円周上の混合分布モデルと交通量データへの応用

加藤 昇吾　　数理・推論研究系 准教授

はじめに

交通量データ

本研究では，時刻として表される以下のデータの統計解析を行う：

観測期間： 2016年6月6日〜7月7日，8月22日〜10月10日の全平日（46日間），

観測場所： 阪神高速3号神戸線の上り線の魚崎–尼崎東間のトラフィックカウンター(図１・左)，

データ： トラフィックカウンターを通過する車両の時刻のデータ(図１・右) ($n = 1121262$).

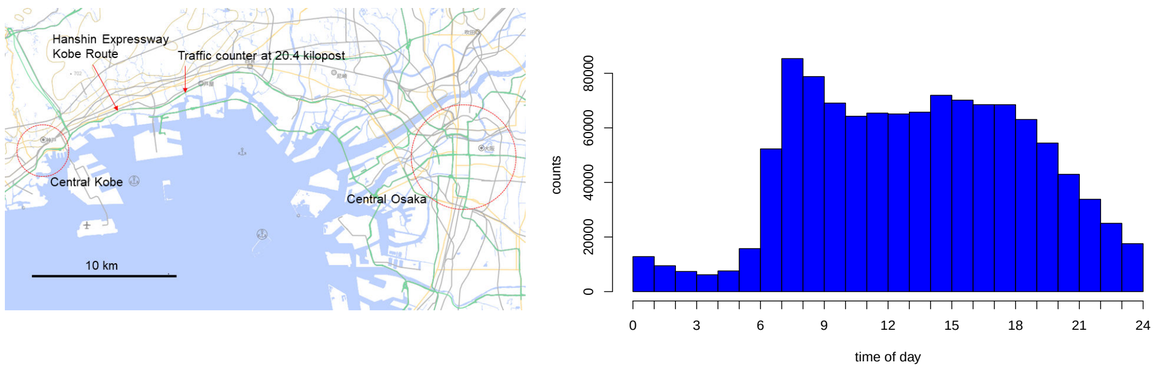

図１．（左）トラフィックカウンターの位置を表す地図．大阪と神戸の間に位置する．
（右）トラフィックカウンターを通過する車両の時刻データのヒストグラム．
※　左図は背景図に地理院地図を使用．右図の元データは阪神高速道路より提供．

研究の目的

時刻は周期性を持つ，つまり，0:00と24:00は同じ時刻を表す．

このような周期性を持つデータのモデル化には，通常の実数値データのための統計モデルをそのまま当てはめることができない．

本研究では，0:00〜24:00の時刻を，（同じく周期性を持つ）0〜2πの角度へと変換し，角度の確率分布によりモデル化を行う．

なお，本研究は，長崎混大氏(東京工業大学)，中西航准教授(金沢大学)，M.C. Jones名誉教授（The Open University, UK）との共同研究である．

円周上の混合分布モデル

定義

本研究では，以下の確率密度関数を持つ円周上の混合分布を考える：

$$f(\theta) = \frac{1}{2\pi}\sum_{k=1}^{m}\pi_k\left\{1 + 2\gamma_k\frac{\cos(\theta-\mu_k)-\rho_k\cos\lambda_k}{1+\rho_k^2-2\rho_k\cos(\theta-\mu_k-\lambda_k)}\right\}, \qquad (1)$$

$$0 \le \theta < 2\pi.$$

ただし，$m\,(\in\mathbb{N})$は混合分布の成分数，$\pi_1,\ldots,\pi_m\,(\in(0,1))$は混合比率 $(\sum_{k=1}^{m}\pi_k = 1)$，$\mu_k,\lambda_k \in [0,2\pi)$，$\gamma_k,\rho_k \in [0,1)$，$(\rho_k\cos\lambda_k-\gamma_k)^2 + (\rho_k\sin\lambda_k)^2 \le (1-\gamma_k)^2$，$k = 1,\ldots,m$.

混合分布(1)の個々の成分は，Kato & Jones (2015)の分布となっている．

Kato & Jones (2015)の分布は，単峰形であり，位置・集中度・歪度・尖度を幅広く調節することが可能な分布として知られる．

確率密度関数の別表現

混合分布の密度関数(1)は，以下のように表すことが可能である：

$$f(\theta) = \frac{1}{2\pi}\sum_{k=1}^{m}\pi'_k\left\{1 + 2\bar{\gamma}_k\frac{\cos(\theta-\mu_k)-\rho_k\cos\lambda_k}{1+\rho_k^2-2\rho_k\cos(\theta-\mu_k-\lambda_k)}\right\} + \frac{\pi'_{m+1}}{2\pi}. \qquad (2)$$

ここに，$\pi'_k = \pi_k\gamma_k/\bar{\gamma}_k$ $(k = 1,\ldots,m)$，$\pi'_{m+1} = 1 - \sum_{k=1}^{m}\pi'_k$，$\bar{\gamma}_k = (1-\rho_k^2)/\{2(1-\rho_k\cos\lambda_k)\}$.

混合分布(2)は識別可能（identifiable）なモデルとなる．

パラメータ推定

$\Theta_1,\ldots,\Theta_n \sim i.i.d.$ 混合分布(2) とする．

修正されたモーメント法

以下の評価関数を考える：

$$\mathrm{ETM}(\boldsymbol{\Psi}) \equiv \sum_{p=1}^{q} w(p)\left|\frac{1}{n}\sum_{j=1}^{n}e^{ip\Theta_j} - E(e^{ip\Theta})\right|^2$$

$$= \sum_{p=1}^{q} w(p)\left|\frac{1}{n}\sum_{j=1}^{n}e^{ip\Theta_j} - \sum_{k=1}^{m}\pi'_k\bar{\gamma}_k\left(\rho_k e^{i\lambda_k}\right)^{-1}\left\{\rho_k e^{i(\mu_k+\lambda_k)}\right\}^p\right|^2.$$

ここで，$\boldsymbol{\Psi} = (\mu_1,\ldots,\mu_m,\rho_1,\ldots,\rho_m,\lambda_1,\ldots,\lambda_m,\pi'_1,\ldots,\pi'_{m+1})$，$w(p)$は重み関数，$q \ge 2m$.

このとき，ETMに基づく新たな推定量を，$\hat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}\in\Omega}{\mathrm{argmin}}\,\mathrm{ETM}(\boldsymbol{\Psi})$ として提案する．ただし，Ωは$\boldsymbol{\Psi}$のパラメータ空間を表す．

【性質】 (i) 一致性を持つ，(ii)（通常のモーメント法と異なり）推定値がパラメータの定義域から外れることがない，(iii) 後述の最尤推定法よりも早く計算できる．

最尤推定

EMアリゴリズムにより最尤推定値を計算することが可能である．

Mステップにおける最大化は，(混合分布ではない) Kato & Jones (2015)の分布の重み付き最尤推定と等しくなる．

【性質】 (i) 有効推定量となる，(ii) 数値実験により，EMアリゴリズムを用いた方が，混合分布の同時密度関数から直接計算するよりも，より安定した解を導くことが示唆されている．

データ解析に用いたパラメータ推定の手順

1: パラメータの取り直しをした混合分布(2)を，ETMに基づいてパラメータ推定を行う．

2: ETMに基づく推定値を初期値として，混合分布(2)の最尤推定値をEMアルゴリズムにより求める．

3: 元の混合分布(1)のパラメータ$\{\gamma_k,\pi_k\}$を，$\hat{\pi}_k = \hat{\pi}'_k/(1-\hat{\pi}'_{m+1})$および$\hat{\bar{\gamma}}_k = \hat{\pi}'_k\hat{\bar{\gamma}}_k/\hat{\pi}_k$を仮定することにより復元する．

交通量データへの応用

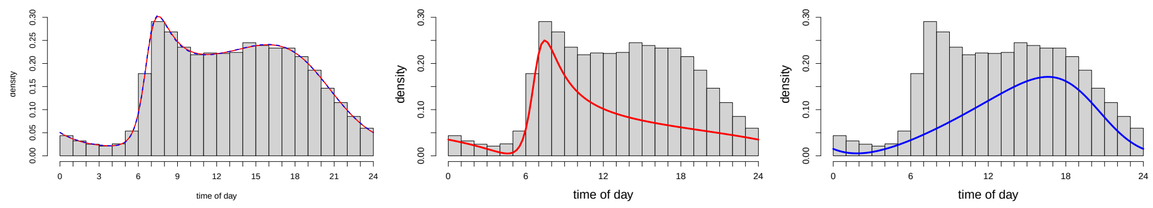

図２．(左) 最尤法により推定された密度関数（実線・赤）とETMに基づいて推定された密度関数（点線・青）（成分数：2）．
（中・右）最尤法により推定された密度関数の１つめの成分（中）と２つめの成分（右）．

成分	$\hat{\mu}_k$ (平均方向)	最頻値	$\hat{\gamma}_k$ (集中度)	$\hat{\alpha}_{2k}$ (尖度)	$\hat{\beta}_{2k}$ (歪度)	$\hat{\pi}_k$ (混合比率)
成分１	10:32	7:28	0.3751	0.1542	−0.2248	0.4845
成分２	15:19	16:37	0.4855	0.0356	0.0888	0.5155

$\hat{\alpha}_{2k} = \hat{\rho}_k\hat{\bar{\gamma}}_k\cos\hat{\lambda}_k$,　$\hat{\beta}_{2k} = \hat{\rho}_k\hat{\bar{\gamma}}_k\sin\hat{\lambda}_k$.