日本統計学会誌
第 43 巻, 第 1 号, 2013 年 9 月
41 頁 ～ 58 頁

特　集

継続時間と離散選択の同時分析のための変量効果モデルとその選択バイアス補正
—Web ログデータからの潜在顧客への広告販促戦略立案—

星野　崇宏*

Joint Random Effect Modeling for Repeated Durations and Discrete Choices with Selection Bias Correction: Application to Promotion Policy Planning for Potential Clients Using Web Access-log Data

Takahiro Hoshino*

ビッグデータの文脈ではほとんど取り上げられていないが，統計学上そして実用上非常に重要な問題として消費者の異質性と選択バイアスがある．本論文では具体例として，1 万 3000 人におよぶ代表性を有する対象者から得られた，年間 1.5 億件ほどのインターネットへのアクセスログデータを利用する．特定のオンラインショッピングサイトに着目し，その広告販促戦略を立案するために，当該サイトへの訪問間隔の個人差と購買確率の高さを表現する変量効果を用いた繰り返しのある継続時間と購買有無の同時分析モデルを提案する．また，競合サイトを利用しているが当該サイトは利用していない潜在顧客に対してどのような広告販促を行うことが有用かを調べるために，潜在顧客における母数推定を行うための共変量調整法として傾向スコアによる重み付き一般化 EM アルゴリズムによる推定法を提案する．提案モデルは個人差を説明する変量効果を有するため，E ステップの計算において積分計算が必要となるが，大規模データではマルコフ連鎖モンテカルロ法などの確率的な数値積分が現実的でない．本研究では近年再び注目を浴びている完全指数型ラプラス近似を利用することで，短時間で大規模データの解析を精度よく行うことができ，実務上示唆のある結果を得ることを示す．

We point out that the current researches in Big data analytics neglect two important issues, consumer heterogeniety and selection bias, in constructing predivtion modeling. We provide a real example in which we must deal with the two issues properly, joint modeling of repeated duration and pruchase behavior. To be more concrete, we apply the joint modeling to a Web access-log dataset from a very large panel study. To plan a promotion policy for potential clients of a online shopping company, we proposed a propensity score weighted generalized EM algorithm of the proposed model, to adjust for covariate differences between potential clients and current clients. The proposed model incorporates random effects expressing unmeasured heterogeniety, which inevitably requires numerical integration. However in large dataset it is not practical to employ Markov chain Monte Carlo methods in random effect modeling. We applied the fully exponential Laplace approximation to the estimation algorithm of the proposed model, found that the algorithm is less computationally expensive, while it provides accurate estimates.

キーワード: ビッグデータ，完全指数型ラプラス近似，個人異質性，選択バイアス，共変量シフト，一般化 EM アルゴリズム，傾向スコア

* 名古屋大学大学院経済学研究科：〒 464-8601 愛知県名古屋市千種区不老町 (E-mail: hoshino@soec.nagoya-u.ac.jp).

1. 問題意識と研究目的

総務省 (2012) はビッグデータを「多量性，多種性，リアルタイム性等が挙げられる」各種データのことと定義し，「異変の察知や近未来の予測等を通じ，利用者個々のニーズに即したサービスの提供，業務運営の効率化や新産業の創出等が可能となる点に，ビッグデータの活用の意義がある」としている．これまでも工学系の学会を中心にビッグデータに関する研究発表は非常に盛んに行われているが（例えば岩爪 (2013) をはじめとする人工知能学会誌の 2013 年 1 月の特集号など），そこでの論点はビッグデータの収集と処理（特に分散処理とリアルタイム処理）に関するものがほとんどであり，総務省の定義によるところでは多量性とリアルタイム性のみが取り上げられ，データの含む多種性についての議論はほとんどなされていない．また予測方法としてはこれまでは機械学習手法の応用が中心であった．

ビッグデータをビジネス分野や公衆衛生・教育・交通行動・減災など行政分野で利活用する場合，データの多種性を考慮した予測への適用が重要である．ここで多種性としてこれまでは得られる変数の多様さが注目されがちだが，実際に個人の行動や意思決定を理解し予測する場合には個人間の違い＝「個人の異質性」が非常に重要である．実際，購買履歴データや医療機関への受診データ，交通行動データなどは，市民や学生，通勤者など個人の行動の結果を反映したものであり，その個人差は極めて大きい．

個人の異質性は個人ごとの予測を行う際に考慮する必要があるだけでなく，情報を得ることができる対象者集団がどのような個人によって構成されるのかによってデータの内容が大きく変わり，そこから得られる解析結果も大きく異なる，という問題にもつながる．本研究の解析例でも示すように，しばしば本来関心のある母集団についてのデータを得ることが不可能である場合は多いが，個人の異質性が大きい場合には得られたデータだけから単純な解析をすると関心対象の母集団の理解を誤る，いわゆる選択バイアスの問題にも直結する．

これまで経済学や経営学，社会学，心理学などの社会科学では個人の異質性の取り扱い，およびデータの選択バイアスの問題の 2 点を無視した解析によって様々なバイアスが生じるということが重要な問題として認識されてきた (例えば星野 (2009)，Angrist and Pischke (2009))．個人の異質性については，マーケティングや計量心理学，生物統計学などの分野では個人の嗜好や意思決定，行動の違いを説明するモデルとして階層モデルや変量効果モデルが発展してきた．計量経済学では，同一個人のデータが繰り返し得られるときに，固定効果や変量効果をモデルに含めることで個人の異質性を除去した解析が行われてきた．また，選択バイアスの問題については計量経済学や疫学では非常に重要な問題としてその除去方法について研究デザインや解析手法の発展がなされてきた．しかしこれらの観点はビッグデータの文脈ではこれまで議論されてこなかった．

本論文ではビッグデータの例として，1 万 3000 人におよぶ代表性を有する調査対象者から得られた，年間 1.5 億件ほどのインターネットへのアクセスログデータを利用する．具体的にはあるオンラインショッピングサイトへの訪問間隔情報と購買有無の情報に対して，個人の異質性を表す変量効果を共通の説明変数とする，繰り返しのある継続時間（生存時間）と購買有無の同時分布のモデリングを適用する．ここで変量効果を導入することで，観測できない異質性に関連する共変量の影響を除去し，関心のある説明変数（具体的にはどのサイトから当該ショッピングサイトを訪問したか）と訪問間隔や購買有無の関係をより精度よく推論することが可能となる．さらに，変量効果は訪問間隔の個人差として理解できるため，訪問間隔の頻繁さが各訪問回ごとの購買確率に影響を与えているかを調べることも可能である．

近年のコンピュータの処理速度の向上により，一般化混合線形モデルや観測変数が離散変数である場合の変量効果モデル，因子分析モデル，動的離散選択モデルなど尤度の計算に積分が必要となるモデルでは，マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo 法：以後 MCMC 法）を用いたベイズ推定が一般的に利用されている．MCMC 法では変量効果や因子，効用などの潜在変数の乱数発生を行う必要があるが，潜在変数は各オブザベーションごとに発生される必要がある．従ってサンプルサイズの増加に従って計算時間が線形に増加するため，MCMC 法は大規模データでは実用的ではない．大規模データの推定においては MCMC 法などと比べて計算量が大幅に少ない近似法である変分ベイズ法 (Jordan *et al.* (1999)) が利用されることがあるが，変分ベイズ法は分散を過小評価するなど近似の精度が低いことが知られている (例えば Wang and Titterington (2004)，Braun and McAuliffe (2010))．

そこで本研究ではラプラス近似を用いた数値積分による解析法を利用する．ラプラス近似自体は積分の近似法として 18 世紀から知られており，統計学でも Tierney and Kadane (1986) によって事後分布のモーメントと分布自体の近似法として提案された古典的な手法の一つである．この 20 年ほどの MCMC 法の発展により，ラプラス近似に関する研究は一時下火となっていたが，大規模データ解析のニーズの高まりから，MCMC 法よりも数値計算量の負荷がはるかに少なく，かつ数理的に性質の明確なラプラス近似を用いた推定法についての研究がここ数年で増加している．例えば Rue *et al.* (2009) は正規分布に従う変量効果モデル（正確にはより一般的な latent Gaussian model）でのラプラス近似の新しい方法として積分段階的ラプラス近似法 (integrated nested Laplace approximation: INLA) を提案し，Fong *et al.* (2010) は一般化混合線形モデルへの INLA の応用を行っている．また Rizopoulos *et al.* (2009) では生存時間と継時データの同時分布に対して，Bianconcini and Cagnone (2012) では一般化線形潜在変数モデルに対して，完全指数型ラプラス近似 (Tierney *et al.* (1989)) を利用した最尤推定法を提案している．また小山 (2012) では一般

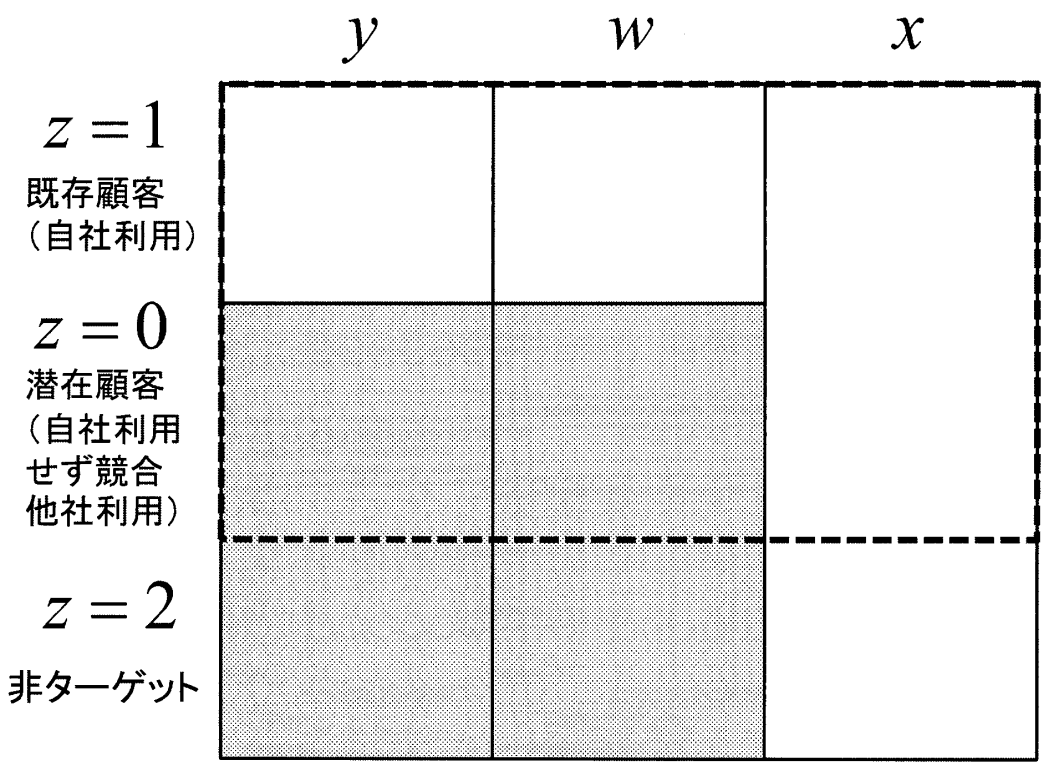

図1　欠測データとしての潜在顧客行動理解での選択バイアス．*灰色は欠測データであることを示す．

化線形モデルの期待値が状態空間モデルに従うような時系列構造を持つモデルを考え，ラプラス近似による推定法を提案している．

このような研究動向と実務的な要請から，本研究では提案モデルの推定値を得るために，完全指数型ラプラス近似を利用したEMアルゴリズムを導出する．これにより今回の解析例のように数十万以上のオブザベーションでも数値積分の必要な推定値の計算が短時間で実行可能となる．

またここで，実務上の関心として，そのWebサービスをすでに利用している消費者（図1の$z=1$）に対してどのように最適なマーケティング施策を実施するかにある場合も多いが，企業側のターゲットになってはいるが実際にはまだ利用していない（または利用回数が頻繁ではない）潜在顧客（図1の$z=0$）にある場合がある．これは(1)そのサービスが普及するにつれて，どのような消費者側の要因，あるいは企業側の行動が利用頻度を早めるかを理解することは，今後企業がどのような広告販促等のマーケティング活動を将来的に実施するべきか，その費用対効果はどの程度か，を考え将来の投資に備えるために必要である，(2)潜在顧客として競合他社の顧客を想定する場合，競合他社からいかに顧客を奪うための効率的なマーケティング活動を知ることが重要となる，などの理由による．しかし当然ながら，$z=1$でのデータの解析を行うだけでは既存顧客についてのマーケティング戦略立案にしかならず，潜在顧客についてどのような行動を行えばよいかについての情

報はわからない．そこで本論文ではこの問題を潜在顧客 $(z=0)$ で利用頻度や購買有無のデータが得られていない欠測データの問題として定式化し，潜在顧客 $(z=0)$ に対するモデル推定を行うために，重み付きの一般化 EM アルゴリズムを用いた重み付き最尤推定を行う．

第 2 節では本論文のモデル設定を説明し，同時分布および基準関数となる重み付き対数尤度を提示する．第 3 節では完全指数型ラプラス近似で E ステップの期待値計算の近似を行う EM アルゴリズムの詳細を示す．第 4 節では解析例のデータの説明と結果の提示を行い，第 5 節ではいくつかの議論を行う．

2. モデル設定

2.1 同時分布の表現

本研究では繰り返しのある継続時間分析 (recurrent event duration analysis, Seethraman and Chintagunta (2003), Bijwaard *et al.* (2006)) または生存分析 (recurrent event survival analysis) モデルと購買有無についてのロジスティック回帰モデルの同時モデルに消費者異質性の変数を共通の説明変数として導入した同時モデルを提案する．

Web マーケティングではパフォーマンスの指標（Key Performance Indicator; KPI）として商品購入や会員登録，資料請求などを含めた概念であるコンバージョンとともに，サイト閲覧時間の長さや，訪問間隔などが単独の KPI として取り上げられるが，複数の KPI，たとえば購入と閲覧間隔の同時分布をモデリングすることで，様々な共変量の影響を除去した上での閲覧間隔の平均的な長さと購入確率の関係を理解することが可能になる．

具体的には y_{ij}^D を個人 i が第 j 回目に特定サイトへ訪問した時から第 $j+1$ 回目に訪問するまでの時間間隔とする．そして $j+1$ 回目でそのサイトからの購入を行った場合に 1，購入していない場合には 0 となる 2 値変数を y_{ij}^B とする．また，個人 i が解析の対象となっている時間幅においてそのサイトを訪問した回数を J_i とする (従って $j=1,\ldots,J_i$ となる)．

ここで個人 i の共変量（以降個人共変量と呼ぶ）の値を $\boldsymbol{x}_i$，個人 i の j 回目から $j+1$ 回目の訪問の間に特定の値を持つ時変共変量を $\boldsymbol{w}_{ij}$，個人 i の訪問間隔の個人差を表す変量効果を f_i とし，f_i は正規分布に従う $f_i \sim N(0,\phi)$ とする．そして訪問間隔がこれらの変数によって説明されるモデルを考える．また，訪問回数が十分複数回観測されるように時間幅を十分にとることで，打ち切りの問題は考えないとする．

本研究では比例ハザードモデルと加速故障時間モデル両方の性質を満たす唯一のモデルであるという利点を有すること，およびマーケティングモデルとしても先行研究で利用されることが多いため，実務的な有用性という観点からワイブルモデルを利用する．具体的には y_{ij}^D がワイブル分布に従うとするが，モデル解釈の明確さを考え，$y_{ij}^{LD}=\log y_{ij}^D$ が以

下の確率密度関数を持つ（ワイブル確率変数の対数であるためガンベル分布に従う）

$$f\left(y_{ij}^{LD}|f_i,\boldsymbol{w}_{ij}\right)=\frac{1}{\sigma}\exp\left[\frac{y_{ij}^{LD}-(\beta_0+f_i+\boldsymbol{w}_{ij}^t\boldsymbol{\beta}_w)}{\sigma}-\exp\left(\frac{y_{ij}^{LD}-(\beta_0+f_i+\boldsymbol{w}_{ij}^t\boldsymbol{\beta}_w)}{\sigma}\right)\right] \tag{2.1}$$

とする (Klein and Moeschberger (2003)). これにより，位置スケールモデルとして表現が可能となり，この分布のモードが $\beta_0+f_i+\boldsymbol{w}_{ij}^t\boldsymbol{\beta}_w$，平均はモードにオイラーの乗数 $\times\sigma$ を加えたものであり，それぞれ説明変数の線形関数で表現できるという利点がある．

さらに $j+1$ 回目のサイト訪問での購入を表す変数 y_{ij}^B はロジスティック回帰モデルに従い，訪問間隔の個人差を表す関数 f_i や共変量に依存する

$$\text{logit}\left[p\left(y_{ij}^B=1|f_i,\boldsymbol{w}_{ij}\right)\right]=\alpha_0+\alpha_f f_i+\boldsymbol{w}_{ij}^t\alpha_w \tag{2.2}$$

とする．

このとき，繰り返しのある継続時間と購買有無の同時分布は

$$p(y_{i1},\ldots,y_{iJ_i}|\boldsymbol{w}_{i1},\ldots,\boldsymbol{w}_{iJ_i})=\int\left(\prod_{j=1}^{J_i}p\left(y_{ij}^{LD}|f_i,\boldsymbol{w}_{ij}\right)p\left(y_{ij}^B|f_i,\boldsymbol{w}_{ij}\right)\right)p(f_i)df_i \tag{2.3}$$

と表現される（但し $y_{ij}=(y_{ij}^{LD},y_{ij}^B)^t$ とする）．ここでパラメータ α_f が小さいほど「サイトへの訪問頻度が高いほど各訪問回において購入する確率が高まる」ことを示す係数と考えることができる．

2.2 選択バイアスの除去

但し，実際には訪問間隔についての情報はその企業サイトを現在利用している消費者についてしか得られない．単に母集団（あるいはターゲット消費者の集団）全体についてのデータが得られていないというだけではなく，特にマーケティングなど実務的な観点からは第1節で述べたように，$z_i=0$ の対象者（期間中に利用していない消費者）についての予測モデルを構築することが重要な場合がある．そこで，z_i を個人 i が解析に利用する時間幅においてある企業サイトを利用している（つまり既存顧客である）場合には1，利用していない（つまり潜在顧客である）場合には0をとるインディケーター変数とすると，$z_i=1$ では $y_{i1},\ldots,y_{iJ_i}$ が観測され，$z_i=0$ では観測されないという欠測問題に帰着する．本研究では欠測は観測可能な個人共変量 $\boldsymbol{x}$ のみに依存し，欠測のある y や $\boldsymbol{w}$ には依存しないという，いわゆるランダムな欠測 (Missing at random) を仮定する．この仮定の下では，下記の重み付き対数尤度

$$\sum_{i=1}^{N}\frac{z_i(1-w(\boldsymbol{x}_i))}{w(\boldsymbol{x}_i)}\log p(y_{i1},\ldots,y_{iJ_i}|\boldsymbol{w}_{i1},\ldots,\boldsymbol{w}_{iJ_i}) \tag{2.4}$$

の最大化によって得られる推定量は $z=0$ を条件付けた場合の同時分布 $p(\boldsymbol{y}|\boldsymbol{w}, z=0)$ のパラメータの一致推定量となる (星野 (2005), Hoshino *et al.* (2006), Wooldridge (2007), Pan and Schaubel (2009)). 但しここで $w(\boldsymbol{x}_i)$ は $\boldsymbol{x}_i$ を所与としたときに $z_i=1$ となる確率である. ここで, 従属変数 y の説明変数として共変量も利用することができる場合, モデルが正しければ上記の重みづけを行わずに観測データの尤度を最大化することで一致性のある最尤推定量を得ることができる. しかし本研究では潜在顧客 $(z=0)$ については $\boldsymbol{w}$ のどの要素が y に影響を与えるかを推論することが解析の目的であり, $\boldsymbol{x}$ と y の回帰関係には関心がない状況である. 従って y と $\boldsymbol{w}$ の周辺回帰モデルの母数推定を行うために $\boldsymbol{x}$ の分布について選択バイアスの除去を行うための重み付き推定を利用している.

また, 第 4 節において紹介する解析例で利用したデータは調査対象者の Web アクセスログすべてを取得しているパネルデータであるため, 実際には潜在顧客 (図 1 の $z=0$) と定義された調査対象者についても y, w や x が得られている. 一方, 実務では自社サイトで実際に商品やサービスを購入した既存顧客 $(z=1)$ についてしかこれらの値が得られない. しかしこのような場合であっても, マーケティングリサーチから, または公表されている統計資料を用いるなどにより潜在顧客における個人共変量の同時分布 $p(x|z=0)$ を知ることができれば,

$$w(\boldsymbol{x}_i)=\frac{p(x_i|z_i=1)p(z_i=1)}{p(x_i|z_i=1)p(z_i=1)+p(x_i|z_i=0)p(z_i=0)} \tag{2.5}$$

から $w(\boldsymbol{x}_i)$ を計算することができる.

2.3 統計モデルとしての新規性

ここで上記のモデルのうち継続時間だけの解析としては, Vaida and Xu (2000) が clustered data に対する変量効果のあるモデルを提案しているが, 彼らの推定法はモンテカルロ積分による EM アルゴリズムを利用している. また Rizopoulos *et al.* (2009) は各対象者ごとに 1 回だけの継続時間と複数回の反復測定から得られる連続変数の同時分布のモデリングを行っている. 本研究のモデルでは上記に示したような Web マーケティングでの問題意識に即して, 繰り返しのある継続時間と購買を表す 2 値変数との同時分布を解析の目的としていること, および選択バイアスの問題を考慮している点が先行研究と異なる.

3. 完全指数型ラプラス近似を用いた一般化 EM アルゴリズム

以後記法の簡便化のために, $\boldsymbol{\alpha}=(\alpha_0, \alpha_f, \boldsymbol{\alpha}_w^t)^t$, $\boldsymbol{\beta}=(\sigma, \beta_0, \boldsymbol{\beta}_w^t)^t$, $\boldsymbol{\theta}=(\boldsymbol{\alpha}^t, \boldsymbol{\beta}^t, \phi)^t$ とする.

ここで重み付き最尤推定量を得るためのスコアベクトルは

$$S(\boldsymbol{\theta}) \equiv \sum_{i=1}^{N} \frac{z_i(1-w(\boldsymbol{x}_i))}{w(\boldsymbol{x}_i)} S_i(\boldsymbol{\theta})$$
$$= \sum_{i=1}^{N} \frac{z_i(1-w(\boldsymbol{x}_i))}{w(\boldsymbol{x}_i)} \int \frac{\partial \log g(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} p(f_i|\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{\theta}) df_i \tag{3.1}$$

但し

$$S_i(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{i1}, \ldots, y_{iJ_i}|\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iJ_i}) \tag{3.2}$$

$$g\left(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta}\right) = \left(\prod_{j=1}^{J_i} p\left(y_{ij}^{LD}|f_i, \boldsymbol{w}_{ij}\right) p\left(y_{ij}^{B}|f_i, \boldsymbol{w}_{ij}\right) \right) p(f_i|\phi) \tag{3.3}$$

であり

$$p(f_i|\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{\theta}) = \frac{g(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta})}{\int g(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta}) df_i}$$

となる．またさらに

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log g(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{j=1}^{J_i} \log p\left(y_{ij}^{B}|f_i, \boldsymbol{w}_{ij}\right)$$
$$+ \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{j=1}^{J_i} \log p\left(y_{ij}^{LD}|f_i, \boldsymbol{w}_{ij}\right) + \frac{\partial}{\partial \phi} \log p(f_i|\phi)$$

となる．完全データのスコア関数は負値を取る可能性があるため，この事後期待値を計算するためにモーメント母関数を用いた完全指数型ラプラス近似 (Tierney *et al.* (1989)) を利用すると，$S_i(\boldsymbol{\theta})$ の r 番目の要素は

$$\hat{S}_{ir}(\boldsymbol{\theta}) = \frac{\partial \log g\left(\hat{f}_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta}\right)}{\partial \theta_r} - \frac{1}{2}\gamma_{ir} \tag{3.4}$$

によって $O(J_i^{-2})$ のオーダーで近似できる (Rizopoulos *et al.* (2009))．但し θ_r は $\boldsymbol{\theta}$ の r 番目の要素，

$$\gamma_{ir} = \boldsymbol{\Sigma}_i^{-1} \left\{ \frac{\partial \boldsymbol{\Sigma}_i}{\partial f_i} \boldsymbol{\Sigma}_i^{-1} \frac{\partial^2}{\partial f_i \partial \theta_r} \log g(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta}) - \frac{\partial}{\partial \theta_r} \boldsymbol{\Sigma}_i \right\} \Bigg|_{f_i=\hat{f}_i} \tag{3.5}$$

$$\boldsymbol{\Sigma}_i = -\frac{\partial^2}{\partial f_i^2} \log g(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i, z_i|\boldsymbol{\theta}) \Big|_{f_i=\hat{f}_i} \tag{3.6}$$

である（本モデルでの上記の具体的な式は補遺参照）．

本研究では以下のステップによる EM アルゴリズムを用いて重み付き最尤推定値を得る．

(1) 初期値を求める．

(2) 1ステップのニュートンラフソン法を用いて式(3.3)を最大化する f_i を $\hat{f}_i$ とする(ニュートンラフソン法に必要な $\log g(f_i, \boldsymbol{y}_i, \boldsymbol{w}_i|\boldsymbol{\theta})$ の 1，2，3 次導関数は補遺に記載).

(3) 完全指数型ラプラス近似で式 (3.1) のスコアベクトルの期待値計算の近似を行う.

(4) スコアベクトルの事後期待値がゼロになる $\boldsymbol{\theta}$ の値を $\hat{\boldsymbol{\theta}}$ とする.

(5) 上記 (2) から (4) までを収束判定基準が満たされるまで続け，満たされた場合には終了する.

シミュレーションによる提案推定法の精度のチェック

上記の提案アルゴリズムの精度を調べるために，シミュレーションを実施した．但しここでは完全指数型ラプラス近似による一般化 EM アルゴリズムの推定精度及び計算速度を MCMC と比較することが目的であり，$S(\boldsymbol{\theta})$ の近似が正確であるかどうかがわかればよいので，選択バイアスは考慮しない状況，つまりすべての対象者での重み $w(\boldsymbol{x}_i)$ は 1 とした．簡単のため w_{ij} は 2 次元でそれぞれ独立を仮定し，そのうち 1 つは確率 0.7 で 1 を取るダミー変数，1 つは標準正規分布に従うとした．サンプルサイズは $N = 5{,}000$ および 100,000 の 2 パターンとした．実際は解析の時間幅と各対象者の訪問間隔に関連するパラメータが決まれば J_i は自動的に決まるが，シミュレーションでは時間幅を設定する意味がないので，ここでは J_i を (1) 平均が 10 となるように下限 3，上限 17 の一様乱数から発生，(2) 平均が 50 となるように下限 5，上限 95 の一様乱数から発生させる 2 つのパターンで発生させた．結果として 4 パターンのシミュレーションを行った．表 1 には各パラメータの真値，提案アルゴリズムと MCMC による推定値，および標準誤差を記載した．MCMC については iteration を 3000 回とし，うち 1000 回を Burn-in phase として除去し，残った乱数列についてすべてのパラメータについて Geweke の収束基準が満たされることを確認した．こ

表 1　シミュレーションによる提案推定法 MCMC の比較.

			σ	β0	βw1	βw2	α0	αf	αw1	αw2	ϕ
		真値	1	3	0.5	0.5	−0.5	0.5	0.3	0.3	1
N=5000 平均10	Laplace	推定値	1.0798	2.9209	0.4803	0.5130	−0.4890	0.4829	0.2795	0.3205	1.1493
		標準誤差	0.2414	0.2857	0.1931	0.1954	0.2272	0.2069	0.1630	0.1723	0.2464
	MCMC	推定値	1.0487	2.9707	0.4930	0.5057	−0.5063	0.4930	0.2912	0.3116	1.0504
		標準誤差	0.2519	0.2924	0.1974	0.1982	0.2528	0.2228	0.1768	0.1778	0.2676
N=5000 平均50	Laplace	推定値	1.0319	2.9710	0.5020	0.4977	−0.4951	0.4954	0.3069	0.2973	1.0403
		標準誤差	0.1094	0.1378	0.0843	0.0834	0.1109	0.0941	0.0761	0.0788	0.1097
	MCMC	推定値	1.0268	2.9751	0.5021	0.4981	−0.4960	0.4966	0.3071	0.2971	1.0429
		標準誤差	0.1127	0.1355	0.0846	0.0904	0.1101	0.0980	0.0822	0.0774	0.1185
N=100000 平均10	Laplace	推定値	1.0539	3.1210	0.4851	0.5101	−0.5088	0.4877	0.3077	0.2985	1.1362
		標準誤差	0.0534	0.0688	0.0424	0.0415	0.0506	0.0442	0.0344	0.0368	0.0532
	MCMC	推定値	1.0282	3.0798	0.4910	0.5051	−0.5051	0.4981	0.3058	0.2983	1.0336
		標準誤差	0.0543	0.0695	0.0437	0.0453	0.0559	0.0476	0.0380	0.0414	0.0560
N=100000 平均50	Laplace	推定値	0.9918	3.0567	0.5020	0.5022	−0.4985	0.5029	0.3029	0.2998	1.0250
		標準誤差	0.0253	0.0305	0.0186	0.0197	0.0227	0.0206	0.0164	0.0171	0.0244
	MCMC	推定値	0.9950	3.0643	0.5011	0.5003	−0.4987	0.5033	0.3025	0.2986	1.0260
		標準誤差	0.0245	0.0318	0.0191	0.0201	0.0249	0.0220	0.0172	0.0184	0.0267

の結果からは，提案アルゴリズムは J_i の平均が 10 の場合には精度がやや落ちる（但し真値を棄却することはない）が，50 の場合には MCMC とほぼ同等の精度があることがわかる．また MCMC では標準誤差が提案アルゴリズムよりも若干大きく出るが，ベイズ推定では漸近理論を用いておらず，他のパラメータの推定の不確実性も考慮した標準誤差が算出されていることが原因と考えられる．上記のすべての推定は SAS/IML 上でコードを作成して Window7 64bit，Intel Core i7-3930K (6 コア・12way/3.20GHz/3.80GHz/12MB)，メモリ 32GB のコンピュータで実行した．推定に要した時間はもっともデータ量の多い $N = 100{,}000$，J_i の平均 50 の条件で提案手法では 10.2 時間であったのに対して，MCMC では iteration が今回 3000 回と少ないにも関わらず 233.1 時間を要した (4 条件での推定時間の比は約 18 倍から 23 倍の間であった)．本モデルに対する MCMC では個人ごとの潜在変数の発生を行っている部分がもっとも計算量を増大させている部分と考えられ，これはほぼサンプルサイズに比例する．提案方法でも個人ごとの潜在変数の推定値計算は行っていることから，結局 MCMC の 3000 回の iteration が数十回のニュートンラフソンに置き換わったことで時間が短縮されたと考えてよい．

これらの結果からは，少なくともこのモデルにおいては MCMC を利用することによる推定精度の向上は，推定時間の増加の割には大きくなく，提案アルゴリズムで十分精度の高い推定を比較的短時間で実行できることがわかった．

4. オンラインショッピングサイトへの訪問間隔と購買有無の同時解析：個人異質性と選択バイアスの考慮

本研究で利用したデータは株式会社ビデオリサーチインタラクティブ社が提供する『インターネット視聴データ』である．『インターネット視聴データ』は，約 13000 人（期間によって多少変動する）が参加するパネル調査であり，調査参加者が自宅のコンピュータから閲覧した全ての URL，サイト閲覧日時，サイトの閲覧時間の長さ，リファラーデータ（どの URL からその URL を閲覧したか）が記録されている web ログデータである．また調査参加者に対しては年に 1 回別途郵送調査も実施していて，その調査ではデモグラフィック属性や商品保有状況，ライフスタイルといった変数も取得しており，2011 年の郵送調査では 9708 人の計画標本に対して 7116 人が回答している．調査参加者は乱数番号法（Random Digit Dialing; RDD）を用いて選ばれており，インターネット閲覧履歴データとしては日本において唯一代表性が担保されているデータであるといえる．また調査参加者は 1 年間継続してこの調査に参加するが，各月ごとに 12 分の 1 のサンプルが入れ替えられるという調査デザインになっている．

本研究では特に，代表的なオンラインショッピングサイトである A 社に着目し，サイトへの訪問間隔を繰り返しのある継続時間ととらえた．また，購買確認画面の URL を特定

し，その URL に対する各個人のアクセス間隔を購買有りとみなした．訪問間隔については，同じ日の訪問であっても，30 分以上後になって別サイトから訪問した場合については，別途の訪問とみなした．

本研究では，2010 年 4 月から 2012 年 2 月までの期間に調査に参加した調査対象者を対象とし，このパネル調査に参加している一年のうち前半で A 社のサイトを 3 回以上訪問しかつ 2 回以上購入利用している調査対象者を “既存顧客”($z = 1$) とし，既存顧客の定義にあてはまらず競合大手オンサインショッピングサイト B 社でこの期間に 2 回以上購入しており，かつ一年のうち後半で A 社で 3 回以上訪問しかつ 2 回以上購入を行った対象者を潜在顧客 ($z = 0$) とした．結果としてここで利用したアクセス数は 196 万 2258 件であった．

また，ここでの個人共変量 $\boldsymbol{x}$ として年代 (5 区分)，性別，居住地域 (9 地域)，職業区分 (6 カテゴリー)，インターネット利用時間の平均，家族構成 (未婚，夫婦だけ，乳幼児のいる夫婦，乳幼児はいないが子供と同居している夫婦，3 世帯家族)，世帯収入区分 (13 分類)，時変共変量 $\boldsymbol{w}_{ij}$ として当該訪問時のアクセス時間（24 時間を 4 分割），経路情報（その訪問時にどのサイトから当該サイトに訪問したか）を利用した．但しこれらの情報だけでは各調査対象者の訪問間隔と購買確率の異質性を説明できない可能性が高いため，個人内で同一の値をとる変量効果を導入する．

またここで経路情報を関心の対象となる時間共変量として利用するのは，どのサイトから自社サイトに流入したかという情報が，企業のバナー広告や検索連動広告への広告費の最適配分や，アフィリエイトサイトへの報酬の決定など企業の Web 上での広告販促戦略に直結するからである．経路情報としては具体的に検索エンジン 3 分類 (yahoo，google，その他)，blog サイト，SNS サイト (twitter，facebook，mixi)，メール，情報サイト (価格コムなど)，その他の 8 カテゴリーを利用した．但し，本来は各 URL ごとに実際にどのような内容のページから遷移して A 社のサイトに流入してきたかを調べる必要があるが，ここでは機械的にドメインレベルで分類している．同様に購入有無も，URL に特定の文字列が現れるかどうかを利用して分類を行っている．またここでサイトへの訪問者はページ閲覧を通常連続で行うので，訪問経路情報は前のページとなり，同じドメインを持つことになるが，ここではこれらの同ドメイン内の訪問経路情報は外し，初めて A 社のサイトに訪問した際の訪問経路を訪問経路情報と定義する．この解析例の問題設定として注意したいこととして，既存顧客と潜在顧客では得ることができる情報がおよび企業の行動が大きく異なるということがある．すでに自社サイトで一定以上の購入を行っている既存顧客 ($z = 1$) についてはメールアドレスがわかっており，かつ購買履歴等の情報からその顧客にカスタマイズされたメールによるクーポンの送付など様々な販促手段が取れる．一方，自社サイトでは購入していない（したがってメールを登録していない，購買履歴がわからないので効果的なプロモーションを行うに足りる情報がない）競合他社（ここでは B 社) の顧客に

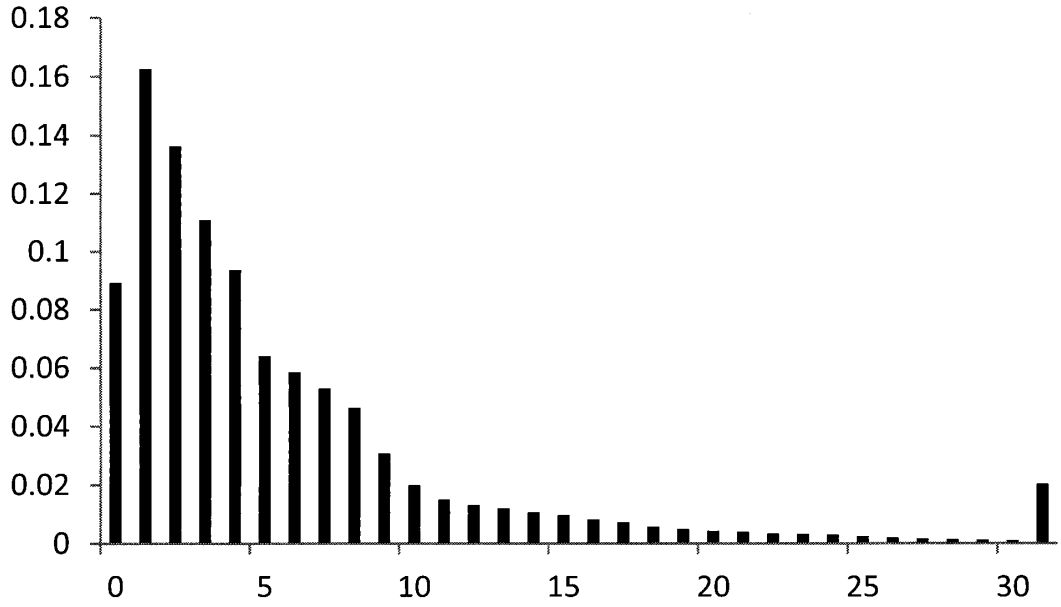

図 2 日別に集計した訪問間隔のヒストグラム.

対してこのような販促活動は行えないので，かれら潜在顧客が自社サイトに訪問し，さらには自社サイトにおいて商品を購入することを促すような Web 上での広告販促戦略を策定することが重要である．従って Web 上での広告販促戦略については既存顧客 ($z = 1$) ではなく潜在顧客 ($z = 0$) の反応が最大になるように行われる方が効果的である可能性が高い．そこで今回は時間共変量である経路情報がどのように自社サイトへの訪問間隔や購買確率に影響を与えているかについて，既存顧客のデータを用いて潜在顧客での結果を推定することを目的とする．

まず今回のデータで A 社にとっての既存顧客 ($z = 1$) とみなせたのは 4483 人であり，一方潜在顧客 ($z = 0$) と見なせたのは 1397 名であった．また，既存顧客の訪問間隔の平均は 8.3 日であった．図 2 に日単位で集計した訪問間隔のヒストグラムを記載している．但し 31 日以降はまとめて 31 日目の部分にその割合を記載している．この図からも，単純なガンベル分布は当てはまらず，消費者異質性を表現する変量効果をモデルに組み込む必要があることが示唆される．

ここでは以下の 4 つの方法で得られた推定値の比較を行った．

(1) 選択バイアスの修正を式 (2.4) によって行った，完全指数型ラプラス近似による重み付き一般化 EM アルゴリズムによる推定

(2) 選択バイアスの修正を行わない完全指数型ラプラス近似による一般化 EM アルゴリズムによる推定

(3) 変量効果を導入せずに，訪問間隔と購買有無を別々に最尤推定

(4) 実際の潜在顧客から得られているデータを利用して行った完全指数型ラプラス近似による一般化 EM アルゴリズムによる推定

4 つの方法による各パラメータの推定値と標準誤差を表 2 に示した．ここで，(4) は実際の潜在顧客から得られた推定値であるため，これを本来知りたい推測の対象であると考えるのが自然である．そこで (4) との推定値の差の二乗の（(3) では変量効果を含んでいないので，変量効果に関連する α_f と ϕ を除いた）全パラメータについての和を二乗誤差和として計算したものも表 2 に記載した．二乗誤差和は (1)，(2)，(3) の順番に大きくなることがわかることから，選択バイアスと個人異質性を同時に考慮した (1) が (4) の結果により近いことがわかる．

またこの表においてアスタリスクは係数がゼロであるとする検定において 5%有意であるものを示している．(1) から (3) と (4) ではサンプルサイズが等しくないので検定の結果だけからは一概に比較はできないが，(2) および (3) の方法で有意になっている係数のいくつか（例えば α_w での Yahoo!など）で (4) で得られた推定値とは正負の符号が逆転しており，潜在顧客について「どのような経路に対して広告を出稿すると潜在顧客により短いインターバルで自社サイトに訪問してもらえるか」または「自社サイトで購入してもらえるか」を理解する際に既存顧客のデータだけを選択バイアスを考慮しないで解析した結果を用いると誤った広告販促戦略を立案してしまう可能性が大きいことが示唆される．具体的には A 社の潜在顧客について訪問間隔を短縮させ，かつ各訪問時ごとに A 社サイトで購入する確率が高まるような経路としては (1) と (4) では google が挙げられるが，(2) と (3) では Yahoo!がそのような関係にある．これは Yahoo!は A 社の既存顧客にとっては有用な経路であることを示しており，既存顧客の維持ならば Yahoo!，潜在顧客の獲得ならば google と，目的によって異なる広告出向先を使い分ける必要があることを示唆している．また，得られたパラメータの推定値を利用して，特定の訪問経路からの流入を増やすような Web 広告を出稿することで，仮にその訪問経路からの流入が 10% 上昇した場合に，訪問間隔や購入確率がどの程度増えるか，といったシミュレーションを実施すれば，広告出稿に対する費用対効果，ROI (Return on investment) を算出することも可能となる．

さてここで，今回は本来解析の対象である潜在顧客についての情報があるため，(4) の結果との類似度から推定法の良さを議論できる．しかし一般に選択バイアスの推定については共変量の値によって各調査対象者がどちらのグループに分類されるかがわかるとする「ランダムな欠測」の仮定の妥当性について検証する必要がある．まずは既存顧客と潜在顧客が個人共変量でどの程度異なるかであるが，今回調整に利用した共変量はすべて既存顧客と潜在顧客で有意に差のある変数である．例えば年代と職業区分の 2 群での違いを図 3 に記載した．

表 2　Web ログデータでの 4 つの解析法での推定値の比較.

		(1) ラプラス近似 選択バイアス修正		(2) ラプラス近似 バイアス修正なし		(3) 変量効果なし		(4) 実データでの 潜在顧客	
		推定値	標準誤差	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差
σ		1.387*	0.01039	1.389*	0.00723	1.476*	0.00695	1.364*	0.00971
β_0		−0.229*	0.00873	−0.234*	0.00721	−0.261*	0.00583	−0.233*	0.01028
β_w	時間帯 12–18	0.165*	0.03624	0.084*	0.02140	0.053*	0.01889	0.178*	0.04865
	時間帯 18–22	−0.089*	0.04009	−0.102*	0.02509	−0.179*	0.02489	−0.092*	0.04803
	時間帯 22–6	−0.137*	0.04093	−0.088*	0.03305	−0.050	0.03107	−0.152*	0.04848
	Yahoo!	−0.028	0.04510	−0.141*	0.03984	−0.111*	0.03791	−0.049	0.06824
	google	−0.299*	0.07304	−0.093	0.06985	−0.101	0.06208	−0.367*	0.08607
	その他検索エンジン	0.109	0.10028	0.034	0.07709	0.060	0.08094	0.130	0.11276
	blog	−0.038	0.06904	−0.110	0.07090	−0.381*	0.05033	−0.093	0.06687
	SNS	−0.091	0.06002	0.009	0.05809	0.109*	0.05034	−0.088	0.07328
	メール	0.093	0.08709	0.034	0.04095	0.051	0.03097	0.121	0.08797
	情報サイト	−0.039	0.05001	−0.019	0.04098	−0.054	0.03887	−0.071	0.07321
α_0		−3.869*	0.29503	−3.739*	0.15600	−3.593*	0.13978	−4.361*	0.40943
α_f		−0.321*	0.11010	−0.208*	0.06982	—	—	−0.409*	0.14132
α_w	時間帯 12–18	−0.110	0.07093	0.050	0.04499	0.069*	0.04109	−0.108	0.09834
	時間帯 18–22	0.051	0.04094	0.065*	0.02499	0.041*	0.02098	0.039	0.07169
	時間帯 22–6	0.210*	0.05097	0.049	0.06097	0.110*	0.05570	0.261*	0.06095
	Yahoo!	−0.019	0.03069	0.109*	0.01610	0.098*	0.01609	−0.048	0.02950
	google	0.081*	0.02483	0.051	0.03329	0.053	0.03192	0.083*	0.03082
	その他検索エンジン	−0.060	0.06094	−0.107*	0.04508	−0.050	0.04604	−0.101	0.09989
	blog	0.101*	0.05098	−0.030	0.03097	−0.060*	0.02806	0.101	0.08426
	SNS	0.179*	0.06093	0.061	0.05093	0.110*	0.04330	0.210*	0.08272
	メール	−0.041	0.07083	0.149	0.03780	0.101*	0.03512	−0.098*	0.04133
	情報サイト	0.080	0.06095	0.008	0.04397	0.076*	0.03799	−0.051	0.07075
ϕ		0.832*	0.05938	0.790*	0.03921	—	—	0.889*	0.08983
(4) との推定値の 2 乗誤差			0.28059		0.71366		1.01456		—

アスタリスクは 5%有意であることを示す.

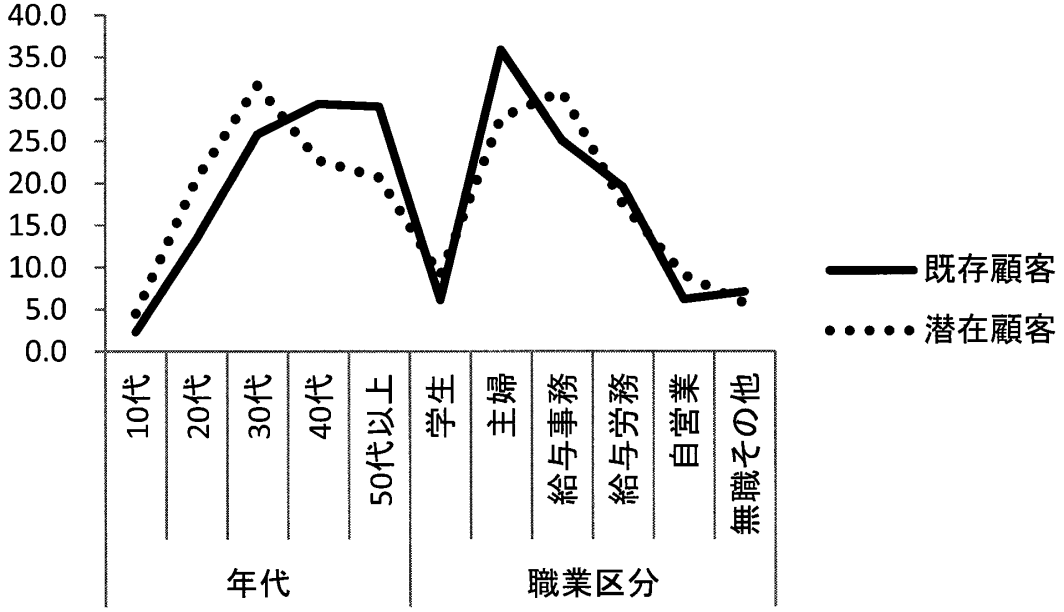

図 3　既存顧客と潜在顧客での年代と職業区分の違い．*縦軸はパーセンテージ．

今回のような傾向スコアを用いた共変量調整の前提条件（この場合はランダムな欠測と同じ）は直接検証することはできないが，間接的な確認法として，共変量によって z の説明が十分に行われるかを調べることが通常である (星野 (2009))．そこでロジスティック回帰分析モデルの説明力の指標としてよく利用される c 統計量を計算したところ，0.7331 とある程度高い値であること，また傾向スコアを用いて重みづけ推定をすることで，各共変量の分布の z による群間差が無視できることから，仮定が間接的に確認されたと考えてよいと思われる．

5. まとめと議論

本論文では大規模データの解析であまり議論されてこなかった個人の異質性と選択バイアスの問題の重要性を指摘し，訪問間隔と購買確率の同時分析においてこれらの問題を変量効果と重み付き推定という形で解決するためのモデリングとアルゴリズムを提案し，大規模なパネル Web ログデータに適用した．シミュレーションから，提案された手法が大規模データにおいて MCMC などと比べても十分な推定精度を有しながらも，少ない計算量で解析が可能であることを示した．また解析例から，異質性と選択バイアスを考慮しない単純な解析では推定に大きなバイアスが生じえることを示した．

実はこれまでも選択バイアスについては，機械学習や人工知能分野でも転移学習 (transfer learning) の一分野として知られており (神嶌 (2010))，特に連続変数を従属変数とする回帰モデルの文脈で独立変数の分布を用いて重み付きを行う共変量シフト (Shimodaira (2000)，Sugiyama *et al.* (2007)) の研究が行われてきた．共変量シフトの重みは傾向スコア (Rosenbaum and Rubin (1983)) を用いた重みとしても理解できるが，実は仮定してい

る回帰モデルが正しければランダムな欠測であることから，重みがなくても一致推定が可能である．したがって共変量シフトは回帰モデルを正しく特定できないことを前提として，予測を行う場合の予測誤差を最小にするために重み付けを行う方法であるといえる．これに対して，統計学で傾向スコアを用いた選択バイアスの除去が盛んに利用されているのは，従属変数と独立変数の回帰関係を推論する際に，単純にバイアスの補正に利用したいだけの共変量を関心のある独立変数とは区別し，共変量と従属変数との回帰関係を仮定せずにロバストな解析を行うことに目的がある．本論文の解析例のように，従属変数の予測に利用できる変数が限定されている場合には，第2節で述べたような理由からモデル仮定の少ない今回のような方法を利用すべきだと考えられる．

また本研究ではzが観測される共変量にのみ依存する「ランダムな欠測」を仮定したが，zが訪問間隔の個人差を表現する変量効果fに依存するという仮定の緩和を行うことは容易であり，欠測データモデルとしては共有パラメータモデル (Follman and Wu (1995)，星野 (2009)) の一種となるが，推定法はここで提案したものを一部修正するだけで実行可能である．

本研究では詳しく議論することができなかったが，共変量と平均パラメータの回帰関係をノンパラメトリックなものとすることで，よりロバストな推定が可能になると期待される．しかし本モデルでは変量効果も存在しているため推定法や多重積分の計算量の問題が生じる．近年関心のない部分にはノンパラメトリックなモデルを想定し，関心があり理論や先行研究の結果から特定できる部分にはパラメトリックなモデルを仮定する「セミパラメトリック」なモデリングが利用されつつあるが，MCMC の計算量の問題さえクリアできれば，セミパラメトリックベイズモデル (例えば Hjort *et al.* (2010)) を用いた選択バイアス調整法 (Hoshino, in press) など，より仮定の少ないロバストかつ推定精度の高い方法を利用することが可能になると考えられる．今後計算機速度のさらなる向上により，今回取り上げたような大規模なデータでのセミパラメトリックベイズモデルへの MCMC の実行も可能になると考えられる．

本研究の解析例の限界としては，URL ごとに実際にどのような内容のページから遷移して A 社のサイトに流入してきたかを調べることができなかったこと，および解析例の潜在顧客の定義では調査に参加した最初の半年以前に A 社ですでに商品購入を行っている対象者，つまり既存顧客がいる可能性を排除できないため，方法 (4) で行われた解析は「本来の潜在顧客を対象にした解析」とは異なる可能性があるという2点が挙げられる．後者については潜在顧客をどのように定義しても同様の問題が生じえるという点で難しいが，前者については URL ごとに内容を特定する形の研究を行うことで対応できる．

謝辞

ビデオリサーチインタラクティブ (株) には本データの使用許可を頂きましたことを感謝致します．また，本論文を執筆する機会を与えて頂きました水田正弘先生，本論文の改稿に当たって有用なコメントを頂いた匿名の査読者に御礼申し上げます．本研究は科学技術振興機構さきがけ「マルチソースデータ高度利用のための統計的データ融合」および科学研究費補助金若手 (A)23680026 の助成を受けたものです．

A. 補遺

変量効果が観測された元でのパラメータ θ のスコア関数は，ガンブル分布を仮定した位置スケールモデルやロジスティック回帰，正規分布モデルのそれらに対応するため省略する．通常の最尤推定では計算しないが完全指数型ラプラス近似に必要な量として（変量効果が得られている場合の）同時分布の変量効果による 1，2，3 次導関数を下記に記載する．

$$-\frac{\partial}{\partial f_i}\log g(f_i,\boldsymbol{y}_i,\boldsymbol{w}_i,z_i|\boldsymbol{\theta}) = -\frac{f_i}{\phi}-\sum_{j=1}^{J_i}\left[\frac{1}{\sigma}+\frac{1}{\sigma}\exp\left[\frac{y_{ij}^{LD}-\boldsymbol{\beta}^t\boldsymbol{w}_{ij}^*}{\sigma}\right]-\alpha_f\left(y_{ij}^B-p_{ij}\right)\right] \tag{A.1}$$

$$\boldsymbol{\Sigma}_i = -\frac{1}{\phi}-\sum_{j=1}^{J_i}\left[-\frac{1}{\sigma^2}\exp\left[\frac{y_{ij}^{LD}-\boldsymbol{\beta}^t\boldsymbol{w}_{ij}^*}{\sigma}\right]+\alpha_f^2 p_{ij}(1-p_{ij})\right] \tag{A.2}$$

$$\frac{\partial}{\partial f_i}\boldsymbol{\Sigma}_i = -\sum_{j=1}^{J_i}\left[\frac{1}{\sigma^3}\exp\left[\frac{y_{ij}^{LD}-\boldsymbol{\beta}^t\boldsymbol{w}_{ij}^*}{\sigma}\right]+\alpha_f^3 p_{ij}(1-p_{ij})(1-2p_{ij})\right] \tag{A.3}$$

但し $\boldsymbol{w}_{ij}^* = (1, f_i, \boldsymbol{w}_{ij}^t)^t$，$p_{ij} = p(y_{ij}^B = 1|f_i,\boldsymbol{x}_i,\boldsymbol{w}_{ij})$ である．実際にはこれらをさらにパラメータについて微分したもの $\partial\boldsymbol{\Sigma}/\partial\theta$ なども計算には必要となるが，たとえば $\boldsymbol{\beta}$ については

$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\Sigma}_i = -\sum_{j=1}^{J_i}\frac{1}{\sigma^3}\boldsymbol{w}_{ij}^*\exp\left[\frac{y_{ij}^{LD}-\boldsymbol{\beta}^t\boldsymbol{w}_{ij}^*}{\sigma}\right] \tag{A.4}$$

α については p_{ij} の導関数が $\boldsymbol{\alpha}p_{ij}(1-p_{ij})$ を利用することから通常のロジスティック回帰分析モデルでの結果を応用すればよいだけであるため，ここでは省略する．

参 考 文 献

Angrist, J. and Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, London.

Bianconcini, S. and Cagnone, S. (2012). Estimation of generalized linear latent variable models via fully exponential Laplace approximation, *J. Multivar. Anal.*, **112**, 183–193.

Bijwaard, G. E., Franses, P. H. and Paap, R. (2006). Modeling purchases as repeated events, *J. Bus. Econ. Stat.*, **24**, 487–502.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice, *J. Am. Stat. Assoc.*, **105**, 324–335.

Follman, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data, *Biometrics*, **51**, 151–168.

Fong, Y., Rue, H. and Wakefield, J. (2010). Bayesian inference for generalized linear mixed model, *Biostatistics*, **11**, 397–412.

Hjort, N. L., Holmes, C., Müller, O. and Walker, S. G. (2010). *Bayesian Nonparametrics*, Cambridge University Press, Cambridge.

星野崇宏 (2005).「欠測群の周辺分布の母数に対する傾向スコアを用いた重み付き M 推定量の提案と介入効果研究への応用」『行動計量学』32 巻 2 号, 121–132.

星野崇宏 (2009).「調査観察データの統計科学：因果推論・選択バイアス・データ融合」岩波書店.

Hoshino, T. (in press). Semiparametric Bayesian estimation for marginal parametric potential outcome modeling: Application to causal inference, *J. Am. Stat. Assoc.*

Hoshino, T., Kurata, H. and Shigemasu, K. (2006). A propensity score adjustment for multiple group structural equation modeling, *Psychometrika*, **71**, 691–712.

岩爪道昭 (2013).「特集「ビッグデータと AI」にあたって」『人工知能学会誌』28 巻 1 号.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. (1999). An introduction to variational methods for graphical models, *Mach. Learn.*, **37**, 183–233.

神嶌敏弘 (2010).「転移学習」『人工知能学会誌』25 巻 4 号.

Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.*, Springer, New York.

小山慎介 (2012).「状態空間法による神経スパイク発火率の推定」『統計数理』60 巻 1 号, 173–188.

Pan, Q. and Schaubel, D. E. (2009). Evaluating bias correction in weighted proportional hazard regression, *Lifetime Data Analysis*, **15**, 120–146.

Rizopoulos, D., Verbeke, G. and Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modeling of survival and longitudinal data, *J. R. Stat. Soc., Ser. B*, **71**, 637–654.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations (with discussion), *J. R. Stat. Soc., Ser. B*, **71**, 319–392.

Seethraman, P. B. and Chintagunta, P. K. (2003). The proportional hazard model for purchase timing: A comparison of alternative specifications, *J. Bus. Econ. Stat.*, **21**, 368–382.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function, *J. Stat. Plan. Inf.*, **90**, 227–244.

総務省 (2012).「情報通信白書」.

Sugiyama, M., Krauledat, M. and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation, *J. Mach. Learn. Res.*, **8**, 985–1005.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *J. Am. Stat. Assoc.*, **81**, 82–86.

Tierney, L., Kass, R. and Kadanae, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *J. Am. Stat. Assoc.*, **84**, 710–716.

Vaida, F. and Xu, R. (2000). Proportional hazards models with random effects, *Stat. Med.*, **19**, 3309–3324.

Wang, B. and Titterington, D. M. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models, *Neural Process. Lett.*, **20**, 151–170.

Wooldridge, J. M. (2007) Inverse probability weighted M-estimation for general missing data problems, *J. Econom.*, **141**, 1281–1301.