多種粒子TASEPにおける統計的粒子識別について

山崎 啓介

東京工業大学大学院知能システム科学専攻

概要

本稿では格子上で同期更新される多種粒子 TASEP に対し時空図の情報を用いて粒子の識別を行う。識別には最尤クラスタリングとベイズクラスタリングの２種類の統計的手法を用いる。格子を通過するまでに前方のセルが空であった回数は負の二項分布の混合モデルで示される。この混合モデルに基づき両手法におけるクラスタリングアルゴリズムを導出した。最尤クラスタリングでは粒子の種数が未知の場合に識別が失敗し、ベイズクラスタリングでは種数決定と識別の両方で有用な結果を得ることをシミュレーションで確認した。

On Statistical Clustering of Multi-Species TASEP

Keisuke Yamazaki

Department fo Computational Intelligence and Systems Science, Tokyo Institute of Technology

Abstract

In this paper, we study statistical clustering of the multi-species totally asymmetric simple exclusion process (TASEP) with parallel dynamics under open boundary conditions. Two statistical methods are applied to clustering: the maximum-likelihood and the Bayes methods. The total number of times the next cell is empty is expressed as a mixture of negative binomial distributions. Clustering algorithms for the both methods are based on this mixture model. According to results from simulation data, we observe that the Bayes method succeeds in detection of the number of species and distinguishing them while the maximum-likelihood method fails when the number is unknown.

1　はじめに

TASEP(totally asymmetric simple exclusion process) は一次元格子上で定義される排他過程であり、交通流を表現する基本的なモデルである。ホップ確率が異なる多種の粒子から構成される TASEP も考案されており、その数理的な挙動が研究されている [1]。多種粒子 TASEP が生成する流れは異なる速度に従う複数の車種が混在する交通流と解釈できる。２種粒子 TASEP を用いて一車線高速道路における低速車と高速車を表現し、定常状態を解析することで車種の混合比と各々の速度を推定する方法が提案された [2]。本稿では多種粒子 TASEP を用いて時空図にある粒子の種類を識別する手法を提案する。これにより道路上の車両が速度を元に何種類のグループに分かれ、各車両がどのグループに属するかを推定することが可能になる。

2　データ特徴量と統計モデル

本稿で考える TASEP を以下のように定義する。粒子は 1 次元 $L+1$ サイトの格子上を同時更新し左から右へと移動する。初期状態では全てのセルが空とする。最初のセルが空のとき、確率 α で粒子が入る。粒子の種類は K 個存在し確率 π_k $(k = 1, \ldots, K)$ に従い流入した粒子の種類を決定する。ここで $\pi_k > 0, \sum_{k=1}^{K} \pi_k = 1$ を満たす。k 番目の種類の粒子のホッ

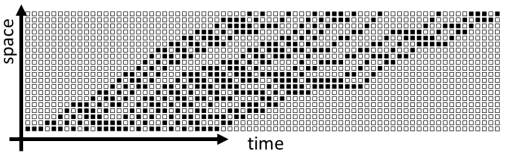

図 1: 粒子 10 個が通過するまでの時空図 ($L = 20$)。2 種粒子 TASEP($K = 2$, $\alpha = 0.8$, $\pi_1 = \pi_2 = 0.5$, $p_1 = 0.9$, $p_2 = 0.6$) により生成した。

プ確率を $p_k > 0$ とし、$k \neq j \Rightarrow p_k \neq p_j$ とする。$L+1$ 番目のセルではどの粒子も確率 1 で外に（右に）ホップする。つまり L 番目のセルに粒子があるとき $L+1$ 番目のセルは常に空である。図 1 に時空図の例を示す。横軸が時間、縦軸が粒子の移動方向を表す。

データは N 個の粒子が通過するまでの時空図とし、いずれの粒子も種類は明かされていないとする。粒子の種類を識別するための特徴量として、前方のセルが空であった回数を用いる。全ての種類において L 番目のセルを通過するまでにホップする回数は共通して L であるが、前方セルが空でかつホップしなかった回数は種類に依存する。前方セルが空の回数はホップした回数としなかったものの和である。つまりホップ確率 p_k が小さい種類ほどこの和が大きくなる。TASEP において移動はベルヌーイ試行（統計的に独立して決定される試行）であるため、ホップするのを「成功」、しないのを「失敗」とみなすと、前方セルが空の回数の分布は「L 回成功するまでに試行した回数」の分布となる。これは統計学で負の二項分布として知られる分布であり、全試行数を x とするとホップ確率 p_k の粒子では

$$p(x|p_k) = \frac{\Gamma(x)}{\Gamma(L)\Gamma(x-L-1)} p_k^L (1-p_k)^{x-L}$$

と表される。ここで $\Gamma(\cdot)$ はガンマ関数である。粒子はどの種類か分からないので全ての可能性を加味し、

$$p(x|\pi, p) = \sum_{k=1}^{K} \pi_k p(x|p_k)$$

が試行回数の分布となる。これは負の二項分布の混合モデルであり、特徴量の分布はこの式で表現されることがわかる。ここで $\pi = \{\pi_1, \ldots, \pi_K\}$ と $p = \{p_1, \ldots, p_K\}$ はパラメータである。

3 クラスタリング手法

n 番目の粒子について前方セルが空であった回数を x_n とすると時空図から特徴量の集合 $X = \{x_1, \ldots, x_N\}$ が得られる。各粒子の種類を表す集合を $Y = \{y_1, \ldots, y_N\}$ とする。ただし $y_n \in \{1, \ldots, K\}$ は n 番目の粒子の種類を表す。X が起こる確率を最大にするパラメータを用いて Y を推定する手法を最尤クラスタリングと呼ぶ。一方、パラメータを積分消去し X と Y の条件付確率 $p(Y|X)$ を求め、これを最大化するように Y を決定する方法をベイズクラスタリングと呼ぶ。以下では両手法について説明する。

3.1 最尤クラスタリング

最尤法では尤度関数を用いてパラメータを最適化し、これを基に粒子の種類を推定する。

尤度関数は以下で定義される。

$$L(\pi, p) = \prod_{i=1}^{N} p(x_i|\pi, p).$$

尤度関数は与えられた特徴量が起こる同時確率であり、これを最大化するパラメータは最尤推定量と呼ばれる。

$$(\hat{\pi}, \hat{p}) = \arg\max_{\pi, p} L(\pi, p).$$

最尤推定量を求める手法は EM(expectation-maximization) アルゴリズム [3] が代表的である。本稿で扱う混合モデルに対する EM アルゴリズムは以下で表される。パラメータの初期値を適当に与え、次に示す E-step と M-step を収束するまで（もしくは所定の回数）繰り返す。

E-step: n 番目の粒子が k 番目の種類である確率 γ_{nk} を負担率と呼び、

$$\gamma_{nk} = \frac{\pi_k p(x_n|p_k)}{p(x_n|\pi, p)}$$

とする。

M-step: 負担率を用いてパラメータの更新を行う。

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}},$$
$$p_k = \frac{L \sum_{n=1}^{N} \gamma_{nk}}{\sum_{n=1}^{N} \gamma_{nk} x_n}.$$

これらのステップを繰り返すごとに尤度関数が増大することが知られている。しかしながら得られた解

真の TASEP	2,1,1,2,1,1,2,2,2,2,2,2,1,2,2,1,1,1,1,1,1,1,1,2,2,2,2,2,1,2,1,1,1,1,2,2,2,2,1,2,1,...
最尤法	2,1,1,2,1,1,2,2,2,2,2,2,1,2,2,1,1,1,1,1,1,1,1,2,2,2,2,2,1,2,1,1,1,1,2,2,2,1,1,2,1,...
ベイズ法	2,1,1,2,1,1,2,2,1,2,2,2,1,2,2,1,1,1,1,1,1,1,1,2,2,2,2,2,1,2,1,1,1,1,2,2,2,2,1,2,1,...

表 1: 上から順に真の TASEP による $y_1,\ldots,y_{40}$、$K=2$ における最尤クラスタリングの結果、ベイズクラスタリングの結果。

が大域解である保証がないため、複数の初期値を試す必要がある。

EM アルゴリズムで得られたパラメータを最尤推定量 $(\hat{\pi},\hat{p})$ とみなし、これを用いて負担率 γ_{nk} を求める。n 番目の粒子について $\gamma_n=\{\gamma_{n1},\ldots,\gamma_{nK}\}$ に従って種類 y_n を決定する。

3.2 ベイズクラスタリング

ベイズ法ではパラメータの事前分布を用いる。本稿では π についてディリクレ分布 $\varphi(\pi|\eta_\pi)$ を、p についてベータ分布 $\varphi(p_k|\eta_p)$ を事前分布とする。事前分布のパラメータ η_π,η_p はハイパーパラメータと呼ばれる。

$$\begin{aligned}\varphi(\pi|\eta_\pi)=&\frac{\Gamma(\eta_\pi)^K}{\Gamma(K\eta_\pi)}\prod_{k=1}^K\pi_k^{\eta_\pi-1},\\ \varphi(p_k|\eta_p)=&\frac{\Gamma(\eta_p)^2}{\Gamma(2\eta_p)}p_k^{\eta_p-1}(1-p_k)^{\eta_p-1},\\ \varphi(p|\eta_p)=&\prod_{k=1}^K\varphi(p_k|\eta_p).\end{aligned}$$

X と Y の同時確率は

$$p(X,Y)=\int\prod_{n=1}^N\pi_{y_n}p(x_n|p_{x_n})\varphi(\pi)\varphi(p)d\pi dp$$

で表され、X が与えられたときの Y の条件付確率は

$$p(Y|X)=\frac{p(X,Y)}{\sum_Y p(X,Y)}\propto p(X,Y)$$

となる。クラスタリングの目的は与えられた X に対する Y を得ることなので、上式の確率に従う Y をサンプリングすればよい。分母は Y に依らないため分子の値に従うサンプリングを行う。一般的に $p(X,Y)$ はパラメータに関する高次元積分を含むため多くの計算量を必要とするが、本稿のモデルは共役事前分布を有し、次のように解析的な積分消去が可能である。表現を簡略化する目的で $-\ln p(X,Y)$ について積分消去を行うと、Y に対するある定数 C を用いて

$$\begin{aligned}-\ln p(X,Y)=&C-\sum_{k=1}^K\ln\Gamma(N_k+\eta_\pi)\\ &+\sum_{k=1}^K\ln\Gamma\Big(\sum_{n=1}^N z_{nk}x_n+2\eta_p\Big)\\ &-\sum_{k=1}^K\ln\Gamma\Big(\sum_{n=1}^N z_{nk}x_n-LN_k+\eta_p\Big)\\ &-\sum_{k=1}^K\ln\Gamma(LN_k+\eta_p)\end{aligned}$$

となる。ここでクロネッカーのデルタ関数を用いて $z_{nk}=\delta_{y_nk}$ とし、$N_k=\sum_{n=1}^N z_{nk}$ とした。これらは (X,Y) が与えられると計算可能な十分統計量である。$-\ln p(X,Y)$ をハミルトン関数としてマルコフ連鎖モンテカルロ法で Y をサンプリングすると条件付確率 $p(Y|X)$ に従う Y が決定できる。サンプリングの際、定数 C は条件付確率の定義より無視できる。

4 シミュレーションデータによるクラスタリング手法の検証

シミュレーションのデータは 2 種粒子 TASEP を用いて発生させた $(K=2)$。サイト数 $L=20$、流入の確率 $\alpha=0.8$、混合比 $\pi_1=\pi_2=0.5$、ホップ確率をそれぞれ $p_1=0.9$, $p_2=0.6$ とし、粒子数 $N=100$ の時空図を作成した。この TASEP を「真の TASEP」と呼ぶ。乱数の異なる 10 個の時空図についてそれぞれクラスタリングを行った。

種類別に色分けされた推定結果の一部を表 1 に示す。粒子 $x_1,\ldots,x_{100}$ に対する種類を y_1 から順に y_{40} まで表示した。表の上から順に真の TASEP が生成した Y、最尤クラスタリングの推定結果、ベイズクラスタリングの推定結果である。推定結果はそれぞれ EM アルゴリズムとマルコフ連鎖モンテカルロ法を 100 回繰り返した後の Y とした。ここでは粒子の種類の数は既知とした。最尤クラスタリング、ベイズクラスタリングともに真の TASEP が生成した

真の TASEP	2,1,1,2,1,1,2,2,2,2,2,2,1,2,2,1,1,1,1,1,1,1,1,2,2,2,2,1,2,1,1,1,1,2,2,2,2,1,2,1,...
最尤法	2,1,1,3,1,1,2,3,4,4,3,2,4,3,4,1,1,1,1,1,1,1,1,4,2,2,2,1,4,1,1,1,1,4,4,2,3,1,4,1,...
ベイズ法	2,1,1,2,1,1,2,2,1,2,2,2,1,2,2,1,1,1,1,1,1,1,1,2,2,2,2,1,2,1,1,1,1,2,2,2,2,1,2,1,...

表 2: $K = 5$ におけるクラスタリング結果。

粒子の種類を概ね正しく推定できていることがわかる。10 個の時空図について最尤クラスタリングでは平均 96.0 個、ベイズクラスタリングでは平均 95.7 個の粒子を正しく識別できた。

次に粒子の種類の数が未知の場合を考える。粒子の混合数を $K = 5$ とし、同様に最尤クラスタリングとベイズクラスタリングを行った。結果の一部を表 2 に示す。最尤クラスタリングでは 4 つの種類に分類しているのに対し、ベイズクラスタリングでは不必要な種類は全て消去され 2 つに分類された。他の時空図に関しても同様の結果を得ており、ベイズクラスタリングが冗長な種類を効率よく刈り込むのに比べ、最尤クラスタリングは 2〜4 種類に分類する結果が得られた。

5　考察とまとめ

粒子数 N が十分大きい場合、統計的漸近論を用いて両手法の比較が行われている [4]。最尤推定量の探索とマルコフ連鎖モンテカルロ法のサンプリングが正しく行えたという仮定の下では、ベイズ法が最尤法よりも精度のよい結果となる。本稿での計算機実験では $N = 100$ であり、これらの仮定が満たされるか否かの判定は困難なため実験結果からクラスタリングの挙動を調べ両手法の特徴について考察する。

最尤法では $K = 2$ の識別に成功し $K = 5$ では失敗することが多い。EM アルゴリズムが最大化する尤度関数は一般的に多峰性を有しており、他の統計モデルにおいても局所解への収束が頻繁に起こることが知られている。本稿の実験でも初期値の影響が大きく結果が不安定であった。また冗長な種類が存在する場合にこの傾向が強いことがわかった。こうした現象に対して統一的な解決策は見つかっておらず、異なる初期値を多く試すなど経験的な方法で回避するほかない。一方、ベイズ法で最大化する条件付確率 $p(Y|X)$ の関数特性は未だ知られておらず、クラスタリングの挙動はわかっていない。今回の $K = 5$ の実験において、全ての時空図で不必要な種類の刈り込みが行われた。特にハイパーパラメータ η_π の値が小さいと刈り込みが早いことが観測された。この性質は実際のデータ解析で非常に有用であるため、そのメカニズムを明らかにしベイズクラスタリングの挙動を詳細に調べることが今後の課題である。

謝辞　本研究の一部は栢森情報科学振興財団研究助成金、科研費 (若手 (B)24700139) の助成を受けたものである。

参考文献

[1] M. R. Evans and T. Hanney, J. Phys. A: Math. Gen. **38**(2005) R195.

[2] 金井政宏, 山崎啓介, 第 18 回交通流のシミュレーションシンポジウム論文集, 9(2012).

[3] A. P. Dempster et al., J. Royal Stat. Soc. B **39**(1977)

[4] K. Yamazaki, arXiv:1204.2069