분포가 복합적인 분포를 이루고 있고, H가 이산적인 경우에는 다음의 식과 같이 표현될 수 있다.

$$f(x) = \sum_{i=1}^{c} H_i(\Theta_i)\ g(x;\Theta_i) \quad \cdots\cdots (2)$$

예를 들어, 어떤 복합확률분포 $f(x)$가 두 개의 일양분포(uniform 분포 ; U(a, b))로써 각각 p1과 p2의 비율로서 복합되어있는 경우라고 한다면 이 관계는 다음의 식으로 표현될 수 있다. 이때 이 식에서 $p_1 + p_2 = 1$ 이어야 한다.

$$f(x) = p_1\, U(a_1, b_1) \; + \; p_2\, U(a_2, b_2) \quad \cdots\cdots (3)$$

따라서, 이런 모형에서 추정하여야 하는 파라메타는 각 일양분포의 파라메타 a_1, a_2, b_1, b_2와 복합분포 비율 p_1, p_2에 대한 추정이 필요하다.

복합확률 분포는 실제 상황에서 다양하게 나타날 수 있다. 가령 어떤 학급의 키를 조사하여 정규분포로써 설명한다고 하자. 그러나 이때 남자와 여자의 키에 대한 확률적 분포의 특성은 서로 다른 경우가 많다고 할 수 있다. 따라서 서로 다른 남녀의 특성을 구분하지 않고 하나의 정규분포로서 설명하는 경우에 비해서, 서로 다른 남녀의 특성을 각각 별도의 정규분포로 정의하고 이를 복합시킨 복합확률분포로서 설명하는 경우의 설명력이 더 나을 것이라고 기대할 수 있다.

교통부문에서도 이러한 예로서는 다양한 사례가 있다고 판단된다. 일반적으로 차량의 도착분포에 대해서 확률분포로서 설명하는 경우가 많고, 이때 일반적으로 교통량이 많은 경우는 차량도착분포는 정규분포의 특성이 주로 나타나고 교통량이 적은 경우는 음지수분포의 특성을 보인다고 설명하고 있다(May, 1990). 그러나 일반적인 가로 상에서 차량의 도착시간에 대한 확률분포를 조사하는 경우 차량의 도착이 차량군을 이루는 경우와 차량군을 이루지 않는 경우로 구분되고, 각각의 도착 차두시간에 대한 확률분포의 특성은 각 경우에 따라 다양한 차이가 발생하여 각 개별적인 특성을 무시하고 단일 확률분포로서 설명하기보다는 복합확률분포로 설명하는 것이 바람직한 경우가 있다.

이외에도 도시부 가로망에서 직진, 좌회전, 우회전하는 교통류의 도착분포에 있어서도 각 개별적인 특성이 다른 경우라고 생각되며, 속도조사의 경우에 있어서도 승용차와 대형차의 평균적인 속도특성이 다른 경우 일반적인 단일 확률분포로서는 설명이 어려운 경우가 있는 등 복합확률분포의 적용이 필요한 다양한 사례가 발생한다고 할 수 있다.

Ⅲ. EM(Expectation Maximization) 알고리즘

일반적인 통계학에서의 확률분포 파라메타의 추정을 위한 방법으로는 최소제곱법(least square estimation)과 최우추정법(maximum likelihood estimation)이 주로 이용된다(Kim et al., 1998; Jeong et al., 1999). 복합확률분포 파라메타의 추정 방법으로는 전통적인 최우추정법의 적용, 베이지안 추정, inversion and error 최소화 방법 등의 적용이 있어왔다(Everitt and Hand, 1981). 본 연구에서는 최우추정법의 한 유형으로서 새롭게 그 유용성이 대두되고 있는 EM 알고리즘을 이용하여 복합확률분포의 파라메타 추정을 시도해 보고 적합성을 검정하였다. 아래의 최우추정법에 대한 수식 표현은 Jeong et al.(1999)등의 통계학 교재를 참고하여 작성하였으며, EM 알고리즘에 표기된 수식은 Dempster et al.(1977)과 Everitt and Hand(1981)을 통하여 재구성하였다.

1. 최우추정법

조사과정를 통해 수집된 n개의 자료를 $\{x_1, \ldots\ldots, x_n\} = X^n$이라고 하고, 이 자료가 파라메타 Θ를 가지는 확률밀도함수 $f(x;\Theta)$를 따르는 경우 최우추정법 (maximum likelihood Estimation)은 자료의 결합확률로써 다음의 식으로 정의되는 우도함수(likelihood function, $\mathcal{L}$)를 최대화하는 파라메타 Θ를 추정하게 된다.

$$\mathcal{L}(\Theta;X^n) = \prod_{i=1}^{n} f(x_i;\Theta) \quad \cdots\cdots (4)$$

위의 식에서 자료 xi에 대해서는 그 값을 가지고 있으므로, 위의 우도함수식은 파라메타 Θ에 대한 함수의 형태를 가지게 되고, 최우추정법에 의한 파라메타의 추정은 이 우도함수를 최대화하는 파라메타 Θ를 추정하게 되므로, 최대점에서 위의 우도함수식의 미분이 0이 되어야 하는 다음의 관계식이 성립하여야 한다.

$$\frac{\partial \mathcal{L}}{\partial \Theta_i} = 0 \quad \cdots\cdots (5)$$

한편, 위에서 정의한 우도함수는 곱의 관계를 가지는 복잡한 형태이므로 미분 과정이 용이하지 않으므로 이 식에 로그를 취한 로그우도함수(L)를 사용한 최대화 관계식을 이용하여도 동일한 결과를 가지게 된다.

$$L = \log \mathcal{L}(\Theta;X^n) \quad \cdots\cdots (6)$$

즉, 최우추정법은 파라메타 Θ의 추정치 $\hat{\Theta}$에 대해 $\hat{\Theta} = \arg\max_{\Theta} \mathcal{L}(\Theta \mid x)$ 또는 $\hat{\Theta} = \arg\max_{\Theta} L(\Theta \mid x)$의 관계로 표현된다.

일반적으로 최우추정법은 이러한 우도함수의 미분을 통해 해석적으로 명시적인 해를 도출하거나, Newton, 또는 Modified-Newton, Quasi -Newton, Newton-Rapson 방법 등의 수치해석 방식을 통해 해를 탐색해 가는 방법(Hill-Climb Method)을 사용하게 된다.

그러나, 이러한 일반적인 최우추정법은 복합확률분포의 경우에 적용이 어렵다. 그 이유는 복합확률분포의 경우 우선 명시적인 해석적 미분의 도출이 어려울 뿐 아니라, 일반적인 수치해석 방법의 초기치의 문제나 국지적 해(local optimization)에 빠지게 되는 오류를 해결하지 못하기 때문이다. 또, 복합확률분포에 대해 일반적인 최우추정법을 적용하는 경우의 해는 국지적 해에 빠지게 되거나, $\mathcal{L}(X^n;\Theta) = \infty$인 경우의 해로 나타나 찾고자 하는 해의 수렴성을 보장해 주지 못하는 것으로 알려지고 있다(Everitt and Hand, 1981).

2. EM 알고리즘

EM 알고리즘 (Expectation-Maximization Algorithm)은 Dempster et al.(1977)에 의해 체계화된 최우추정법의 수행을 위한 추정 알고리즘의 한 방법(Dempster et al., 1977)으로, Newton의 방법 등 일반적인 해의 탐색 과정으로는 추정이 어려운 경우에 사용될 수 있고, 또 특히 불충분한 자료(incomplete data 또는 missing data)가 사용되는 경우에 그 유용성이 입증된 알고리즘이다.

EM 알고리즘은 조사를 통해 주어진 자료 X^n이 불충분하다고 하고 위의 우도함수 $\mathcal{L}$은 불충분 우도함수로 간주한다. 따라서 완전한 자료 Z는 알려져 있지만 불충분한 자료 X와 알려져 있지 않은 자료 Y에 의해 완전한 자료 $Z=(X, Y)$로 구성된다고 표현한다. 따라서, EM알고리즘은 알려지지 않은 자료 Y와 추정되어야 하는 파라메타 Θ에 대해서는 초기값으로 가정하고 먼저 반복적으로 완전한 자료 Z에 대한 로그우도함수의 기대값(Q)을 다음의 관계로부터 구한다. 이 과정을 E-단계 (expectation step)라고 하고 이 식에서 Θ^{i-1}은 이전 단계의 파라메타를 표시하고 추정될 파라메타는 Θ로 표현되고 있다.

$$\text{E-step}: \quad Q(\Theta, \Theta^{i-1}) = E\left[\log f(X, Y \mid \Theta) \mid X, \Theta^{(i-1)}\right] \quad \cdots\cdots (7)$$

이렇게 현 단계의 확률(우도함수)에 대한 기대값이 주어지면, 우도함수가 최대화되는 방향으로 다음 단계에 사용될 파라메타들을 계산한다. 이 관계를 M-단계라고 하고 다음의 식으로 표현된다.

$$\text{M-Step}: \quad \Theta^i = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)}) \quad \cdots\cdots (8)$$

즉, M-단계에서는 우도함수의 최대화를 위해 기대값(Q)을 최대화하는 다음 단계에서 사용할 파라메타를 탐색하는 과정이라고 설명할 수 있다. 일반적으로 M-단계에서 기대값 Q를 직접 최대화하기보다는 최대화에 부합하는 방향으로 파라메타 탐색을 수행하며, 이러한 관계를 GEM (Generalized EM)이라고 하며, 다음의 식으로 표현한다.

$$\text{GEM의 M-Step}: \quad Q(\Theta^{(i)}, \Theta^{(i-1)}) > Q(\Theta, \Theta^{(i-1)}) \quad \cdots\cdots (9)$$

따라서 EM 알고리즘은 미지의 값(파라메타와 missing data)에 대해서는 일정 값을 가정하고, 우도함수의 기대값을 계산하는 E-단계와 다음 단계에서 사용할 파라메타를 추정하는 M-단계의 반복적인 계산과정을 통해 일정 수준의 수렴성이 보장될 때까지 반복계산을 수행하여 파라메타를 추정하는 과정으로 설명할 수 있고, 보통 일반적 최우추정법에 비해 반복계산의 회수는 많은 것으로 알려지고 있다. 한편, 이러한 EM 알고리즘의 수렴성은 일반적인 여러 문헌에서 증명되어 있다(Dempster et al., 1977; McLachlan and Krishnan, 1997).

본 연구에서는 일반적인 최우추정법으로는 추정이 어려운 복합확률분포에 대한 파라메타 추정을 위하여 EM 알고리즘을 적용하였다. 복합확률분포에 대한 우도함수 $\mathcal{L}$는 다음 식의 형태로 표현할 수 있다.

$$\mathcal{L}(p, \Theta_1, \ldots, \Theta_c; X^n) = \prod_{j=1}^{n} \left[\sum_{i=1}^{c} p_i g_i(x_j; \Theta_i) \right] \quad \cdots\cdots (10)$$

한편, 이 우도함수에 대한 로그우도 함수는 Lagrange 상수 λ를 도입하여 다음과 같이 표현할 수 있다.

$$L(p, \Theta_1, \ldots, \Theta_c; X^n) = \log \mathcal{L}(p, \Theta_1, \ldots, \Theta_c; X^n) - \lambda\left(\sum_{i=1}^{c} p_i - 1\right) \quad \cdots\cdots (11)$$

우도함수를 최대화하기 위해 다음의 미분 관계식을 이용한다.

$$\frac{\partial L}{\partial p_k} = \sum_{j=1}^{n} \frac{g_k(x_j;\Theta_k)}{f(x_j)} - \lambda = 0 \quad \cdots\cdots (12)$$

$$\frac{\partial L}{\partial \Theta_{ik}} = \sum_{j=1}^{n} p_k \frac{\partial\, g_k(x_j;\Theta_k)/\partial\Theta_{ik}}{f(x_j)} = 0 \quad \cdots\cdots (13)$$

한편, 베이즈의 정리로부터

$\Pr(k \mid x_j) = \frac{p_k g_k(x_j;\Theta)}{f(x_j)}$ 이 성립하고, Lagrange 상수 λ는 n으로 유도된다. 이들의 관계로부터 다음의 두 가지 식을 도출할 수 있으며 이 두 식은 각각 EM알고리즘의 E-단계와 M-단계에 해당한다.

E-Step : $$\widehat{p_k} = \frac{1}{n} \sum_{j=1}^{n} \Pr(k \mid x_j) \quad \cdots\cdots (14)$$

M-Step:

$$\sum_{j=1}^{n} \Pr(k \mid x_j) \frac{\partial\ \log\ g_k(x_j;\Theta_k)}{\partial\Theta_{ik}} = 0 \quad \cdots\cdots (15)$$

따라서, 위의 두 식의 반복적인 적용을 통해 파라메타 추정과정의 수행이 가능하게 된다.

Ⅳ. 복합확률분포의 파라메타 추정방법

본 연구에서는 두 개의 단순한 정규분포가 복합확률분포를 이루는 경우를 사례로 하여 파라메타 추정 방법을 제시하고자 한다.

만약 복합분포를 이루는 개별적인 확률밀도함수가 평균 μ, 분산공분산행렬 Σ을 가지는 다항정규분포(multi-variate normal distribution, $g_i(x;\Sigma_i,\mu_i)$)를 따른다고 한다면, 이들의 복합확률분포함수($f(x;p,\Sigma,\mu)$)는 다음의 관련 식으로 표현할 수 있다. 아래에 표기된 복합확률 분포와 관련된 수식은 Everitt and Hand(1981)을 통하여 재구성하였다.

$$f(x;p,\Sigma,\mu) = \sum_{i=1}^{c} p_i g_i(x;\Sigma_i,\mu_i) \quad \cdots\cdots (16)$$

여기서, $p=(p_1,p_2,\ldots p_{c-1})$이고 $0 < p_i < 1$,

$$p_c = 1 - \sum_{i=1}^{c-1} p_i \quad (17)$$

$$g_i(x;\Sigma_i,\mu_i) = (2\pi)^{-d/2} \mid \Sigma_i \mid^{-1/2} \exp[-1/2 \times (x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)] \quad \cdots\cdots (18)$$

분산 - 공분산(variance-covariance)행렬 : $\Sigma = \{\Sigma_1, \Sigma_2, \ldots, \Sigma_c\}$

평균 : $\mu = \{\mu_1, \mu_2, \ldots, \mu_c\}$

한편, 위에서 표시한 다항정규분포는 다음과 같은 확률밀도함수로 표현되는 단일 변수에 의한 평균(μ)과 표준편차(σ)를 파라메타로 가지는 정규분포(normal distribution, N)가 다수의 변수를 가지는 경우로 확장된 관계로 이해할 수 있다.

$$N(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu)^2/\sigma^2\right] \quad \cdots\cdots (19)$$

본 연구에서는 위와 같이 두 개의 단일 정규분포가 각각 p_i의 결합비율로 복합된 복합확률분포에 대해 EM 알고리즘을 적용하여 파라메타를 추정하였으며, 이 과정을 통해 위에서 설명한 EM 관계식이 다음과 같이 적용되었다.

E-Step

$$P(s \mid x_i) = \frac{p_s\, g_s(x_i;\Sigma_s,\mu_s)}{f(x_i;\, p,\, \Sigma,\mu)} \quad \cdots\cdots (20)$$

M-Step

$$\widehat{p_k}=\frac{1}{n}\sum_{i=1}^{n}\widehat{P}(k \mid x_i),\ k=1,\ldots,c-1 \quad \cdots\cdots (21)$$

$$\widehat{\mu_k}=\frac{1}{n\,\widehat{p_k}}\sum_{i=1}^{n}\widehat{P}(k \mid x_i)\,x_i\ ,\quad k=1,\ldots,c \quad \cdots\cdots (22)$$

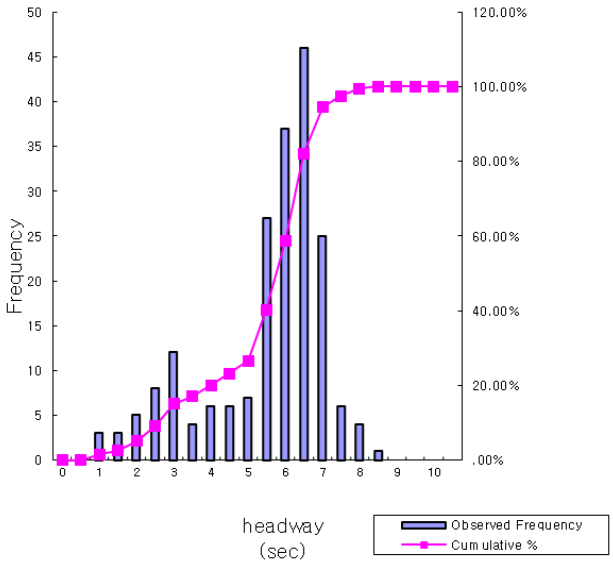
$$\widehat{\Sigma_k}=\frac{1}{n\,\widehat{p_k}}\sum_{i=1}^{n}\widehat{P}(k \mid x_i)(x_i-\widehat{\mu_k})(x_i-\widehat{\mu_k})'\quad k=1,\ldots,c \quad \cdots\cdots (23)$$

한편, 본 연구에서는 MATLAB 프로그램을 사용하여 이 EM 관계식들이 적용된 프로그램을 작성하여 두 개의 정규분포에 대한 복합확률분포의 파라메타 추정 과정을 수행하였다(Kim, 2001).

Ⅴ. 차두시간 조사 및 분석결과

1. 조사의 내용

본 연구의 적용 사례 분석을 위한 조사를 양양군 내 7번 국도 상에서 2003년 5월에 실시하였으며, 수동 스톱워치로 0.1초 단위로 1시간 정도 조사하여 200개 차량의 도착 차두시간(time headway)를 수집하였고 이에 대한 조사결과를 <Table 1>과 <Fig. 1>에 제시하였다. 조사지점은 단속류인 국도 7호선(왕복 4차로)의 송현교차로 - 상운교차로 사이 지점(상양혈리 부근)에서 조사하였고, 두 교차로간 간격이 5km 이상으로 길어 거의 연속류로 통과되는 지점의 특성을 가진다. 구간 중간에 양양국제공항으로 진출입하는 입체교차로인 손양입체교차와 8군단 진입교차로가 있으나 차량 진행에 영향이 없을 것으로 판단된다. 조사방법 및 시간은 수동 스톱워치로 0.1초 단위로 약 1시간 정도 조사하였다.



<Fig. 1> Observed Vehicle Time Headway Distribution Data

<Table 1> Statistics of Observed Headway

Statistics	
Mean	5.273225
Standard Error	0.113384
Median	5.78005
Mode	5.6057
STD Deviation	1.603485
Variance	2.571165
Kurtosis	0.370674
Skewness	-1.09303
Range	7.4424
Min.	0.8437
Max.	8.2861
Sum	1054.645
Number of Data	200

위의 조사 자료를 살펴보면 대략 2.5초 ~ 3초 사이에 하나의 자료 밀집 구간이 있고 6초 ~ 6.5초 부근에 또 다른 자료의 밀집 구간이 나타나고 있어 단일의 확률분포로서는 설명이 어려울 것으로 판단된다. 이러한 복합적 분포를 가지는 차두시간의 관측 결과는 일반적인 상황은 아니며 조사지점에서 조사 당시의 특수한 상황에 따른 관측결과일 수도 있다고 생각된다. 그러나, 연속류의 통행 특성에 유사하며 통행차량이 많지 않았던 본 조사지점과 같은 경우에서 이러한 복합분포를 가지는 자료가 나타날 가능성은 충분히 있을 수 있다고 판단된다. 다만, 이 결과가 일반적인 상황이라고는 판단하기에는 아직 좀 더 많은 자료로 보완할 필요가 있다고 생각되며, 이러한 복합분포를 보이는 차두시간의 일반적 도착분포에 대한 규명은 추후 연구대상으로

하고, 본 연구에서는 이러한 다소 특이한 차두시간 조사 결과에 대해서도 차두시간 확률분포를 추정할 수 있도록 하는 방법인 EM 알고리즘의 적용 가능성에 대해서만 연구의 초점을 맞추고자 한다.

2. 추정 결과

본 연구에서는 위의 조사 자료에 대해 두개의 정규분포가 복합된 복합 확률분포를 적용하여 EM 알고리즘을 통하여 관련 파라메타를 추정하였다. 다음의 <Table 2>에는 각 반복 수행에 따른 파라메타 추정 결과가 제시되어 있다. 추정과정의 수행을 위해 적용한 초기값은 평균차두시간에 대해서 각각 3초와 6초, 분산에 대해서는 모두 1.0을 적용하였으며 두 개의 정규분포 복합비율에 대해서는 초기값으로 0.3과 0.7을 사용하였다.

초기값의 결정은 이후 탐색과정에서의 수렴성에 영향을 줄 수 있기 때문에 신중하게 결정해야 하며 본 연구에서는 조사된 자료로부터 임의(random)의 직관적인 기준으로 결정하였기 때문에 분석에 영향을 줄일 수 있는 방법으로 결정하였다.

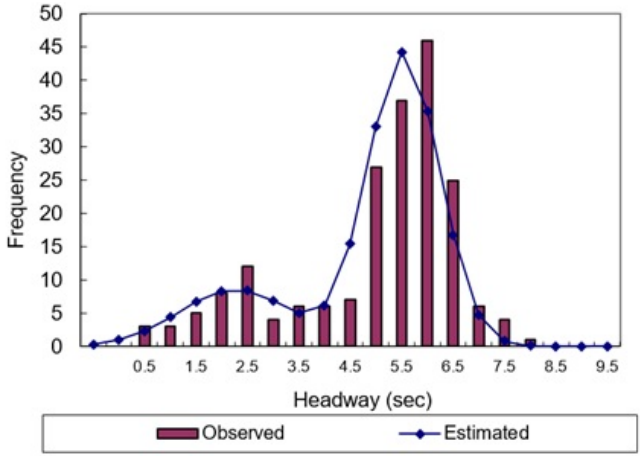
<Table 2>를 통해 보면 EM 알고리즘은 모두 22회의 반복수행과정을 통해 두개의 정규분포 파라메타가 추정되었으며, 첫 번째 정규분포는 평균차두시간 2.7626과 분산 1.1727의 추정 결과로 나타났고, 두 번째의 정규분포에 대해서는 평균차두시간 6.0371초이며 분산이 0.4784로 추정되었다. 한편 두 개의 정규분포가 복합되는 비율은 0.2333과 0.7667인 것으로 추정되었다. 따라서 위의 자료에 대한 복합확률분포의 확률밀도함수는 다음의 식으로 표현될 수 있다.

$$f(x) = 0.2333 \times N(x\ ; 2.7626,\ 1.1727) + 0.7667 \times N(x\ : 6.0371, 0.4784) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\ (24)$$

<Table 2> Result of EM Iteration for the Estimation of the Mixture Distribution

Iteration	Approach Value	Mean 1	Mean 2	Var. 1	Var. 2	Ratio of 1
0	-	3.0	6.0	1.0	1.0	0.3
1	9.0274	2.0082	6.6725	0.9227	0.5944	0.2225
2	9.2329	2.6334	6.0288	1.2845	0.8872	0.2225
3	0.3507	2.5769	6.0447	1.0030	0.5413	0.2169
4	0.1619	2.7389	5.9750	1.0635	0.4974	0.2225
5	0.0527	2.7610	5.9922	1.1162	0.4928	0.2268
6	0.0249	2.7600	6.0104	1.1411	0.4879	0.2293
7	0.0125	2.7610	6.0205	1.1536	0.4841	0.2308
8	0.0074	2.7617	6.0268	1.1610	0.4818	0.2317
9	0.0045	2.7621	6.0307	1.1655	0.4804	0.2323
10	0.0029	2.7623	6.0332	1.1683	0.4796	0.2327
11	0.0018	2.7624	6.0347	1.1700	0.4791	0.2329
12	0.0011	2.7624	6.0356	1.1710	0.4788	0.2331
13	0.0007	2.7625	6.0362	1.1716	0.4787	0.2331
14	0.0004	2.7625	6.0366	1.1720	0.4786	0.2332
15	0.0003	2.7625	6.0368	1.1723	0.4785	0.2332
16	0.0002	2.7625	6.0369	1.1724	0.4784	0.2333
17	0.0001	2.7625	6.0370	1.1725	0.4784	0.2333
18	0.0001	2.7625	6.0371	1.1726	0.4784	0.2333
29	0.0000	2.7626	6.0371	1.1726	0.4784	0.2333
20	0.0000	2.7626	6.0371	1.1726	0.4784	0.2333
21	0.0000	2.7626	6.0371	1.1727	0.4784	0.2333
22	0.0000	2.7626	6.0371	1.1727	0.4784	0.2333

<Fig. 2>는 관측자료와 파라메타 추정결과를 통한 복합확률분포를 함께 표현하고 있다.



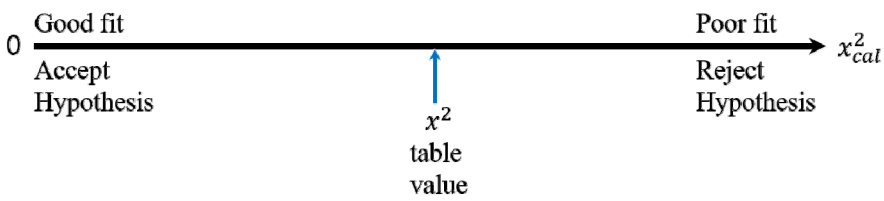
<Fig. 2> Estimated Result of the Mixture Distribution Compared to the Observed Data

3. 적합도 검정

예측자료의 분포가 관측자료의 분포를 얼마나 잘 설명하고 동질한 지에 대한 적합도 검정이 필요하여 본 연구에서는 가장 대표적인 적합도 검정방법인 χ^2 검정을 사용하여 복합확률분포가 관측분포와 얼마나 적합한 지를 수행하였다. χ^2 검정은 다음과 같이 χ^2 값을 계산하여 그 값의 크기에 따라 예측분포가 관측분포와 얼마나 유사한 동질성을 가지는지를 평가할 수 있는 통계학적 방법이다.

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \cdots\cdots (25)$$

각 차두시간별 관측돗수와 예측돗수가 정확하게 같으면 χ^2 값은 0으로 계산되고 이는 2개의 분포가 완전히 동질한 분포라는 것을 나타낸다. 그리고 연구가설은 「관측된 분포와 예측분포 사이의 차이는 없다」이고 χ^2 값이 작을수록 관측돗수와 예측돗수의 차이가 작아져서 유의수준에 따른 χ^2 기준치보다 작으면 연구가설은 채택되고 관측분포와 예측분포는 거의 유사한 분포를 나타내는 것을 의미한다.



<Fig. 3> Comparing Chi-Square Calculated and Table Values

<Table 3>에는 예측된 분포의 관측자료에 대한 설명력을 의미하는 적합도 검정결과를 표시하고 있다. 이 표에서 보듯이 복합확률분포의 χ^2값은 30.43으로 계산되어 유의수준 5%의 χ^2값 28.9에 의해서는 기각되었으나 유의수준 1%의 χ^2값 34.8에 대해서는 채택되는 것으로 나타나 유의수준 1%에서 복합확률분포의 설명력이 있는 것으로 나타났다. 한편, <Table 1>에 나타난 조사자료의 통계값을 이용하여 단일 정규분포에 대한 검정결과는 χ^2값이 102.19로 나타나 어떠한 경우에도 채택되지 못하여 전혀 설명력이 없는 분포로 판단된다. 즉, 단일 확률분포로서는 주어진 자료를 설명하기가 어렵다고 판단된다. 한편, Pearson TypeⅢ 분포에 대한 검정과정도 수행하였으나 검정결과에서 역시 적합도를 만족하지 못하는 것으로 나타났다.

<Table 3> Goodness-of-Fit Test for the Estimated Mixture Distribution

Time Headway	Frequency	Mixture distribution		Single Normal distribution	
		Estimated frequency	χ^2	Estimated frequency	χ^2
0	0	0	0.00	0	0.00
~ 0.5	0	1	1.00	0	0.00
~ 1	3	2	0.50	1	4.00
~ 1.5	3	4	0.25	1	4.00
~ 2	5	7	0.57	2	4.50
~ 2.5	8	8	0.00	4	4.00
~ 3	12	8	2.00	7	3.57
~ 3.5	4	7	1.29	11	4.45
~ 4	6	5	0.20	16	6.25
~ 4.5	6	4	1.00	20	9.80
~ 5	7	5	0.80	23	11.13
~ 5.5	27	14	12.07	25	0.16
~ 6	37	33	0.48	24	7.04
~ 6.5	46	44	0.09	21	29.76
~ 7	25	35	2.86	18	2.72
~ 7.5	6	17	7.12	10	1.60
~ 8	4	5	0.20	8	2.00
~ 8.5	1	1	0.00	5	3.20
~ 9	0	0	0.00	2	2.20
~ 9.5	0	0	0.00	2	2.20
~ 10	0	0	0.00	0	0.00
Sum	200	200	30.43	200	102.19
χ^2, 0.05 = 28.9, χ^2, 0.01 = 34.8					

Ⅵ. 결론 및 향후 연구

본 연구에서는 이상의 내용을 통하여 최우추정법의 한 분야인 EM 알고리즘이 교통분야에 적용가능한지 소개하고 교통류에서 확률분포를 통하여 설명하는 차량의 도착 차두시간 분포에 대해 한 예로서 두 가지 단순 정규분포가 조합된 복합확률분포에 대하여 EM 알고리즘을 통하여 관련 파라메타를 추정하고 실제 차두시간 분포를 잘 모사할 수 있는지를 조사, 분석하였다.

분석결과 일반적으로 차두시간의 표현에 이용되는 기존의 단일확률분포로서는 설명이 어려운 경우에도 EM 알고리즘을 통해 추정된 파라메타를 이용한 복합확률분포는 주어진 실제 교통자료의 차두시간 분포와

거의 유사하게 모사하였고 유의수준 1%에서 적합도 검정(χ^2 test)도 신뢰성이 확보되어 설명력이 높은 것으로 나타났다. 따라서 본 논문에서 이용된 차량도착 차두시간분포의 경우 이외에도 기존에 사용되던 단일 확률분포로서는 설명이 어려운 경우에서 EM 알고리즘의 적용을 시도해 볼 수 있을 것으로 판단된다.

기존 두 가지의 차두시간 분포를 조합하는 복합확률분포의 경우, 차량군과 비차량군의 비율, 평균과 표준편차와 같은 통계치를 사용하여 복합확률분포를 표현하였는데 이는 조합하는 단일 확률분포의 선택과 이러한 분포의 통계치 결과에 따라 복합확률분포의 형태가 결정되고 적합성이 많이 좌우된다.

본 연구에서는 반복계산을 통한 복합확률분포의 파라메타를 추정하는 방식이므로 선택되는 단일분포에 어느 정도 독립적이고 본 연구에서 조사된 차두시간 자료처럼 2가지 분포가 매우 특이한 형태를 가진 복합분포도 잘 묘사하는 것이 특징이라 할 수 있다.

본 연구에서는 교통분야에서의 EM 알고리즘의 적용성을 평가하고자 단순한 두 개의 정규분포를 이용하는 복합확률분포를 가지고 실제 교통자료의 차두시간 분포를 설명하는 것을 대상으로 연구되었으나, 보다 다양한 복합확률분포에 대해서도 이 알고리즘의 검증이 필요하다고 판단되고, 또한 원래 EM 알고리즘이 가지고 있는 장점 중의 하나인 불충분한 자료(incomplete data 또는 missing data)에 대한 알고리즘의 적용성에 대해서도 실증적인 추가평가가 필요하다고 생각된다.

ACKNOWLEDGEMENTS

본 논문은 대한교통학회의 제44회 추계학술발표회에 게재되었던 논문을 수정·보완하여 작성하였습니다.

REFERENCES

Dempster, A. P., Laird, N. M. and Rubin, D. B.(1977), “Maximum Likelihood from Incomplete Data via EM Algorithm”, *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1. pp.1-22.

Everitt, B. S. and Hand, D. J.(1981), *Finite Mixture Distributions*, Chapman and Hall, pp.4, 9, 25.

Jeong, S. I., Kim, B. S., Park, S. G. and Yoo, Y. G.(1999), *Probability and Statistics for Engineers and Scientists*, Cheongmoon-Gak Publications, p.306. (in Korean)

Kim, C. G.(2001), *MATLAB How to Use and It's Applications*, Kyowoo-Sa Publications. (in Korean)

Kim, W. C., Kim, J. J., Park, B. W., Park, S. H., Song, M. S., Lee, S. Y., Lee, Y. J., Jeon, J. W. and Cho, S. S.(1998), *Modern Statistics* (4th ed.), Youngji Cultural Publication, p.279. (in Korean)

May, A. D.(1990), *Traffic Flow Fundamentals*, Prentice Hall, pp.14-35.

McLachlan, G. J. and Krishnan, T.(1997), *The EM Algorithm and Extensions*, John Wiley & Sons.