The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 795-804

Estimation of the OD Traffic Intensities in Dynamic Routing Network: Routing-Independent Tomography[1)]

Seung-Gu Kim[2)]

Abstract

In this article, a tomography for the estimation of the origin-destination(OD) traffic intensities in dynamic routing network is considered. Vardi(1996)'s approach based on fixed route is not directly applicable to dynamic routing protocols, which arises from the fact that we cannot access the route at every observation time. While it uses link-wise traffics as the observations, the proposed method considers the triple of ingress/outgress/relayed traffics data at each node so that we can transform the problem into a routing-independent tomography. An EM algorithm for implementation and some simulated experiments are provided.

Keywords : network tomography, dynamic routing, EM algorithm.

1. 서론

네트워크 토모그래피(Network Tomography)는 Y. Vardi(1996)에 의해 처음으로 소개되었다. 라우팅 행렬(routing matrix)로서 출발지와 목적지 사이의 경로가 사전에 규정된 네트워크 토폴로지(network topology)를 표현하여 링크(link) 단위의 트래픽 자료(traffic data)의 관측을 바탕으로 OD(origin-destination) 트래픽 밀도(traffic intensity)들을 추정하였다. 이 문제는 의료영상에서 Y. Vardi(1984)가 개발하였던 토모그래피와 아주 흡사하여 "네트워크 토모그래피"란 명칭을 사용하였다.

네트워크 토모그래피는 인터넷의 확산과 더불어 극히 최근에야 주목을 받기 시작하여 지금은 다양한 목적에 적합한 토모그래피가 여러 각도에서 연구되고 있는데, 크게 세 방향으로 분류할 수 있다. 첫째는 Y. Vardi(1996)을 시작으로 한 C. Tebaldi and M. West(1998), J. Cao 등(2000) 및 G. Laing and B. Yu(2003)에 의해 연구된 OD 트래픽 밀도 추정문제이며, 둘째는 M. Coates and R. Nowak(2001)에 의해 연구된 링크 노드(link nodes)들 사이의 지연 특성(특히 지연분포 delay distributions) 추정 문제, 마지막으로 R. Castro 등(2003)의 토폴로지 식별(topology identification) 문제라 할 수 있다. 이 세 가지 분야는 모두 네트워크에서 간접 관측된 자료를 사용하여 네트워크 내부 특성을 탐구한다는 목적을 가지고 있다. 이 중에 본 연구의 관심사는 OD 트래픽 밀도 추

1) This research was supported by Sangji University Research Fund, 2002.

2) Associate professor, Department of Applied Statistics, Sangji University, 220-702
E-mail : sgukim@mail.sangji.ac.kr

정문제이다.

불과 10년 전만 하더라도 인터넷 망은 상대적으로 간단하여 라우터는 OD간의 경로를 거의 정적(static or fixed)으로 선정하였지만 최근에는 네트워크 규모가 방대해짐에 따라 동적으로 경로가 설정되는 동적 라우터(dynamic router)가 보편적으로 사용되고 있다. 따라서 링크 사이의 라우팅 토폴로지가 고정되어 있거나, 그렇지 않으면 링크 상의 고정된 확률정보가 주어졌을 때에 만 가능한 고정 라우팅 토모그래피는 문제에 직면하게 된다. 왜냐하면 인터넷 상의 네트워크가 동적 라우팅 관계를 가질 때 임의의 관측시점에서 라우팅 정보를 얻는 것은 불가능하며, 오랜 관측으로부터 고정된 링크확률을 얻는다는 것은 실제 가동 중인 라우터의 부하를 과중하게 하므로 현실적으로 어렵기 때문이다.

본 연구에서는 이 문제를 해결하기 위해 네트워크의 토폴로지를 정의하는 라우팅 행렬(link routing matrix)에 의존하지 않는 토모그래피 기법을 제안한다. 기존의 방법과 제안된 방법의 큰 차이는 사용된 관측치에 있는데, 기존의 방법이 라우팅에 종속된 링크 단위의 트래픽 자료를 관측치로 사용하는 반면, 제안된 방법은 라우팅 형태에 무관한 노드의 입력(ingress), 출력(outgress) 트래픽 및 경유 트래픽(relayed traffic)을 관측치로 사용한다. 현실적으로 어떤 시점에서 각 노드의 경유 트래픽의 측정은 라우터의 성능에 큰 지장을 초래하지 않고 처리될 수 있다. "라우팅 무관(routing-independent) 토모그래피"란 1) 네트워크 내의 노드들을 연결하는 링크들에 대해 고정된 트래픽 관측치를 사용하지 않고, 2) 링크의 관측치 보다는 노드 별 관측치를 한번이 아닌 일정 횟수에 걸쳐 얻어지는 경우의 토모그래피를 지칭하는 의미로 사용한다

한편 본 연구에서는 망 내의 지연이나 손실은 무시할 정도로 작다고 가정한다. 다음 장에서는 네트워크 토모그래피의 기본개념을 소개하며, 3-4 장에서는 제안된 방법과 EM 알고리즘에 의한 토모그래피 기법을 실었고, 5 장에서는 동적 라우팅을 따르는 가상의 네트워크를 만들어 제안된 기법에 대한 모의실험 결과를 제공하였다. 마지막으로 6 장에서 결론과 더불어 제안된 기법의 제한점을 논의하였다.

2. 기본개념과 고정 라우팅 네트워크 토모그래피 소개

네트워크는 두 개 이상의 노드(node)와 각 노드를 연결하고 있는 링크(link)들로 구성된다. 링크는 메시지의 전송로를 의미하며, 본 논문에서 각 노드는 단말기(sink node) 기능과 메시지의 경유가 가능한 서브네트워크(sub-network) 혹은 라우터(router) 기능 등을 모두 수행하는 장치를 의미한다. 한편 모든 메시지는 출발지인 원점노드와 도착지인 종점노드를 갖는다고 가정한다.

m 개의 노드로 구성된 네트워크에서 $X_{j_1j_2}\,(j_1, j_2=1,\cdots,m; j_1\neq j_2)$를 원점노드 j_1에서 종점노드 j_2로 전송된 OD 트래픽의 양이라 하자. 트래픽의 크기는 패킷(packets)의 개수 혹은 바이트 카운트(byte counts)로 측정하므로 $X_{j_1j_2}\geq 0$을 만족하며, $X_{j_1j_2}\sim$ Poisson$(\lambda_{j_1j_2})$를 따른다고 가정한다. 그리고 n 개의 각 링크에서 관측된 링크 트래픽 $\boldsymbol{U}=(U_1, U_2,\cdots, U_n)^T$은 미관측 자료 $\boldsymbol{X}=(X_{12}, X_{13},\cdots, X_{p,p-1})^T$와

$$\boldsymbol{U}=\boldsymbol{AX}\ ,\ \text{단}\quad \boldsymbol{X}\geq \boldsymbol{0} \tag{2.1}$$

의 관계를 갖는다. 여기서 $n\times M(=m(m-1))$ 크기의 행렬 $\boldsymbol{A}=a_{ij}$는 라우팅 행렬로서 토폴로

지를 정의하고 있다. 즉, a_{ij}는 주소가 $j=(j_1, j_2)$인 OD 트래픽이 i 번째 링크를 경유하면 1, 그렇지 않으면 0을 나타낸다. 예제 1.1에서 간단한 네트워크와 라우팅 행렬을 사용하여 상세히 설명하였다.

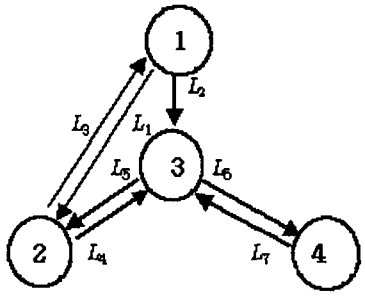

그림 1.1 Vardi의 네트워크

예제 1.1 그림 1은 Y. Vardi(1996)에서 소개된 것으로서, 4 개의 노드(1-4번)와 7 개의 (유향)링크 $L_1 \sim L_7$으로 구성된 네트워크이다. 그리고 라우팅 행렬을

$A=$

	(1,2)	(1,3)	(1,4)	(2,1)	(2,3)	(2,4)	(3,1)	(3,2)	(3,4)	(4,1)	(4,2)	(4,3)
$L_1(1\rightarrow 2)$	1	0	0	0	0	0	0	0	0	0	0	0
$L_2(1\rightarrow 3)$	0	1	1	0	0	1	0	0	0	0	0	0
$L_3(2\rightarrow 1)$	0	0	0	1	0	1	1	0	0	1	0	0
$L_4(2\rightarrow 3)$	0	0	0	0	1	0	0	0	0	0	0	0
$L_5(3\rightarrow 2)$	0	0	0	0	0	0	1	1	0	1	1	0
$L_6(3\rightarrow 4)$	0	0	1	0	0	1	0	0	1	0	0	0
$L_7(4\rightarrow 3)$	0	0	0	0	0	0	0	0	0	1	1	1

과 같이 정의하고 잇다. 행렬 A의 행은 7 개의 링크를 나타내고 있고, 열은 12 개의 OD 트래픽 주소를 표현하고 있다. 예를 들어 두 번째 행(L_2)은 노드 1에서 노드 3를 연결하고 있는 링크를 나타내는데, 이 링크를 경유하는 OD는 (1,3), (1,4), (2,4) 임을 의미하고 있다. 그리고 6 번째 열은 OD의 주소가 (2, 4) 즉 원점인 노드 2에서 종점인 노드 4로 가는 경로가 $2\rightarrow 1\rightarrow 3\rightarrow 4$ ($L_3\rightarrow L_2\rightarrow L_6$) 임을 나타낸다. 이와 같이 라우팅 행렬은 네트워크의 경로정보를 제공하고 있다. □

망 내에 지연이나 트래픽 손실 등이 없다고 가정한다면, OD 트래픽 X가 주어지면 라우팅 행렬 A에 의해 링크 트래픽 U가 결정된다. OD 트래픽 토모그래피의 목적은 K 번 반복 관측된 링크 트래픽 자료 $(U^{(1)T}, U^{(2)T}, \cdots, U^{(K)T})^T$를 사용하여 OD 트래픽 밀도 $\lambda=(\lambda_{12}, \lambda_{13}, \cdots, \lambda_{m,m-1})^T$를 비음의 제약(non-negative constraints) 하에서 추정하는 것이다. 물

론 각 노드에는 OD 주소 (j_1, j_2)를 식별하여 $X_{j_1 j_2}$를 집계하는 어떠한 장치도 갖고 있지 않다고 가정한다(사실 집계 장치가 있다면 토모그래피를 할 이유가 없다).

Y. Vardi(1996)과 같은 기존의 기법은 관측 시점 $k(=1,\cdots,K)$에서 기지인 라우팅 행렬 $\boldsymbol{A}$가 고정되어 있다는 가정 하에 OD 트래픽 밀도 $\underline{\lambda}$를 추정한다. 그런데 현실에서는 다음과 같은 문제점들이 있다. 첫째, 관측시점에서 가동중인 노드로부터 트래픽을 수집하는 것이 노드의 본래의 전달 및 교환 기능 상당히 저하시킬 수 있기 때문에 라우팅 행렬 $\boldsymbol{A}$을 정확히 안다는 것은 매우 어렵다. 둘째, 라우팅이 동적일 경우 모든 관측 시점에 대해 $\boldsymbol{A}^{(1)}=\boldsymbol{A}^{(2)}=\cdots=\boldsymbol{A}^{(K)}=\boldsymbol{A}$으로 고정되어 있지 않다. 이 경우 관측 시점 k에 대해 라우팅 행렬 $\boldsymbol{A}^{(k)}$의 상태방정식(state equation)을 파악하여 문제를 해결하는 방식을 생각해 볼 수는 있으나 다양하고 복잡한 인터넷 환경 하에서는 거의 현실성이 없다.

본 연구에서는 링크 단위의 자료 대신 각 노드에서 관측되는 입출력 트래픽과 경유 트래픽을 측정함으로써, 네트워크 내부의 경로 정보를 고려하지 않는 방법 다시 말해서 라우팅 행렬 $\boldsymbol{A}$에 무관한 토모그래피를 도입하는 것이 이 문제를 해결하는 최상의 방법이라 판단하여, 이를 다음 장에서 다룬다.

3. 노드 트래픽 관측모형

먼저 표기의 단순성을 위해 잠시 관측시점을 나타내는 위첨자 (k)를 생략하기로 한다. 노드 $j(=1,\cdots,m)$에서 관측되는 트래픽을 입/출력을 고려하여 Y_j^I, Y_j^O로 나타내자. 앞 장의 U_i가 고정되어 있는 각각의 개별 링크에서 측정한 트래픽인 반면, Y_j는 링크의 연결상태를 고려하지 않고 단지 노드 j에 입력되거나 출력되는 트래픽을 측정한 것이다. 그래서 Y_j^I나 Y_j^O는 노드 j에 연결된 모든 링크에 대한 U_i의 합계로 생각할 수 있다. 그리고 Y_j^I는 종점이 j인 트래픽(X_j^I)과 노드 j를 경유하는 트래픽(R_j^I)의 합으로 구성되어 있고, 마찬가지로 Y_j^O는 원점이 j인 트래픽(X_j^O) 그리고 노드 j를 경유하는 트래픽 (R_j^O)의 합이 된다. 지연이 무시할 정도로 작다고 가정하면 $R_j^I=R_j^O=R_j$이므로 노드 j에서 (Y_j^I, Y_j^O, R_j)를 관측하게 된다. 그래서 $\mathcal{L}_j^I$와 $\mathcal{L}_j^O$를 각각 입력과 출력 시 노드 j에 연결된 모든 링크의 집합이라 할 때,

$$Y_j^I=\sum_{i\in\mathcal{L}_j^I}U_i=X_j^I+R_j=\sum_{h\neq j}X_{hj}+R_j \tag{3.1}$$

≡ (종점이 노드 j인 OD 트래픽 총계)+(노드 j를 경유하는 트래픽)

$$Y_j^O=\sum_{i\in\mathcal{L}_j^O}U_i=X_j^O+R_j=\sum_{h\neq j}X_{jh}+R_j \tag{3.2}$$

≡ (원점이 노드 j인 OD 트래픽 총계)+(노드 j를 경유하는 트래픽)

의 관계를 갖는다. 만약 링크들이 고정 라우팅에 의해 정의된다면, 경유 트래픽 R_j는 X_{ij}들의 선형식으로 정확히 표현할 수 있다. 예를 들어, 예제 1.1에서 노드 2에 입력/출력되는 트래픽은 각각

$$Y_2^I = U_1 + U_5 = (X_{12} + X_{32} + X_{42}) + (X_{31} + X_{41}) = X_2^I + R_2$$

$$Y_2^O = U_2 + U_3 = (X_{21} + X_{23} + X_{24}) + (X_{31} + X_{41}) = X_2^O + R_2$$

이므로, $R_2 = X_{31} + X_{41}$으로 결정된다. 물론 라우팅이 고정되어 있지 않은 경우 R_2가 항상 X_{31}과 X_{41}으로 구성된다고 확신할 수 없다. 그렇지만 X_2^I와 X_2^O의 구성요소는 라우팅에 관계없이 항상 고정될 것이다. 다시말해서, 노드 2의 입출력 트래픽 Y_2^I, Y_2^O는 라우터의 변이와 무관한 성분인 노드 트래픽 X_2^I, X_2^O와 라우팅에 따라 변이하는 성분인 경유 트래픽 R_2로 구성된다.

일반적으로, 관측시점 k에서 어떤 행합산연산행렬 $\boldsymbol{E}^{(k)}$에 대해

$$\boldsymbol{Y}^{(k)} = \boldsymbol{E}^{(k)}\boldsymbol{U}^{(k)} = \boldsymbol{E}^{(k)}\boldsymbol{A}^{(k)}\boldsymbol{X}^{(k)} = \boldsymbol{H}\boldsymbol{X}^{(k)} + \boldsymbol{B}^{(k)}\boldsymbol{X}^{(k)} \tag{3.3}$$

혹은

$$\boldsymbol{Y}^{(k)} = \boldsymbol{H}\boldsymbol{X}^{(k)} + \boldsymbol{R}^{(k)}, \quad k = 1, \cdots, K \tag{3.4}$$

의 관계가 성립한다. 단, $\boldsymbol{R}^{(k)} = \boldsymbol{B}^{(k)}\boldsymbol{X}^{(k)}$이며, $\boldsymbol{B}^{(k)}$는 $\boldsymbol{X}^{(k)}$로부터 $\boldsymbol{R}^{(k)}$로의 선형변환을 나타낸다. 그리고 $\boldsymbol{H}$는 $2m \times m(m-1)$ 크기의 행렬로서 0 혹은 1을 원소로 하는데, 각 행은 입력 혹은 출력 트래픽의 성분을 1로써 나타내고 있다. 앞으로 $\boldsymbol{H}$를 "I/O 트래픽 행렬" 그리고 $\boldsymbol{B}$를 "경유 트래픽 행렬"이라 부르겠다. 식 (3.3)의 관계는 임의적이므로 라우팅 행렬 $\boldsymbol{A}^{(k)}$에 대해 행렬 $\boldsymbol{H}$와 $\boldsymbol{B}^{(k)}$는 항상 존재하며, 식 (3.4)는 관측시점 k에 종속된 관측 트래픽 $\boldsymbol{U}^{(k)}$가 선형변환 $\boldsymbol{E}^{(k)}$을 통해 $\boldsymbol{A}^{(k)}$에 무관한 I/O 트래픽 $\boldsymbol{H}\boldsymbol{X}^{(k)}$ 성분과 $\boldsymbol{A}^{(k)}$에 종속된 경유 트래픽 $\boldsymbol{R}^{(k)}$ 성분의 합으로 나타낼 수 있음을 말해주고 있다. 이 것은 곧 만약 노드 트래픽 $\boldsymbol{Y}^{(k)}$와 경유 트래픽 $\boldsymbol{R}^{(k)}$를 관측할 수 있다면, 라우팅 행렬 $\boldsymbol{A}^{(k)}$에 관계없이 고정된 I/O 트래픽 행렬 $\boldsymbol{H}$에 대한 관측모형 $\boldsymbol{Y}^{(k)} - \boldsymbol{R}^{(k)} = \boldsymbol{H}\boldsymbol{X}^{(k)}$을 바탕으로 한 토모그래피가 가능함을 의미한다.

예제 1.2 예제 1.1의 라우팅 행렬 $\boldsymbol{A}$에 대한 I/O 트래픽 행렬과 경유 트래픽 행렬은 각각

$\boldsymbol{H} =$

	(1,2)	(1,3)	(1,4)	(2,1)	(2,3)	(2,4)	(3,1)	(3,2)	(3,4)	(4,1)	(4,2)	(4,3)
L_1^I	0	0	0	1	0	0	1	0	0	1	0	0
L_1^O	1	1	1	0	0	0	0	0	0	0	0	0
L_2^I	1	0	0	0	0	0	0	1	0	0	1	0
L_2^O	0	0	0	1	1	1	0	0	0	0	0	0
L_3^I	0	1	0	0	1	0	0	0	0	0	0	1
L_3^O	0	0	0	0	0	0	1	1	1	0	0	0
L_4^I	0	0	1	0	0	1	0	0	1	0	0	0
L_4^O	0	0	0	0	0	0	0	0	0	1	1	1

$B=$

	(1,2)	(1,3)	(1,4)	(2,1)	(2,3)	(2,4)	(3,1)	(3,2)	(3,4)	(4,1)	(4,2)	(4,3)
L_1^I	0	0	0	0	0	1	0	0	0	0	0	0
L_1^O	0	0	0	0	0	1	0	0	0	0	0	0
L_2^I	0	0	0	0	0	0	1	0	0	1	0	0
L_2^O	0	0	0	0	0	0	1	0	0	1	0	0
L_3^I	0	0	1	0	0	1	0	0	0	1	1	0
L_3^O	0	0	1	0	0	1	0	0	0	1	1	0
L_4^I	0	0	0	0	0	0	0	0	0	0	0	0
L_4^O	0	0	0	0	0	0	0	0	0	0	0	0

과 같다. 행렬 $\boldsymbol{H}$의 각 행과 열은 서로 다르며, 주소 (j_1, j_2)에 대응하는 열에서 행 $L_{j_1}^O$와 $L_{j_2}^I$의 원소만 1이고 나머지는 0이라는 점을 주목하자. 따라서 행 $L_{j_1}^O$는 오직 행 $L_{j_2}^I$와 내적 1을 가지며, 다른 행과의 내적은 0이다. 한편, 이 문제에서는 라우팅이 고정적이므로 관측 시점 k에 관계없이 $\boldsymbol{B}^{(k)}=\boldsymbol{B}$가 된다. ■

이제 문제는 식 (3.4)의 관계로부터 모수인 OD 트래픽 밀도 $\underline{\lambda}$를 식별할 수 있는 지 그리고 의미있는 추정이 가능한 지에 대한 의문으로 귀결된다. 모수 $\underline{\lambda}$에 대한 추정문제는 다음 장에서 다루기로 하고, 먼저 식별가능성을 확인한다.

식 (3.4)의 식별성은 Y. Vardi(1996)가 제공한 정리를 이용하면 간단히 증명할 수 있다. 먼저, $X_{ij}\sim$ Poisson(λ_{ij})를 따를 때, 식 (3.3)의 관계로부터

$$Y_i-R_i\sim \text{Poisson}\left(\sum_j h_{ij}\lambda_j\right)$$

를 따름은 자명하다. 둘째, $\boldsymbol{H}$의 모든 열은 서로 다르다. 왜냐하면 $j=(j_1, j_2)$번째 열은 원점 j_1과 종점 j_2에 대응한 출력과 입력원소만이 1고 나머지는 0이므로 열들은 서로 같지 않다. 셋째, $\boldsymbol{H}$의 행과 열의 개수에 대해 $2^{2m}>m(m-1)+1$을 만족한다. 따라서 모수 벡터 $\underline{\lambda}$는 서로 다른 관측에 대해 식별가능하다.

4. EM 알고리즘

이제 우리의 목적은 관측자료 $(Y_j^{I(k)}, Y_j^{O(k)}, R_j^{(k)})$, $j=1,\cdots,m$; $k=1,\cdots,K$ 가 주어졌을 때, 식 (3.4)의 관계를 바탕으로 비음의 제약 하에서 모수 $\underline{\lambda}$의 최우추정치를 구하는 것이다. 먼저 $Y_j^{I(k)}-R_j^{(k)}(\geq 0)$, $Y_j^{O(k)}-R_j^{(k)}(\geq 0)$를 관측치로 사용할 것이고, 두 자료는 포아송 분포를 따르므로 이 장에서는 일반성을 잃지 않고 $R_j^{(k)}=0$으로 놓기로 하자.

일반적으로 토모그래피에서 로그-우도

$$\sum_{k=1}^{K} \log f_Y(\boldsymbol{y}^{(k)}; \underline{\lambda}) \tag{4.1}$$

로부터 직접 최우추정치 $\hat{\underline{\lambda}}$를 얻는다는 것은 매우 어렵다. 그래서 $\boldsymbol{X}$ 를 완전자료(complete data) 그리고 관측값

$$\boldsymbol{Y} = (Y_1, \cdots, Y_{2m})^T = (Y_1^I, Y_1^O, \cdots, Y_m^I, Y_m^O)^T$$

을 불완전 자료(incomplete data)로서 $\boldsymbol{Y}^{(k)} = \boldsymbol{H}\boldsymbol{X}^{(k)}$의 관계를 구성하는 EM 알고리즘으로부터 추정치를 구하는 것이 일반적이므로, 먼저 이 자료구조 하에서 EM 알고리즘을 구해보자.

$Y_i^{(k)} \sim$ Poisson($\boldsymbol{h}_i, \underline{\lambda}$)을 따른다. 단, $\boldsymbol{h}_i$는 행렬 $\boldsymbol{H}$의 i 번째 행을 나타낸다. 이제 EM 알고리즘은 E-step에서

$$\begin{aligned} Q(\underline{\lambda}|\underline{\lambda}^{(t)}) &= \sum_{k=1}^{K} E\left[\log f_X(\boldsymbol{X}^{(k)}; \underline{\lambda}) | \boldsymbol{y}^{(k)}, \underline{\lambda}^{(t)}\right] \\ &\propto \sum_{k=1}^{K} \sum_{j=1}^{M} \left\{-\lambda_j + E\left[X_j^{(k)} | \boldsymbol{y}^{(k)}, \underline{\lambda}^{(t)}\right] \log \lambda_j\right\} \end{aligned} \tag{4.2}$$

을 구하고, M-step에서

$$\lambda_j^{(t+1)} = \frac{1}{K} \sum_{k=1}^{K} E\left[X_j^{(k)} | \boldsymbol{y}^{(k)}, \underline{\lambda}^{(t)}\right], \quad j = 1, \cdots, M \tag{4.3}$$

를 계산한다. 여기서 조건부 기대값은 첫째, 확률변수 $X_j^{(k)}$은 $\boldsymbol{Y}^{(k)}$의 $2m$ 개의 원소 중에 오직 $(Y_{j_2}^I, Y_{j_1}^O)$에만 종속된다는 사실(예제 1.2 참조)과 D. Karlis(2003)의 다변량 포아송분포에 대한 몇 가지 결과들을 이용하면

$$\begin{aligned} E\left[X_j^{(k)} | \boldsymbol{y}^{(k)}, \underline{\lambda}^{(t)}\right] &= E\left[X_j^{(k)} | y_{j_2}^{I(k)}, y_{j_1}^{O(k)}, \underline{\lambda}^{(t)}\right] \\ &= \lambda_j^{(t)} \frac{p(y_{j_2}^{I(k)} - 1, y_{j_1}^{O(k)} - 1; \mu_0^{(t)}, \mu_1^{(t)}, \mu_2^{(t)})}{p(y_{j_2}^{I(k)}, y_{j_1}^{O(k)}; \mu_0^{(t)}, \mu_1^{(t)}, \mu_2^{(t)})} \end{aligned} \tag{4.4}$$

단,

$$\mu_0^{(t)} = \lambda_j^{(t)}, \quad \mu_1^{(t)} = \sum_{c \neq j} h_{j_2 c} \lambda_c^{(t)}, \quad \mu_2^{(t)} = \sum_{c \neq j} h_{j_1 c} \lambda_c^{(t)} \tag{4.5}$$

과 같이 얻을 수 있다. 단,

$$p(x, y; \mu_0, \mu_1, \mu_2) = e^{-\mu_0 - \mu_1 - \mu_2} \prod_{i=1}^{\min x, y} \frac{\mu_0^i}{i!} \frac{\mu_1^{x-i}}{(x-i)!} \frac{\mu_2^{y-i}}{(y-i)!}$$

는 모수가 (μ_0, μ_1, μ_2)인 이변량 포아송 확률분포함수이며, $(h_{j_1,c}, h_{j_2,c})$은 행렬 $\boldsymbol{H}$의 c 번째 열에서 OD 주소 $j = (j_1, j_2)$에 대응하는 원소이다.

그런데 식 (4.4)의 E-step에서 계산해야 하는 이변량 포아송 확률분포가 $y_{j_2}^{I(k)}, y_{j_1}^{O(k)}$ 값이 크면 계산시간이 오래 걸릴 뿐 아니라 정확도가 떨어지는 것으로 알려져 있다. 더구나 식 (4.3)의 M-step에서 k에 관한 합이 자료에 직접 작용하지 않기 때문에 매 단계마다 관측치의 개수인 K

번만큼 이변량 포아송 확률밀도값을 분모와 분자에 계산해 넣어야 한다. 그래서 관측치의 개수가 작을 때조차 거의 실현할 수 없을 정도의 계산시간 문제를 야기한다.

이 문제를 해결하기 위해 완전자료를 $X_j^I, X_j^O; j=1,\cdots,M$와 같이 확장한다. 단, $X_j = X_j^I + X_j^O$ 이며, $X_j^I \sim$ Poisson$(w_j\lambda_j)$ 및 $X_j^O \sim$ Poisson$((1-w_j)\lambda_j)$를 따른다 하자. 여기서 $0 \le w_j \le 1$인 상수이다. 이때,

$$Q(\underline{\lambda}|\underline{\lambda}^{(t)}) \propto \sum_{k=1}^{K}\sum_{j=1}^{M}\left\{-\lambda_j + \log\lambda_j\left(E\left[X_j^{I(k)}|y_{j_2}^{I(k)}, \underline{\lambda}^{(t)}\right] + E\left[X_j^{O(k)}|y_{j_1}^{O(k)}, \underline{\lambda}^{(t)}\right]\right)\right\} \quad (4.6)$$

이 되며,

$$E\left[X_j^{I(k)}|y_{j_2}^{I(k)}, \underline{\lambda}^{(t)}\right] = y_{j_2}^{I(k)}\frac{w_j\lambda_j^{(t)}}{\lambda_j^{(t)} + \sum_{c\neq j}^{M} h_{j_2c}\,\lambda_c^{(t)}} = \lambda_j^{(t)}\frac{w_j\, y_{j_2}^{I(k)}}{\sum_{c=1}^{M} h_{j_2c}\,\lambda_c^{(t)}} \quad (4.7)$$

$$E\left[X_j^{O(k)}|y_{j_1}^{O(k)}, \underline{\lambda}^{(t)}\right] = y_{j_1}^{O(k)}\frac{(1-w_j)\lambda_j^{(t)}}{\lambda_j^{(t)} + \sum_{c\neq j}^{M} h_{j_1c}\,\lambda_c^{(t)}} = \lambda_j^{(t)}\frac{(1-w_j)\, y_{j_1}^{O(k)}}{\sum_{c=1}^{M} h_{j_1c}\,\lambda_c^{(t)}} \quad (4.8)$$

를 얻을 수 있다. 따라서 EM 알고리즘은

$$\begin{aligned}\lambda_j^{(t+1)} &= \frac{1}{K}\sum_{k=1}^{K}\left\{E\left[X_j^{I(k)}|y_{j_2}^{I(k)}, \underline{\lambda}^{(t)}\right] + E\left[X_j^{O(k)}|y_{j_1}^{O(k)}, \underline{\lambda}^{(t)}\right]\right\} \\ &= \lambda_j^{(t)}\left[\frac{w_j\,\bar{y}_{j_2}^{\,I}}{\sum_{c=1}^{M} h_{j_2c}\,\lambda_c^{(t)}} + \frac{(1-w_j)\,\bar{y}_{j_1}^{\,O}}{\sum_{c=1}^{M} h_{j_1c}\,\lambda_c^{(t)}}\right], \quad j=1,\cdots,M \end{aligned} \quad (4.9)$$

으로 얻을 수 있다. 단, $\bar{y}_j = \sum_{k=1}^{K} y_j^{(k)}/K$ 를 나타낸다.

EM 알고리즘 식 (4.9)는 $\lambda_j^{(t)} > 0$을 만족하는 추정치를 제공해야한다. 이 것은 초기치에 의해 만족시킬 수 있는데, 식 (4.9)의 괄호 내의 성분에서 알 수 있듯이, 초기 추정치 $\lambda_j^{(0)}$을 양의 값으로 취하는 한 언제나 비음의 조건을 만족하는 해를 얻게 된다.

0과 1사이의 값을 취하는 가중치는 사전에 분석자에 의해 선택되는 모수로서, w_j=1/2를 선택하면, 노드의 입/출력 트래픽 정보를 동등하게 사용함을 반영한다. 만약 w_j=0 혹은 1을 취하면 각각 출력 혹은 입력 트래픽 자료만을 이용하여 트래픽 밀도를 추정하게 된다. 한편, 노드 j에 따라 서로 다른 가중치를 사용할 수 있는데, 저자의 경험으로 볼 때 $w_j^{(t)} = \mu_{j1}^{(t)}/(\mu_{j1}^{(t)} + \mu_{j2}^{(t)})$이 비교적 좋은 추정 결과를 나타내었다.

5. 모의실험

예제 1.1에서 소개된 링크 라우팅 행렬을 $\boldsymbol{A}_1$이라 하고, OD (1,2), (1,3) 및 (1,4)의 경로 L_1과 L_2 및 $L_2 \rightarrow L_6$를 각각 $L_2 \rightarrow L_5$과 $L_1 \rightarrow L_4$ 및 $L_1 \rightarrow L_4 \rightarrow L_6$으로 수정하여 라우팅 $\boldsymbol{A}_2$하고,

OD (2,3) 및 (2,4)의 경로 L_4와 $L_3 \rightarrow L_2 \rightarrow L_6$을 각각 $L_3 \rightarrow L_2$ 및 $L_4 \rightarrow L_6$로 바꾸어 A_3라 하자. 즉, A_2와 A_3는 행렬 A의 열 성분을 각각

(1,2)	(1,3)	(1,4)
0	1	1
1	0	0
0	0	0
0	1	1
1	0	0
0	0	1
0	0	0

(2,3)	(2,4)
0	0
1	0
1	0
0	1
0	0
0	1
0	0

으로 치환한 것이다. 그리고 두 성분을 동시에 치환한 행렬을 A_4라 하자. 동적 라우터를 $A_d = \{A_1, A_2, A_3, A_4\}$과 같이 정의하고, 관측 때마다 동적 라우터의 성분 중 하나가 임의적으로 선택되도록 하였다.

12 개의 OD 트래픽 $X^{(k)} = (X_{1,2}^{(k)}, X_{1,3}^{(k)}, \cdots, X_{4,3}^{(k)})^T$은 밀도 $\lambda = (10, 20, \cdots, 120)^T$인 포아송 분포로부터 독립적으로 자료를 생성하여 만들었으며, $U = A_d X$에 의해 링크 트래픽 자료를 얻었다. 그리고 식 (3.3)-(3.4)의 관계로부터 경유 트래픽을 뺀 노드 트래픽 자료 $Y^{(k)} = ((Y^{I(k)} - R^{(k)})^T, (Y^{I(k)} - R^{(k)})^T)^T$, $k = 1, \cdots, 1000$를 사용하여 식 (4.9)의 EM 알고리즘에 적용하여 밀도 λ를 추정하였다. 단, 초기해는 $\lambda^{(0)} = (1, 1, \cdots, 1)^T$로 하였다.

한편, 모의실험은 관측치의 개수가 $K = 20, 100, 1000$인 경우에 대해서 각각 200 번씩의 추정치를 구하여 추정치의 평균 $\mathrm{av}(\hat{\lambda}_j)$와 표준편차 $\mathrm{sd}(\hat{\lambda}_j)$을 측정하였다. 그 결과를 표 5.1에 나타내었다.

표 5.1의 실험 결과는 라우팅이 동적이라 하더라도 측정된 관측치 즉 I/O 노드 트래픽과 경유 트래픽을 사용하여 매우 만족스러운 토모그래피가 가능함을 보여주고 있다. 특히 밀도 값이 작은 $\lambda_{1,2} \approx 10$의 경우를 제외하고 밀도 추정치는 표본의 크기와 관계없이 불편성(unbiasedness)을 만족하는 것으로 보인다. 그리고 참 밀도의 값이 커질수록 추정치의 변이가 커지는 경향을 뚜렷이 보이고 있으나, 표본의 크기가 커짐에 따라 변이는 작아지고 있다. 이 것은 최우추정량의 일치성(consistency) 특징을 그대로 나타내고 있는 것으로 파악할 수 있다.

6. 결론 및 추가논의

본 연구에서는 동적 라우팅을 따르는 네트워크에서 OD 트래픽 밀도 추정을 위한 토모그래피 기법을 제안하였다. 기존의 연구는 링크 단위의 트래픽을 관측치로 사용하기 때문에 문제가 라우팅 종속적일 수밖에 없으며, 따라서 동적 라우팅 하에서는 토모그래피가 거의 불가능하다. 본 논문은 관측치로서 링크 단위의 트래픽 대신 각 노드에서 관측할 수 있는 I/O 노드 트래픽과 경유 트래픽을 사용하여 라우팅에 무관한 고정 변환 관계로부터 OD 트래픽 밀도 추정이 가능함을 보였다. 그러나 본 연구에서 실험한 네트워크는 노드의 수가 4 개로 비교적 크기가 작은 것이다. 이 때 고정 변환행렬 H의 크기는 8×12로서 완전위수행렬(full-ranked matrix)이 아니다.

표 5.1 관측치의 개수에 따른 모의실험결과: 밀도의 참값(제1열), 평균(제2열) 및 표준편차(제3열).

관측치수	$n=20$			$n=100$			$n=1000$		
OD 주소	$\overline{\lambda_j}$	av $(\hat{\lambda_j})$	sd $(\hat{\lambda_j})$	$\overline{\lambda_j}$	av $(\hat{\lambda_j})$	sd $(\hat{\lambda_j})$	$\overline{\lambda_j}$	av $(\hat{\lambda_j})$	sd $(\hat{\lambda_j})$
(1,2)	9.90	17.28	0.65	9.80	17.32	0.28	9.97	17.32	0.08
(1,3)	19.25	19.46	0.72	19.65	19.42	0.31	20.03	19.41	0.09
(1,4)	30.05	23.26	0.77	30.79	23.29	0.37	29.96	23.29	0.12
(2,1)	40.75	41.19	1.04	41.24	41.28	0.41	40.08	41.27	0.15
(2,3)	50.60	49.79	1.23	50.21	49.68	0.52	49.93	49.70	0.17
(2,4)	60.50	58.98	1.42	59.76	59.02	0.64	60.17	59.07	0.20
(3,1)	69.50	68.40	1.37	69.82	68.40	0.65	69.70	68.35	0.21
(3,2)	76.95	74.02	1.56	81.49	73.98	0.70	80.04	74.00	0.23
(3,4)	90.50	97.74	1.72	88.69	97.60	0.90	89.39	97.64	0.29
(4,1)	101.30	100.30	1.66	102.11	100.42	0.81	99.66	100.40	0.24
(4,2)	107.40	108.60	1.85	108.61	108.69	0.91	109.82	108.75	0.28
(4,3)	118.90	121.17	2.10	121.07	120.89	0.93	120.39	120.94	0.28

따라서 추정 상에 정칙화 과정(regularization)이 필요하다. 다행히 적정한 반복횟수의 EM 알고즘은 어느 정도 정칙화의 성질을 가지므로, 본 연구의 모의실험에서는 큰 문제를 보이지는 않았다. 그러나 노드의 수가 큰 네트워크에서의 정칙화에 관한 연구가 필요할 것으로 판단된다.

참고문헌

[1] Cao, J., Davis, D., Wei, S.V. and Yu, B.(2000), Time-Varying Network Tomography: Router Link Data, *Journal of American Statistical Association*, vol 95, 1063-1075.

[2] Castro, R., Coates, M., Liang, G., Nowak, R. and Yu, B.(2003), Network Tomography: recent developments, *Statistical Science(invited)*

[3] Coates, M.J. and Nowak, R.(2001), Network delay distribution inference from end-to-end unicast measurement, *Proceeding of IEEE Int. Conference on Acoustics, Speech and Signal Processing*, May 2001.

[4] Karlis, D.(2003), An EM Algorithm for Multivariate Poisson Distribution and related Models, *Journal of Applied Statistics*, vol. 30, no. 1, 63-77.

[5] Liang. G. and Yu, B.(2003), Maximum Pseudo Likelihood Estimation in Network Tomography, *IEEE transaction on Signal Processing*(Special Issues on Data Networks).

[6] Tebaldi, C. and West, M.(1998), Bayesian inference on network traffic using link count data(with discussion), *Journal of American Statistical Association*, vol. 93, 557-576.

[7] Vardi, Y.(1996), Network tomography: Estimating source-destination traffic intensities from link data, *Journal of American Statistical Association*, vol. 91, 365-37.

[2003년 8월 접수, 2003년 9월 채택]