Vol.2018-IS-143 No.2
2018/3/5

電子商取引サイトにおける顧客到着現象の混合分布推定

佐藤　聖[a)]

概要：一般にマーケットにおける顧客到着現象は指数分布に従っていると言われている．本研究では電子商取引サイトにおいて顧客到着現象の混合分布推定を目的をする，推定の結果，混合数 4 において一番近い分布を推定することができた．今回使用した顧客到着現象のデータでは大部分が形状パラメータ k=2，尺度パラメータ θ =7 に従うガンマ分布であることが推定できた．今回の推定結果から，電子商取引サイトにおいても顧客到着現象は指数分布に個々人が従っているものだと推測できる．

キーワード：混合分布，指数到着

Mixed distribution estimation of customer arrival phenomenon on e-commerce site

Sei Sato[a)]

Abstract:
In general it is said that customer arrival phenomenon in the market follows exponential distribution. In this research, we aim to estimate mixed distribution of customer arrival phenomena at e-commerce site. As a result of estimation, it was possible to estimate the distribution closest in mixture number 4. In the data on the customer arrival phenomenon used this time, it was estimated that the gamma distribution mostly corresponds to the shape parameter k = 2 and the scale parameter θ = 7. Based on the estimation result this time, it can be inferred that even at the e-commerce site, the customer arrival phenomenon is one in which each individual follows exponential distribution.

Keywords: mixed distribution，Exponential arrival

1. はじめに

近年，コンピューターや通信技術の発達と共に，容易に様々なデータの収集が行えるようになった．得られた膨大なデータはビックデータと呼ばれ．そこから新たな知見を得られることが多くなった．そのためビックデータの収集は盛んに行われており．それぞれのデータセットに合わせたデータ解析手法が必要とされている．このデータ解析の主要な技術の一つとして混合分布が知られている．混合分布推定では EM アルゴリズムを用いたパラメータ推定がよく行われている．EM アルゴリズムは Dempster[2] らによって定式化されたもので，与えられた混合数と，初期パラメータに基づいて，逐次的にモデルの対数尤度を増加させて，局所最尤推定に収束することが知られている．

確率分布間の距離推定を利用して混合分布推定を行う方法がある．Sugiyama[1] によって取り上げられているように確率分布間の距離を推定する手法が多く存在している．統計学や機械学習の分野において多くの場合で用いられているカルバック・ライブラー (KL) 距離をはじめとして，ピアソン距離，相対 PE 距離，L^2 距離などが挙げられている．本研究ではその中でも L^2 距離を用いて，実際の電子商取引サイトにおける顧客到着現象の実データを混合分布を用いてモデル化することが目的である．一般に事象の生起間隔の確率は指数分布に従っていることは知られている．今回用いる実データではそれぞれの顧客到着現象が指数分布に従っていることをデータの不完全性から断定することができない．そこで各顧客が指数分布に従って到着してい

a) 16rmd14@ms.dendai.ac.jp

ると仮定し，指数分布の和の分布であるガンマ分布を用いて，全体の到着現象をモデル化すること目指す．これにより，電子商取引サイトにおいて，顧客がどのような到着現象を行っているのかを混合分布を用いて明らかにする．

2. モデルについて

2.1 指数分布

本研究で考える顧客の到着現象は指数分布で考えることができる．顧客は前回の到着の影響を受けず，次の到着に影響を与えない（無記憶性）とする．この時、ある到着から次の到着までの時間間隔 t は指数分布に従い以下のように定義される．

$$f(x) = \lambda e^{-\lambda t} \tag{1}$$

このように，指数分布はある事象の生起間隔の確率を表す．生起間隔とは，ある事象が起こって次にまた発生するまでの間隔である．

2.2 ガンマ分布

指数分布に従う，独立な確率現象の和はガンマ分布で表すことができる．ガンマ分布は形状パラメータを k，尺度パラメータを θ を用いて (2) 式のように定義される．

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \tag{2}$$

k=1 の時，ガンマ分布はパラメータ θ を平均値とする指数分布を表す．顧客がパラメータ θ を平均値に持つ互いに独立な指数分布に従う時，指数分布の和がガンマ分布と考えることができる．

2.3 混合分布

混合分布は複数の分布の組み合わせを表現する際に用いられる．ガンマ分布における混合分布式は混合比率 w を用いて (3) 式のように表すことができる．

$$\begin{aligned} \hat{f_X}(x) &= w_1 f_X(x; k_1, \theta_1) + \ldots + w_n f_X(x; k_n, \theta_n) \\ &= \sum_{j=1}^{n} w_j f_X(x; k_j, \theta_j) \end{aligned} \tag{3}$$

3. 混合分布推定

3.1 データセット

本研究で使用したデータは「Happi+(ハピタス)」というポイントサイトの実データを用いる．

このサイトで行われたサービスの利用をコンバージョン (conv) として，登録された各ユーザーの利用した日数を計測している．2014 年 1 月 1 日から 2015 年 12 月 31 日までの間の 2 年間における約 4.6 万人のユーザーの利用データを用いる．このデータは顧客の個人情報は特定できないデータとなっている．

図 1 ポイントサイト「Happi+」

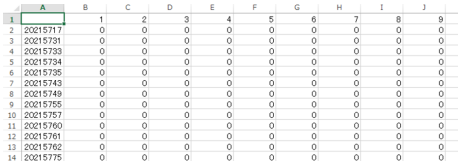

図 2 データセット

はじめにこのデータから各ユーザーごとに到着時間間隔が分かるようにデータを変換する．縦が各ユーザーを表し，横が日付で各ユーザーがコンバージョンしたかを表す．ユーザーの到着時間間隔を求めたデータから，それぞれの平均値を求めたその和の分布を求める．この際，コンバージョンした 2 回目以降の到着間隔を使用した．これは一度しか利用しなかったユーザーもいるため，平均とすることができないためである．これを満たすユーザーは約 3.5 万人で，このデータを本研究では利用した．

3.2 L^2 距離

本研究では，実データと混合分布モデルとの差異を評価する手法として L^2 距離を用いて行う．求めた 2 分布間の距離を Ev として評価を行う．L^2 距離の特徴として Sugiyama[1] で挙げられているように，非負性と対称性を持っており，数学的な距離を表している。それぞれの 2 確率分布が有界であれば、その差も有界である非常に安定性が高いという特徴がある．本研究で扱うデータはガンマ分布を想定しており，正の分布である．このことから非負性である L^2 距離は非常に適している．

次に L^2 距離を用いた Ev の求め方を (3) 式に示す．

$$\begin{aligned} Ev &= \int_i^n \left(f(x) - \hat{f_X}(x) \right)^2 dx \\ &= \sum_{i=1}^{n} \left(f(x_i) - \hat{f_X}(x_i) \right)^2 \\ &= \sum_{i=1}^{n} \left(f(x_i) - \sum_{j=1}^{m} w_j \hat{f_X}(x; k_j, u_j) \right)^2 \end{aligned} \tag{4}$$

Ev が低いほど実データのグラフと，モデルが一致するパラメータであることが推定できる．本研究では Ev*を最も

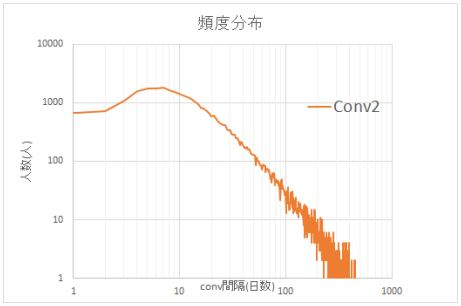

図 3 実データ

混合数	fx(k, θ ,w)
1	(2,7,1)
2	(9,6,0.1)(2,7,0.9)
3	(9,7,0.1)(7,1,0.1)(2,8,0.8)
4	(4,9,0.1)(7,9,0.1)(7,1,0.1)(2,7,0.7)
5	(3,4,0.1)(5,8,0.1)(7,9,0.1)(7,1,0.1)(2,7,0.6)

表 1 混合数と各パラメータ

Ev が低くなったパラメータとする．

3.3 混合分布推定手順

はじめに、各ユーザーごとに求めた平均到着間隔の日数の頻度を数える．日数の頻度を総データ数で割った物を頻度確率として，fx（実データ）とする．fx は経験確率密度分布である．この実データのグラフを図 3 に示す．
次に，(2) 式を用いて推定を行うガンマ分布の形状パラメータ k，尺度パラメータ θ をそれぞれ 1 から 10 の範囲を 1 刻みで設定する．また混合比率 w は 0.1 刻みで設定をする．これらの範囲で (k, θ ,w) を組み合わせた計算を行いそれぞれ Ev を求めていく．この操作を混合数 1 から混合数 5 までそれぞれ行う．この操作の中で一番 Ev が低くなった値を Ev*として各混合数ごとの Ev*を求めた．
ここで，図 1 の実データに注目すると，conv 間隔が 1 日の時と，100 日以降の時にばらつきが見られることが分かる．そこで，本研究では 2 日から 100 日にかけて範囲を限定して，混合分布推定を行う．(2) 式における i=0，n=100 として計算を行った．

4. 結果

4.1 混合数

混合数 1 から混合数 5 にかけての最も Ev が低くなった各パラメータ，比率は表 1 のように得られた．
混合数 1 における Ev*(2,7,1) と実データの比較図は図 4 のように得られた．
混合数 2 における Ev*(9,6,0.1)(2,7,0.9) と実データの比較図は図 5 のように得られた．
混合数 3 における Ev*(9,7,0.1)(7,1,0.1)(2,8,0.8) と実データの比較図は図 6 のように得られた．
混合数 4 における Ev*(4,9,0.1)(7,9,0.1)(7,1,0.1)(2,7,0.7) と実データの比較図は図 7 のように得られた．
混合数 5 における Ev*(3,4,0.1)(5,8,0.1)(7,9,0.1)(7,1,0.1)(2,7,0.6) と実データの比較図は図 8 のように得られた．
混合数 1 から混合数 5 までを合わせたグラフを図 9 に示す．並べてグラフをみて分かるように，混合数が増えるにつれて実データとモデルと徐々にフィッティングしている範囲が増えているのが分かる．

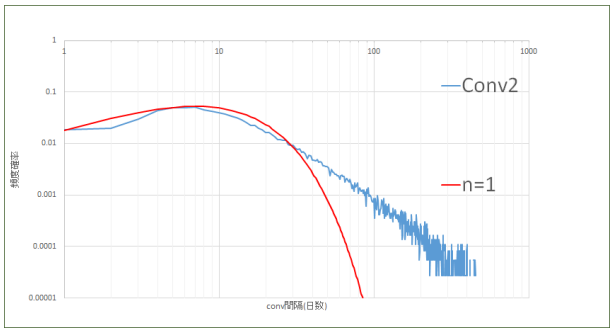

図 4 混合数 1

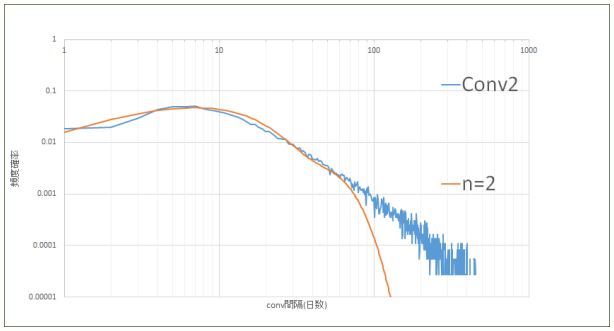

図 5 混合数 2

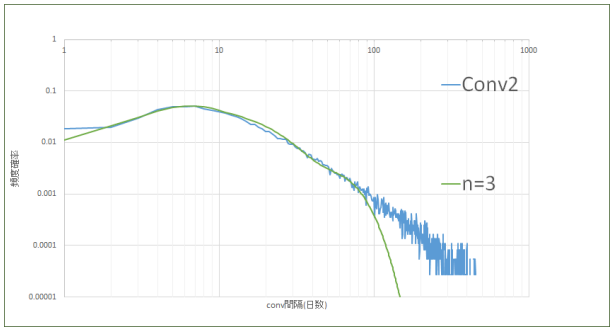

図 6 混合数 3

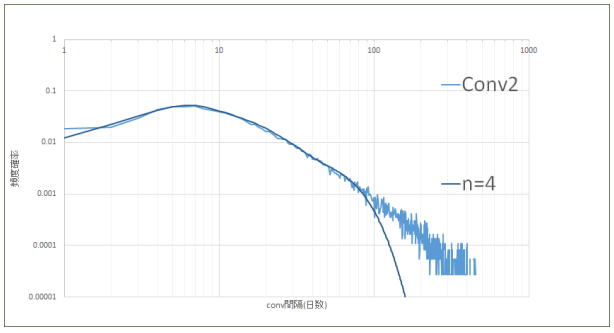

図 7 混合数 4

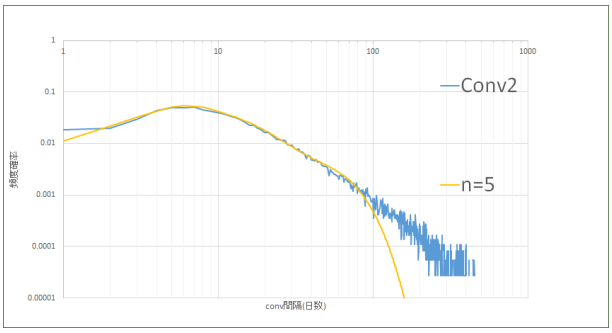

図 8　混合数 5

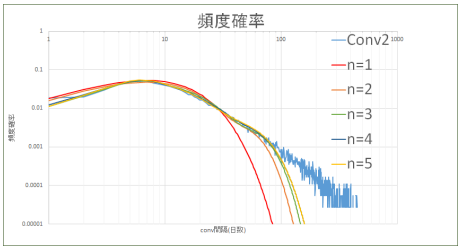

図 9　混合分布推定結果

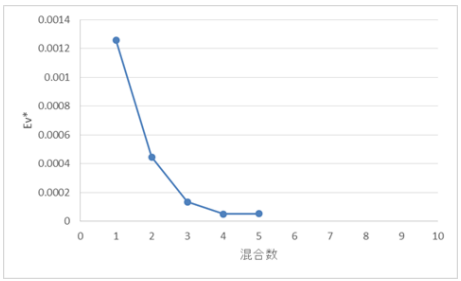

図 10　Ev*と混合数

混合数	Ev*
1	0.00125839531420286
2	0.000446109622700299
3	0.000134820782485993
4	0.0000513040384466298
5	0.0000543105650408125

表 2　混合数と Ev*

4.2　Ev*と推移

図 10 に Ev*の混合数 1 から混合数 5 までの推移グラフを示す．また表 2 に Ev*の値の一覧を示す．図 10 から見て分かるように，混合数の増加とともに Ev*は減少している．混合数 4 から混合数 5 にかけては微量の増加が見られた．

5.　考察

はじめに，各混合数におけるパラメータについて考えてみる．混合数 1 から混合数 5 にかけて必ず共通として fx(2,7) が共通して推定されている所に注目したい．fx(2,7) は常に混合比率においても 1.0，0.9，...，0.6 と，混合分布の大部分を占めている，この結果から大多数の顧客は，同じガンマ分布に従って，到着していると考えることができる．ここでは混合数 3 においては fx(2,8) となってはいるが、fx(2,7) ほとんど同じであると考えて良いだろう．次に，混合数 3 から混合数 5 にかけて fx(7,1) が共通して推定されている．また混合数 4，混合数 5 において fx(7,9) が共通して比率 0.1 で推定されている．こちらも同様に顧客到着現象の固定された集団であると考えることができると言えるだろう．このことから本研究の実データでは，顧客の大部分である 6,7 割は fx(2,7) のガンマ分布，加えてそれぞれ 1 割の顧客が fx(7,1)，fx(7,9) のガンマ分布に従って到着していると考えられる．合わせると約 9 割ほどの顧客は指数到着であると言えるだろう．

次に，図 9 の Ev*の推移と合わせて注目したい．混合数を増やすほどに詳細に分布のパラメータを設定できるため，混合数が多くなるほど，普通は実データのとの一致度合いが上がっていくと考えられる．結果からは混合数が増えていくに伴って、Ev*も減少し，実データとの一致度合いも上がっているのが分かる．しかし図 7 の結果から見て分かるように，EV が徐々に下がっていったのち，混合数 5 から上昇したことが分かる．混合数が 5 になった時，EV が混合数が 4 の結果よりも高くなったのである．つまり，ここでは混合数が 4 の時，図 4 で示した結果のように一番フィッティングが高い結果を示したのである．このことから混合数が増えるごとに EV 以降は上昇する，または Ev*の値は停滞するだろうと考えられる．従って混合数 4 が本研究で用いた実データにおいて一番良い混合数であると言えるだろう．

6.　結論

以上の結果から混合分布を用いて電子商取引サイトにおける顧客到着現象を概ねモデル化できたと言えるだろう．混合数が増えるほどに，Ev がより 0 に近づき，実データのとフィッティング度合いも上がっていることが結果から示されており，L2 距離を用いて混合分布を推定することができたと言える．その中でも混合数 4 の時が実データと一番フィッティング性のあるグラフが示された．パラメータ推定手順において説明したように，本研究ではデータの仕様から 2〜100 日までのデータに限定して推定を行っている．本研究の目的である顧客到着現象のモデル化は，この範囲内に限定してできていると言えるだろう．

上述の通り 1 日目と 101 日目以降においてはモデル化ができていない．これに関しては二点理由があげられる．まず一つ目はデータの不足によるためである．不足した顧客到着現象のデータから平均到着時間間隔の頻度分布を導出したが，頻度分布に関しても穴あきが生じている．そのた

めデータの正確性を測るために本研究では推定範囲を限定することにしていた．二つ目は，データにべき性がある可能性があるためである．101 日目以降においては，さらに長い期間のデータ扱うことを考えれば，長い期間利用しないユーザーが発生しうる可能性は大いにあるだろう．そうすると，一回の到着間隔が 300 日，500 日... などと増えた場合に平均到着時間は長くなり，大きな値がまばらに発生するだろう．そのため，101 日目以降に関してはべき分布特性などを含めたモデル化を行うことが必要になるだろう．

謝辞　本論文は著者が東京電機大学大学院理工学研究科情報学専攻在籍中に得た研究成果をまとめたものである．研究にあたって，指導頂いた上浦基助教に心からの感謝を表する．また同じ研究室で多くの議論を交わした同大学院所属の関根氏を始めとする研究室の方々に対しても心からの感謝を表する．

参考文献

[1] Sugiyama, DistanceApproximationbetweenProbability-Distributions ：RecentAdvancesinMachine Learning 日本数理応用学会論文誌,

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, maximum likelihood from incomplete data via the EM Algorithm *Journal of the Royal Statistical Society. Series B (Methodological) Vol. 39, No. 1 (1977), pp. 1-38*,