MMRC
DISCUSSION PAPER SERIES

MMRC-J-183

消費者行動理論にもとづいた 個人レベルの RF 分析： 階層ベイズによる Pareto/NBD モデルの改良

東京大学　大学院経済学研究科・経済学部
阿部　誠

2007 年 11 月

東京大学21世紀COE〔センター・オブ・エクセレンス〕
ものづくり経営研究センター

消費者行動理論にもとづいた個人レベルの RF 分析：階層ベイズによる Pareto/NBD モデルの改良

東京大学　大学院経済学研究科・経済学部

阿部　誠

2007 年 11 月

An Individual Level RF Analysis based on Consumer Behavior Theory:

A Hierarchical Bayes Framework on the Pareto/NBD Model

ABSTRACT

This research extends a Pareto/NBD model of customer-base analysis using a hierarchical Bayesian (HB) framework to suit today's customized marketing. The proposed HB model presumes three tried and tested assumptions of Pareto/NBD models: (1) a Poisson purchase process, (2) a memoryless dropout process (i.e., constant hazard rate), and (3) heterogeneity across customers, while relaxing the independence assumption of the purchase and dropout rates and incorporating customer characteristics as covariates. The model also provides useful output for CRM, such as a customer-specific lifetime and survival rate, as by-products of the MCMC estimation.

Using two different types of databases --- music CD for e-commerce and FSP data for a department store, the HB model is compared against the benchmark Pareto/NBD model. The study demonstrates that recency-frequency data, in conjunction with customer behavior and characteristics, can provide important insights into direct marketing issues, such as the demographic profile of best customers and whether long-life customers spend more.

Key words: CRM, One-to-One Marketing, Bayesian method, MCMC, data augmentation

要約

RFM 分析で使われるリーセンシー(直近の購買からの経過時間)とフリークエンシー(購買頻度)のデータから、一般的な消費者行動の仮定に基づいて、ある時点での顧客の生存確率を推定する。既存の経験ベイズに基づいた Pareto/NBD モデルを階層ベイズの枠組みに改良し、購買率と離脱率を表すパラメータに共変量を組み込むことによって、マーケティングに有益な知見が得られる。実証研究として、日米２種類の顧客購買データを使い、このモデルを評価する。

キーワード：マーケティング、階層ベイズ、MCMC 法、データ補完

１．はじめに

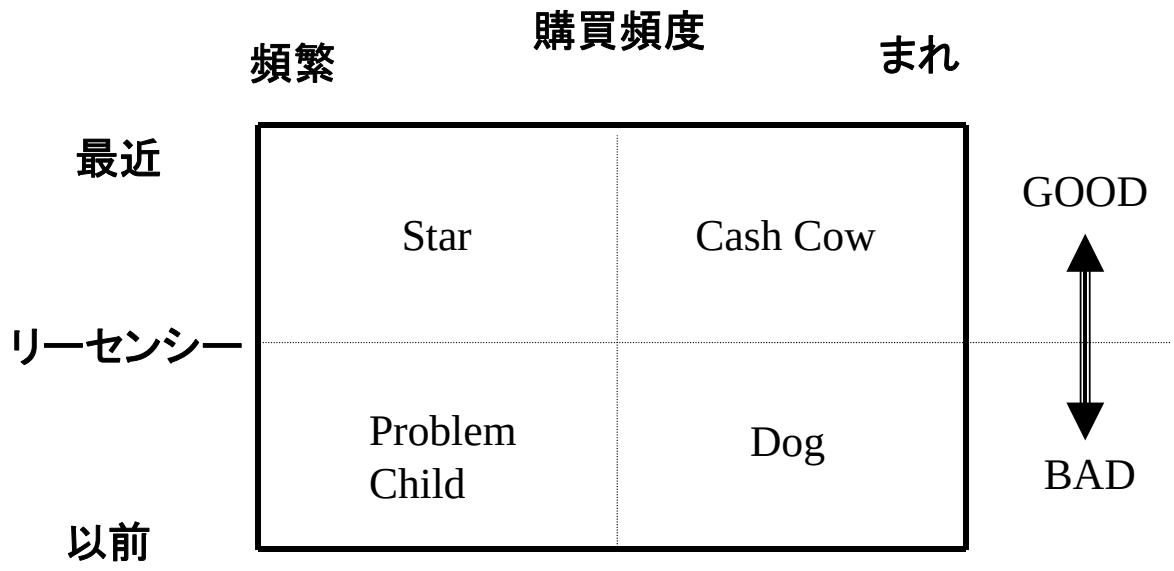
マーケティングで重要な概念である顧客の生涯価値(customer lifetime value)を計算するには、顧客の離脱率を把握する必要がある (Blattberg and Deighton 1996)。しかし離脱する顧客は単に購買を止めるだけで、特に年会費などの支払い義務がなければ、わざわざ離脱を申告することはしない。通常このような場合、企業は独自の経験則に基づいて、例えば顧客が３ヶ月購買しなければ離脱したと判断したりする。実務家の間でよく使われる RFM(recency, frequency, monetary-value)分析では[1]、『リーセンシー=3 ヶ月』のようなアドホックで一律なルールが基本になっているが、ここには２つの問題がある。第１に、このルールが主観的なことである。なぜ２ヶ月や４ヶ月でなく、3 ヶ月なのだろうか？ ２つ目の問題は、マーケティングの基本的概念である顧客の異質性を無視していることである。同じ 3 ヶ月のリーセンシーでも、購買間隔が長い顧客は離脱の心配が少ないが、購買間隔が短い顧客は離脱している可能性が高いであろう。つまり離脱率の推測には顧客の異質性に配慮する必要がある。

図 1 は既存の RF 分析を、企業にとっての顧客の魅力度(貢献度)の観点から、ボストン・コンサルティング・グループ社の戦略ポートフォリオ・マトリクスとして描いたものである。

[1] RFM 分析は、購買データからリーセンシー（直近の購買からどのくらいの時間が経ったか？)、フリークエンシー（購買頻度)、１回の平均購買金額の３つの基準で顧客をセグメント分けする手法のことである。

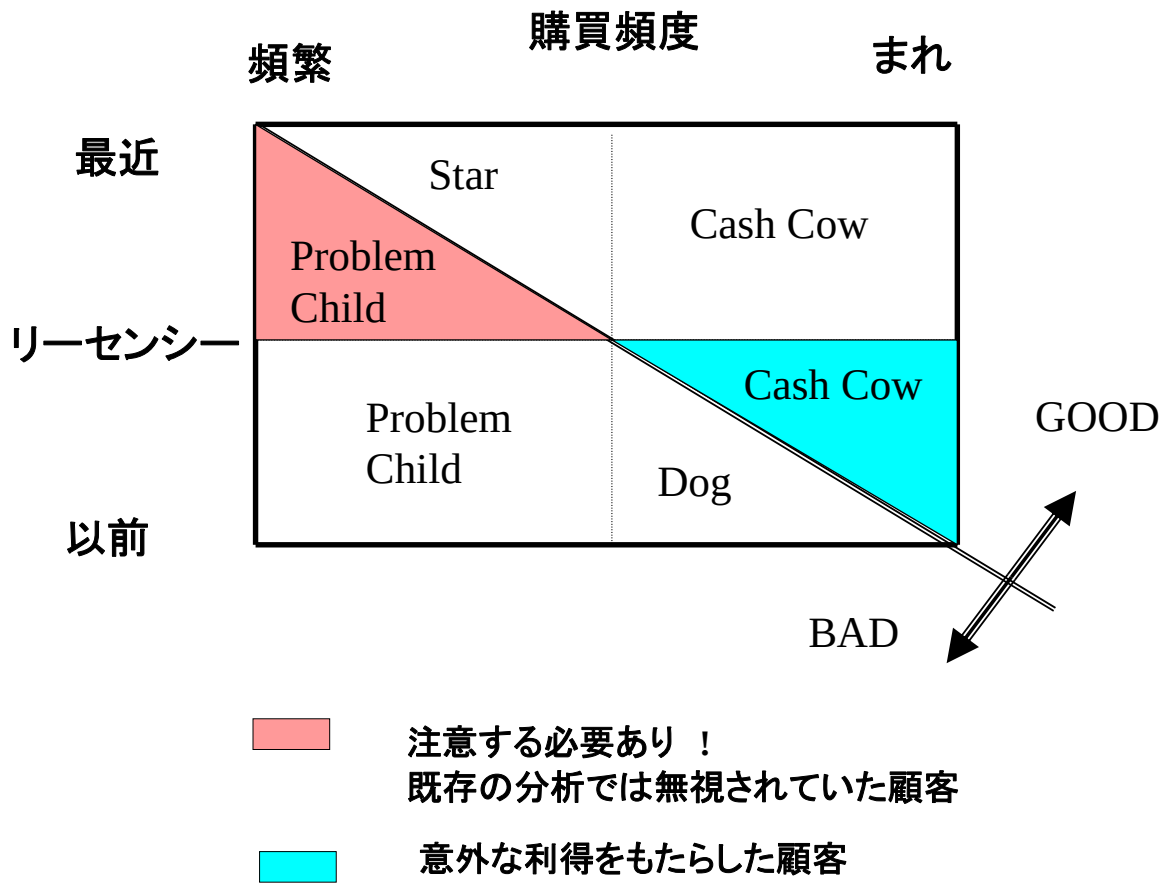
まず、あるリーセンシーの閾値(例えば 3 ヶ月)によって生存顧客(Star と Cash Cow)と離脱顧客(Problem Child と Dog)を区別し、さらに生存顧客は購買頻度(フリークエンシー)に基づいて優良顧客(Star)と通常顧客(Cash Cow)に識別される。この方法ではリーセンシーとフリークエンシーが独立に考慮されている。

図１．既存の RF 分析



これに対して、リーセンシーとフリークエンシーの両方を同時に考慮すると、図 2 のように解釈が変わってくる。既存の分析では Star と区分されていた購買頻度が高くかつ最近購買した顧客(左上の区分)の中でも、購買頻度に見合うようなリーセンシーを示さない顧客は Problem Child として早急な対応が求められる。また、既存の分析では Dog と区分されていた購買頻度が低くかつ最近購買していない顧客(右下の区分)の中でも、その人の購買頻度に対して比較的最近購買していれば、企業に予想外の売上げをもたらしたとして Cash Cow に分類される。既存の RF 分析では、これら 3 角形の 2 セグメントの顧客に対して企業は特別な配慮をしなかったため、利得機会を失っていた。

図 2. 本研究で提案している RF 分析



この問題に関して、Schmittlein, Morrison and Colombo (1987)（この先からは略して SMC と呼ぶ）は一般的な消費者行動の仮定に基づいた Pareto/NBD モデルを提案し、リーセンシーとフリークエンシーの関係を導くことによって、ある時点における生存確率を顧客別に求めた。このモデルでは、消費者行動として(1)ポアソン・プロセスにしたがう購買行動（購買率を表すポアソン・パラメータ＝λ）と(2)離脱までの時間が指数分布にしたがう無記憶的な離脱行動（離脱率を表す指数パラメータ＝μが一定）が仮定されている。さらに顧客の異質性を考慮して、λとμがそれぞれ独立なガンマ分布にしたがう混合分布モデルとなっている。その後、この研究の流れで数本の論文が発表されたが(Fader, Hardie and Lee 2005a, 2005b, Reinartz and Kumar 2000, 2003, Schmittlen and Peterson 1994)、近年の情報技術の進歩により可能となった個人別対応のマーケティング（ワン・トゥー・ワン・マーケティング）はこの研究分野の

重要性を飛躍的に高めた。

本研究は、このマイクロ・マーケティングにより相応しいように、SMC とその一連の研究の流れである行動理論にもとづいた RF 分析の概念を発展させたものである。SMC が採用した消費者行動の仮定はそのまま残し、階層ベイズのフレームワークに基づいて個人別のパラメータを想定することによって顧客の異質性をモデル化する(Abe 2006,2008)。具体的には、(1)顧客の異質性を表す混合分布をシミュレーション手法に委ね、(2)顧客の生存時間や離脱の有無を示す観測不能な指標をモデルの潜在変数として取り込む。顧客の異質性を積分によって解析的に総計する必要がないため、以下のメリットが生じる。

１．概念の単純化

SMC が論文での一番重要な結果であると主張している顧客別生存確率の公式とその難解な導出(彼らの論文中の式(11)～(13)と付録)が不要になる。

２．パラメータ推定の単純化

Pareto/NBD モデルの混合分布パラメータの推定は複雑で、最尤法も含めていくつか提案されており、Schmittlein and Peterson (1994)でも詳細に検証されている。これらが不要になる。

３．計算の単純化

Pareto/NBD モデルでは、パラメータの推定と生存確率の計算に、ガウス・ハイパージオメトリックという非標準的な関数の数値を繰り返し使うが、通常のソフトには存在しないため、数値手法で近似的に求めなければならない。計算を簡便化するために、Fader, Hardie, and Lee (2005a)は Pareto/NBD モデルを近似する BG/NBD モデルを提案した。

４．モデルの柔軟性

この論文で提案するモデルは、Pareto/NBD モデルの仮定のひとつである購買率と離脱率パラメータの分布の独立性を必要としない。データがこの仮定を満たしていなければ、Pareto/NBD モデルのパラメータ推定にはバイアスがかかる可能性がある。本モデルは、2 つのパラメータが相関しているデータに対しても適用できることに加えて、下記の(6)で説明するように独立性の統計的仮説検定をも行うことができる。

５．潜在変数の個人別推定

購買率 λ と離脱率 μ は個人ごとに推定される。3.1.節で説明するが、経験ベイズのフレームワークに基づいた Pareto/NBD モデルでは、これらのパラメータを求めることは計算上非常に負荷が高い。個人 i の(λ_i, μ_i)の事後平均を散布図として描くことによって、Pareto/NBD モデルの独立性の仮定を評価することが可能である。Pareto/NBD モデルからは容易に得られないその他の潜在変数として、生存時間の期待値と一年後の顧客維持率がある。

６．正確な誤差の推定

本研究で用いられた MCMC 法によるベイズのフレームワークでは、漸近理論を使わずにパラメータを点推定ではなく事後分布として求めるため、統計的仮説検定のための誤差が正確に推定できる。第 4 節の実証分析では、log(λ)と log(μ)の相関係数の事後分布から、その独立性を統計的に検定する。

７．モデルの発展性

λ や μ が顧客の共変量の関数となる階層モデルの構築と推定が容易である。

(a) Schmittlein and Peterson (1994)は産業コードで顧客をセグメント分けして、セグメントごとに Pareto/NBD モデルを推定することにより、実務に有益な示唆が得られることを示した。本論文で提案するモデルでは、セグメント変数を階層的に組み込むことによって全てのセグメントを同時推定できるため、データの自由度を最大限有効に活用できる。また説明変数としてカテゴリー変数以外を組み込むことも可能である。

(b)　顧客特性が顧客の生存時間におよぼす影響を調査するために、Reinartz and Kumar (2003)は 2 ステップ・アプローチを提案した。まず Pareto/NBD モデルを使って RF データから顧客の生存時間を予測し、第２ステップではその生存時間を従属変数、顧客特性を説明変数とした比例ハザードモデルを構築するというものである。離脱率 μ を顧客特性の関数とした階層モデルでは、これを 1 ステップで分析できるため、統計的仮説検定を適用するための誤差も正しく推定される。

８．正式なベイズのパラダイム

SMC のアプローチは通称、経験ベイズと呼ばれ、データが尤度関数と事前分布の推定の両方に使われるため、パラメータの精度が過大に推定される傾向がある。サンプル数が多かったり、事前分布を他のデータから推定したりすれば問題は少ないが、ベイズのパラダイムでは経験ベイズは階層ベイズの近似と理解されている。(Gelman, Carlin, Stern and Rubin 1995, p.123)

この論文は以下の構成になっている。まず第 2 節で提案モデルを SMC の Pareto/NBD モデルと比較しながら説明した後、第 3 節で MCMC シミュレーション法によるモデルの推定方法を紹介する。第 4 節では米国の E コマースと日本のデパートにおける顧客購買記録データを使った実証分析を行い、ベンチマークである Pareto/NBD モデルとの比較を試みる。第 5 節では研究の結論とモデルの限界を述べる。

2. モデル

2.1. 消費者行動の仮定

［仮定 1］購買はポアソン・プロセスに従う

この仮定は、購買が過去に何時起きたかに関係なく、ランダムに発生することを意味する。Ehrenberg (1972, 1988) の研究以来、この無記憶的なゼロ次の購買プロセス（これに対して購買発生が過去 1 期の状態したがうのであれば 1 次のマルコフ・プロセスになる）は多くのデータでロバストなことが確認されている(Bass, Givon, Kalwani, et. al 1984)。ただしこの仮定は周期性のある購買には当てはまらないため、単一カテゴリーよりは多カテゴリーの購買行動に適用するべきである。

［仮定 2］顧客の生存時間は指数分布に従う

これは、離脱が過去の生存時間に関係なくランダムに起きるという無記憶性を意味する。この仮定の妥当性は、離脱が企業に対する飽き、競合企業への乗り換え、転居、死去などの様々な理由によって起きることと、一度購買が観測されるということは顧客の生存が確認されて離脱プロセスがリセットされる、という 2 点から支持される。

［仮定 3］顧客の異質性

購買頻度を表すポアソンのパラメータと生存時間を表す指数分布のパラメータは、顧客によって異なる。Pareto/NBD モデルでは 2 変量独立ガンマ分布が仮定されていたが、以下の理由から本研究では 2 変量対数正規分布を仮定する。

(a)経営上、有益な知見を得るために、本研究ではこれらのパラメータを顧客特性に関する共変量の関数とした階層ベイズモデルへの拡張を行う。その際、パラメータに多変量正規分布を仮定したベイズ回帰分析は多くのモデルで採択されており、推定が容易である。

(b) 多変量正規分布の共分散行列から、log(λ)と log(μ)の相関が推定できる。相関のある 2 変量ガンマ分布は標準的な分布でないため、推定が複雑である(Park and Fader 2004)。

(c) Pareto/NBD モデルを用いた過去の全ての研究では、離脱率 μ のガンマ分布の形状母数が全て 1 以下と推定されているが、その場合、生存時間の期待値は無限大を意味する。顧客はいずれ離脱することを考慮すると、これは直感に反するため、対数正規分布の方がより適切であろう。

2.2. 数学的表記

図 3. 購買履歴データの表記

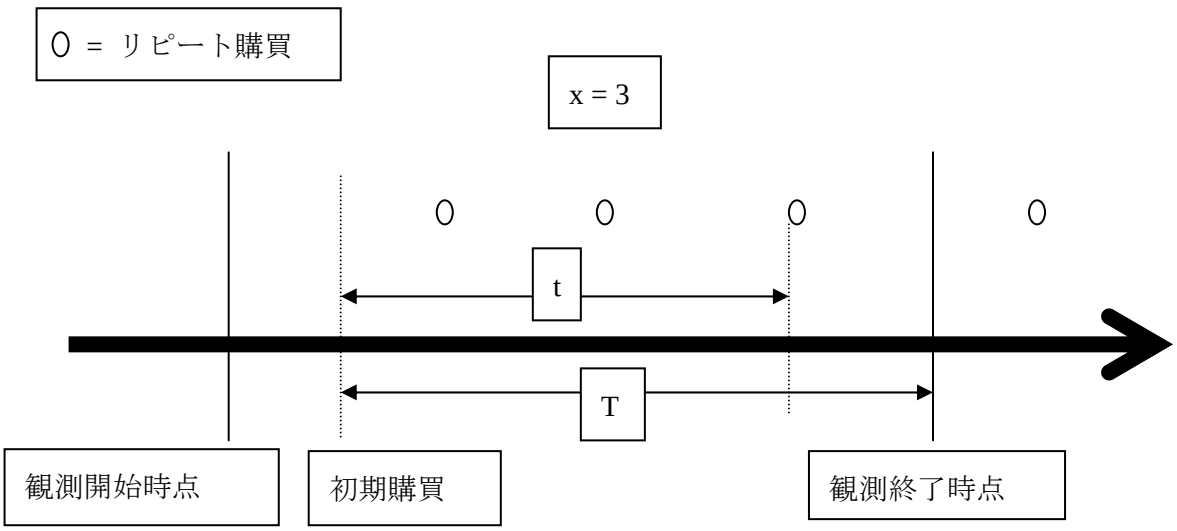


図３は SMC が用いた RF データ（x, t, T）の表記であり、本論文でもそれにしたがう。最初の購買は時間０に発生し、その顧客の購買履歴は時間 T まで記録される。x は期間(0, T]に発生したリピート購買の回数(初回の購買を含まない)を表し、最後のリピート購買(x 回目)は時点 t に起きる。したがって、リーセンシーは T-t と定義できる。τは顧客の生存時間を表すが、データがＴ以降打ち切られているため、τは観測されない。ここでの数学的表記を用いると、前節のモデルの仮定は以下のように表される。

(A1)
$$P[x|\lambda]=\begin{cases}\dfrac{(\lambda T)^x}{x!}e^{-\lambda T} & \text{if } \tau > T \\ \dfrac{(\lambda \tau)^x}{x!}e^{-\lambda \tau} & \text{if } \tau \le T\end{cases} \qquad x = 0,1,2,..$$

(A2)
$$f(\tau)=\mu e^{-\mu\tau} \qquad \tau \ge 0$$

(A3)
$$\begin{bmatrix}\log(\lambda)\\ \log(\mu)\end{bmatrix} \sim MVN\left(\theta_0=\begin{bmatrix}\theta_\lambda\\ \theta_\mu\end{bmatrix}, \Gamma_0=\begin{bmatrix}\sigma_\lambda^2 & \sigma_{\lambda\mu}\\ \sigma_{\mu\lambda} & \sigma_\mu^2\end{bmatrix}\right)$$

ここでは *MVN* は多変量正規分布を表す。

λはポアソン・プロセスのパラメータで、E[x]=λT なので、λは「単位期間あたりの購買頻度」と解釈できる。μ は指数分布のパラメータで、E[τ]=1/μ なので、μは大雑把に「離脱率」を表すと解釈できる。これらの仮定から、経営上有益な顧客レベルの統計値、たとえば生存時間の期待値、１年後の維持率、観測終了時点での生存確率、などが付録に導かれている。

2.3. 説明変数を導入するための階層モデルへの拡張

購買頻度パラメータλと離脱率パラメータμを顧客特性の関数としてモデル化することによって、購買頻度の多い顧客やロイヤルティーの高い（つまり生存時間の長い）顧客のプロファイルに関する知見が得られる。顧客特性が人口統計的変数であれば、まだ購買記録のない新規顧客を獲得する場合のターゲットに関する情報が得られる。一番シンプルなモデルは、λとμの対数が説明変数の線形となる以下の回帰モデルである。

(A3') $$\begin{bmatrix}\log(\lambda_i)\\ \log(\mu_i)\end{bmatrix} \equiv \theta_i = \beta' d_i + e \qquad \text{where } e \sim MVN(0, \Gamma_0)$$

d_i は K×1 ベクトルで、顧客 i のK個の特性を表す。βは K×2 のパラメータ・ベクトル、e は2×1の誤差項で平均0、共分散行列 Γ_0 の多変量正規分布にしたがう。この回帰モデルは、前節の θ_0 を $\beta' d_i$ に置き換えたものである。特別なケースとして d_i が 1 のスカラーの場合、切片のみで説明変数を含まないため、(A3)と同等になる。

３．モデルの推定

3.1. 潜在変数の導入

経験ベイズの枠組みに基づいた Pareto/NBD モデルでは個人レベルのλとμを推定するのが困難であるという理由を考察することによって、適切な推定のアプローチが導かれる。Pareto/NBD モデルでは、

事前分布： $\lambda_i \sim \text{Gamma}(r, \alpha)$, $\mu_i \sim \text{Gamma}(s, \beta)$

顧客が T_i の時点で生存している場合、それぞれの事後分布は、

$$\lambda_i | \text{data}_i \sim \text{Gamma}(r+x_i, \alpha+T_i), \qquad \mu_i | \text{data}_i \sim \text{Gamma}(s, \beta+T_i)$$

顧客が T_i より前の yi の時点で離脱している場合、それぞれの事後分布は、

$$\lambda_i | \text{data}_i \sim \text{Gamma}(r+x_i, \alpha+y_i), \qquad \mu_i | \text{data}_i \sim \text{Gamma}(s+1, \beta+y_i)$$

となる。したがって、データからは観測されない情報(顧客がT_iの時点で生存あるいは離脱しているのかと、離脱している場合はその時間y_i)が分からない限り、上記の関係を使ってλ_iとμ_iをベイズ的に更新することは出来ない。今回のモデルでは、これらの観測されない情報を潜在変数として導入する。表記を単純にするために、この先の説明では、顧客を表す添え字iを省く。潜在変数zを、顧客が観測終了時点Ｔで生存していれば１、そうでなければ０（つまり離脱）と定義する。さらにz=0 の場合、もう一つの潜在変数として離脱時刻y(<T)を定義する。もし、zとyが既知であれば、RFデータ(x, t, T)の尤度関数はx>0 の場合、以下

のような単純な形になる[2]。

z=1 のケース (顧客はＴの時点で生存)

$$P\left(x\text{回目の購買が時点}t\text{で発生＆Tまで生存＆期間}[t,T]\text{に購買が発生しない}\right)$$
$$=\frac{\lambda^x t^{x-1}}{\Gamma(x)}e^{-\lambda t}\times e^{-\mu T}\times e^{-\lambda(T-t)}$$
$$=\frac{\lambda^x t^{x-1}}{\Gamma(x)}e^{-(\lambda+\mu)T}$$

z=0 のケース (顧客は y<Ｔの時点で離脱)

$$P\left(x\text{回目の購買が時点}t\text{で発生＆期間}[t,\mathrm{y}]\text{に購買が発生しない＆}y\text{の時点で離脱}(y<T)\right)$$
$$=\frac{\lambda^x t^{x-1}}{\Gamma(x)}e^{-\lambda t}\times e^{-\lambda(y-t)}\times \mu e^{-\mu y}$$
$$=\frac{\lambda^x t^{x-1}}{\Gamma(x)}\mu e^{-(\lambda+\mu)y}\qquad (\mathrm{t}\le y\le T)$$

２つのケースを統合した、よりコンパクトな尤度関数は式(１)のように表現される。

(1)　　$$L(x,t,T\mid\lambda,\mu,z,y)=\frac{\lambda^x t^{x-1}}{\Gamma(x)}\mu^{1-z}e^{-(\lambda+\mu)\{zT+(1-z)y\}}$$

ただしデータからはｚもｙも観測されないので、これらを欠損データと見なして、データ補完を使った MCMC 法(Tanner and Wong 1987)によって推定する。ｚをシミュレーションで発生させるための確率、つまり顧客がＴの時点で生存している確率は式(２)になる(導出は付録を参照)。

[2] x=0 は、リピート購買が無いことを意味するため t=0 になる。したがって $\Gamma(x=0)$と t^{x-1} は定義されない。この場合の尤度関数は$z=0$の場合 $e^{-(\lambda+\mu)T}$，$z=1$の場合 $\mu e^{-(\lambda+\mu)y}$ となる。したがって、(1)式は$L(x,t,T\mid\lambda,\mu,z,y)=\mu^{1-z}e^{-(\lambda+\mu)\{zT+(1-z)y\}}$と表される。

(2) $$P[\tau > T \mid \lambda, \mu, T, t] = P[z = 1 \mid \lambda, \mu, T, t] = \frac{1}{1 + \frac{\mu}{\lambda + \mu}\left[e^{(\lambda+\mu)(T-t)} - 1\right]}.$$

3.2. データ補完による推定

このモデルではパラメータλとμは顧客別に推定されるため、顧客を表す添え字 i（i=1, …, I）を戻す。顧客別パラメータ θ_i = [log(λ_i), log(μ_i)]'は、(A3')式のように平均$\beta' d_i$、共分散行列 Γ_0 の多変量正規分布にしたがう。ここでの目的は、観測されたリーセンシーとフリークエンシーデータからパラメータ$\{\theta_i, y_i, z_i, \forall i; \beta, \Gamma_0\}$を推定することである。

3.3. 事前分布

(A3')式から、λ_i とμ_i の事前分布は多変量対数正規分布に設定する。ハイパ・パラメータβとΓ_0は、事前分布としてそれぞれ多変量正規分布と逆ウィッシャー分布を仮定した標準的なベイズ回帰モデルとする。

$$\beta \sim MVN(\beta_0, \Sigma_0), \quad \Gamma_0 \sim IW(\nu_{00}, \Gamma_{00})$$

定数 $(\beta_0, \Sigma_0, \nu_{00}, \Gamma_{00})$ は拡散事前分布となるような値を選択する。

3.4. MCMC ステップ

パラメータ$\{\theta_i, \tau_i, z_i, \forall i; \beta, \Gamma_0\}$の推定は、前節の事前分布に基づいて MCMC 法で行う。これは、各パラメータを残りのパラメータが既知と仮定した条件付確率密度から逐次的に乱数発生させ、このプロセスを何回も繰り返すことによって、収束した分布はパラメータの同時確率密度になることが知られている。実際のステップは以下になる。

[1] $\theta_i^{(0)}\ \forall i$ の初期値を決める.

[2] 各顧客 i に対して

[2a] (２)式に基づいて $\{z_i \mid \theta_i\}$ を乱数発生させる.

[2b] もし $z_i = 0$ の場合, 切断指数分布から $\{y_i \mid z_i, \theta_i\}$ を乱数発生させる.

[2c] (１)式に基づいて $\{\theta_i \mid z_i, y_i\}$ を乱数発生させる.

[3] 多変量ベイズ回帰モデルによって $\{\beta, \Gamma_0 \mid \theta_i, \forall i\}$ を更新する.

[4] 収束が得られるまでステップ[2]~[3] を繰り返す.

以下に各ステップの詳細を説明する。

[2a]　θ_i は前回の繰り返しで得られた λ_i と μ_i を指数変換して求め、その θ_i を（２）式に代入することによって乱数を発生させるための $P(z_i = 1)$ が求められる。

[2b]　$z_i = 0$ は顧客 i が最終購買 t_i の後、観測終了時点 T_i より前に離脱したことを意味する。よって生存時間 y_i は仮定(A2)によりパラメータ μ_i の指数分布にしたがうが、それは $t_i < y_i < T_i$ の範囲に限定されなければならない。

[2c]　発生された z_i と y_i から (1)式の尤度関数を計算し、それに事前分布を乗じることで λ_i と μ_i を乱数発生させるための事後分布が得られる。ここでの事前分布（対数正規分布) は尤度関数（(1)式）に対して共役でないため、独立ＭＨアルゴリズムによって、まず λ_i、そして次に μ_i を発生させる。提案分布としては、受容確率が４０％程度になるように分散を任意に指定できる対数正規分布を用いた。

[3]　多変量ベイズ回帰モデルは標準的な手法なので、テキストブックなどを参照して欲しい (Congdon 2001; Gelman, Carlin, Stern, and Rubin 1995; Rossi, Allenby, and McCulloch 2005)。

４. 実証分析

本論文では提案モデルを HB(hierarchical Bayes) モデルと呼び、実際のデータを用いて既存の

Pareto/NBD モデルとの比較、検証を試みる。データは、米国 CDNOW の E コマースと日本のデパートの FSP(フリークエント・ショッパーズ・プログラム)から収集された顧客購買記録である。FSP は通称、ポイントカード制度とも呼ばれている。

4.1. CDNOW E コマース・データ

このデータは Fader, Hardie and Lee (2005a, 2005b)で使われたもので、CDNOW のウェブサイトで買われた音楽ＣＤの顧客別購買履歴を 78 週間分(1/1/97~6/30/98)集めたものである。データベースは、最初の 12 週間に CDNOW のメンバーになった 2357 人分の購買記録が含まれている。Fader, Hardie and Lee と同様に、最初と最後の 39 週間をそれぞれモデルの推定と検証に使った。したがって、観測期間(T)は顧客がいつメンバーになったかによって 27 週間～39 週間と異なる。過半数(60%, 1411 人)の顧客は推定期間にリピート購買をしておらず(つまり x=0)、モデル検証のデータとしては難しいものとなっている。データベースには顧客の人口統計的情報が含まれていないため、初期購買金額($)をモデルの説明変数として用いた。推定用データの記述統計が表１に示されている。

表 1.　CDNOW：データの記述統計

	平均	標準偏差	最小	最大
リピート購買数	1.04	2.19	0	29
観測期間 T (日数)	229.01	23.29	189	272
リーセンシー (T-t) (日数)	181.09	77.11	0	272
初期購買金額 (ドル)	32.99	34.66	0	506.97

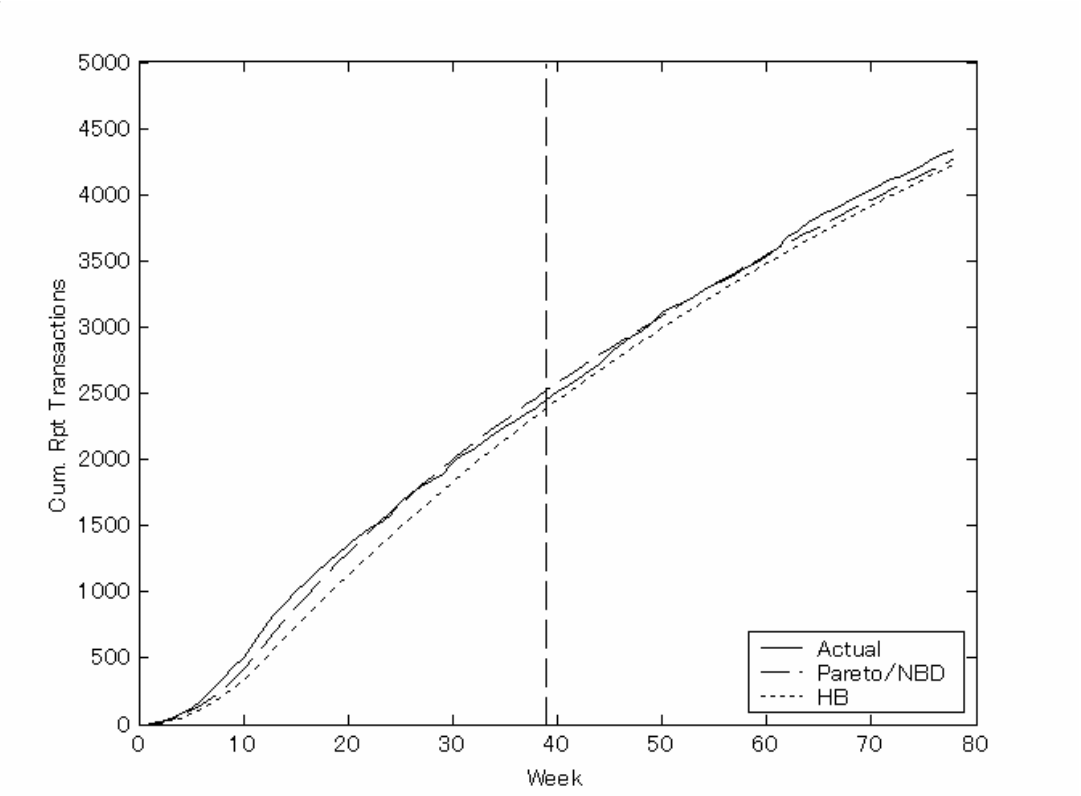
MCMC ステップは 14,000 回繰り返され、そのうち最後の 4,000 サンプルをパラメータの事後分布の構築に使った。収束はグラフ上での目視に加えて、Geweke のテスト(Geweke 1992)で確認した。推定用データに対するフィットと検証用データでの予測精度に関して、HB モデルと Pareto/NBD モデルとを比較した。非集計レベルにおけるモデルの精度指標として、顧客別に推定された購買回数と実際に観測された回数の相関係数と平均二乗誤差(MSE)を用いた。集計レベルでの精度指標としては、週別の累積購買回数を平均２乗誤差率の平方根(RMS)で評価した。表２にその結果が報告されている。非集計レベルでは両モデルの精度は似ているが、集計レベルの指標では Pareto/NBD モデルの方が若干、優れている。このことは週別累積購買回数を時系列にプロットした図４からも確認できる。グラフ内に描かれた垂直の点

線は、推定期間と検証期間の境界を表している。

表 2.　CDNOW：モデルのフィット

精度指標		Pareto/NBD	HB モデル
非集計レベルの指標			
相関係数	検証データ	0.63	0. 62
	推定データ	1.00	0.98
MSE	検証データ	2.57	2.61
	推定データ	0.64	0.58
集計レベルの指標			
時系列 RMS	検証データ	55.2	97.5
	推定データ	68.2	167.6
	全データ	61.9	136.7

図 4.　CDNOW：週別累積リピート購買数



非集計レベルの精度は、推定期間の購買回数ごとに顧客をグループ化し、検証期間での平均購買回数をプロットした図５からも視覚的に確認できる。同様のグラフは Fader, Hardie and Lee (2005a, b)でも採用されている。

図 5.　CDNOW：推定期間の購買回数別にみた検証期間の期待購買回数

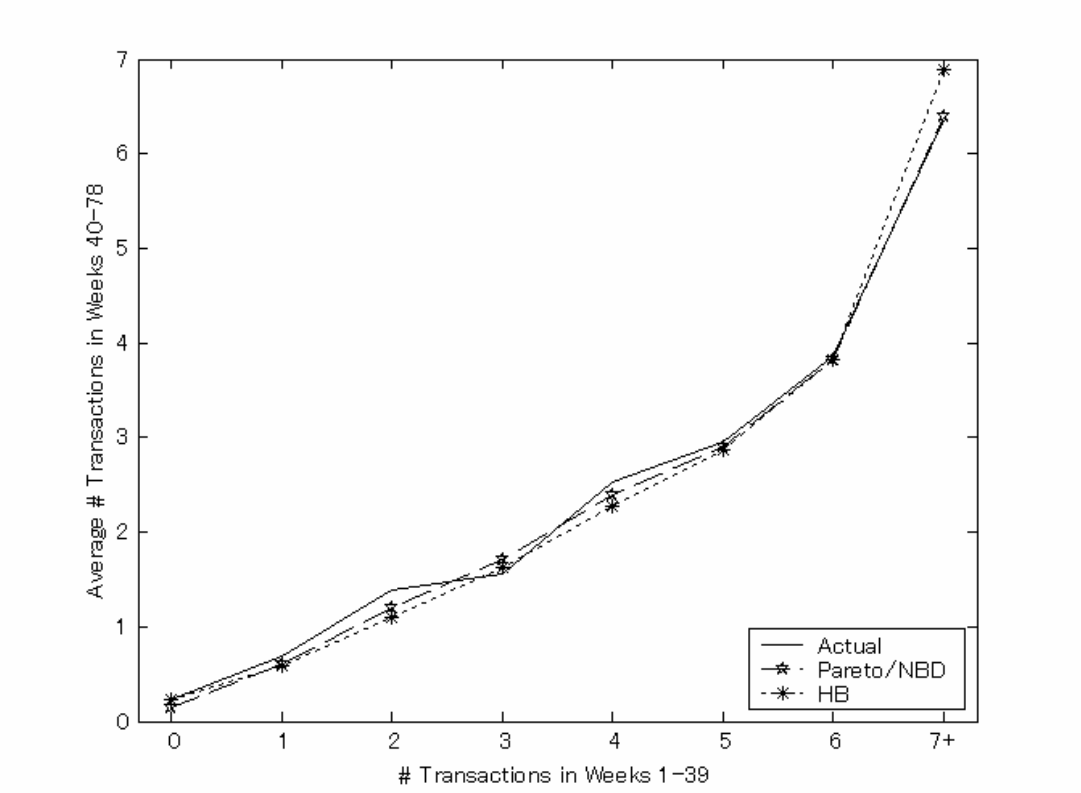


表 3 は HB モデルのパラメータの推定結果を表したもので、事後平均とカッコ内に標準誤差の目安として 2.5%と 97.5%の領域が示されている。左辺はλとμの対数であることに注意されたい。

表 3.　CDNOW：モデルの推定結果

(カッコ内の数字は 2.5％と 97.5％の領域を表す)

＊ は５％で有意を表す

		HB
購買頻度	**切片**	-4.19 (-4.33,-4.06)
λ	**初期購買金額 ($ 10^{-3})**	3.16* (1.54, 4.79)
離脱率	**切片**	-4.36 (-4.60, -4.15)
μ	**初期購買金額 ($ 10^{-3})**	-0.042 (-1.15, 1.06)
相関係数 (log(λ), log(μ))		-0.07 (-0.34, 0.26)
周辺対数尤度		-1385

唯一の説明変数である初期購買金額は、頻度に対して有意に正となっている。最初の購買金額が高い顧客ほど、その後より頻繁に購買することを意味する。この説明変数は離脱率に対しては有意でないため、初期購買金額が高くても低くても顧客は同様に離脱する傾向にある。また、ハイパ・パラメータの共分散行列から導かれた log(λ)と log(μ)の相関が-0.07 で、有意でないことが分かる。このことは、図 6 に示された 4000 回の MCMC ステップから得られた相関係数の分布でも確認できる。図 7 は各顧客のλ_i とμ_i (i=1,…,2357)の事後平均を散布図としてプロットしたものである。購買頻度パラメータλの値は離脱率パラメータμと比較して顧客間により大きな違いがあるが、この２つのパラメータの間に特別な関係は見受けられない。よって、このデータでは Pareto/NBD モデルの仮定であるλとμの分布の独立性が満たされていると言えよう。

図 6.　CDNOW：MCMC 法で推定された log(λ)と log(μ)の相関係数のヒストグラム

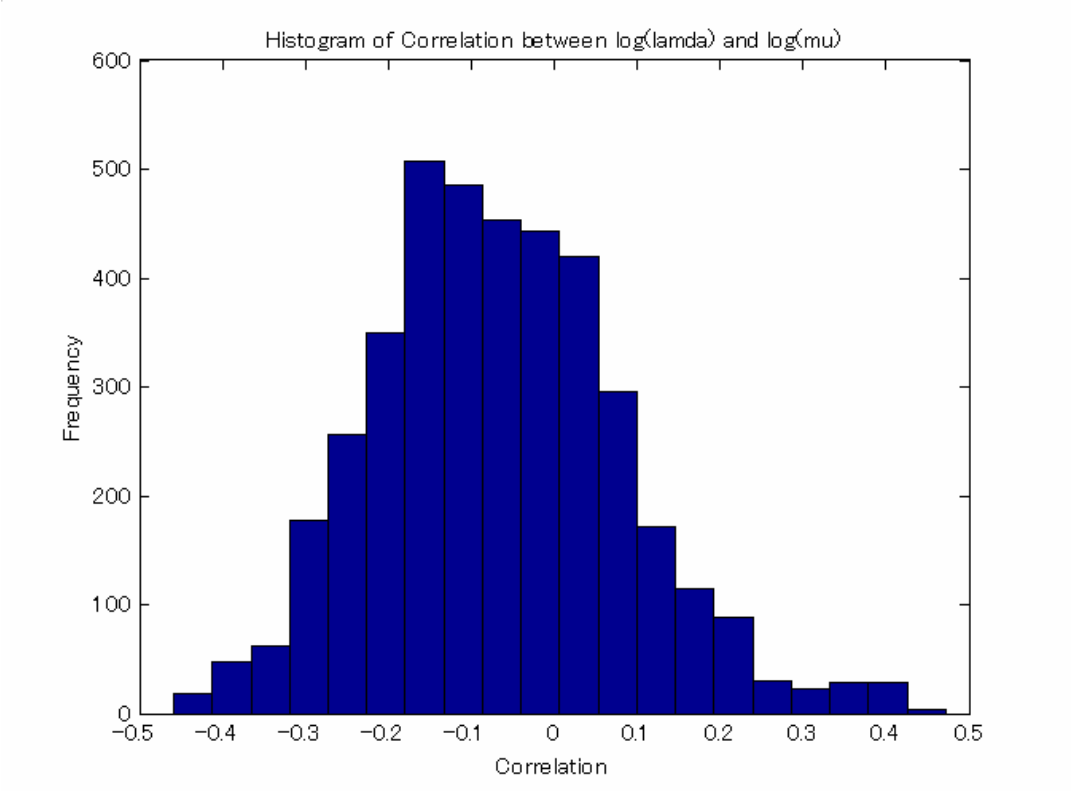


図 7.　CDNOW：λとμの顧客別事後平均の散布図

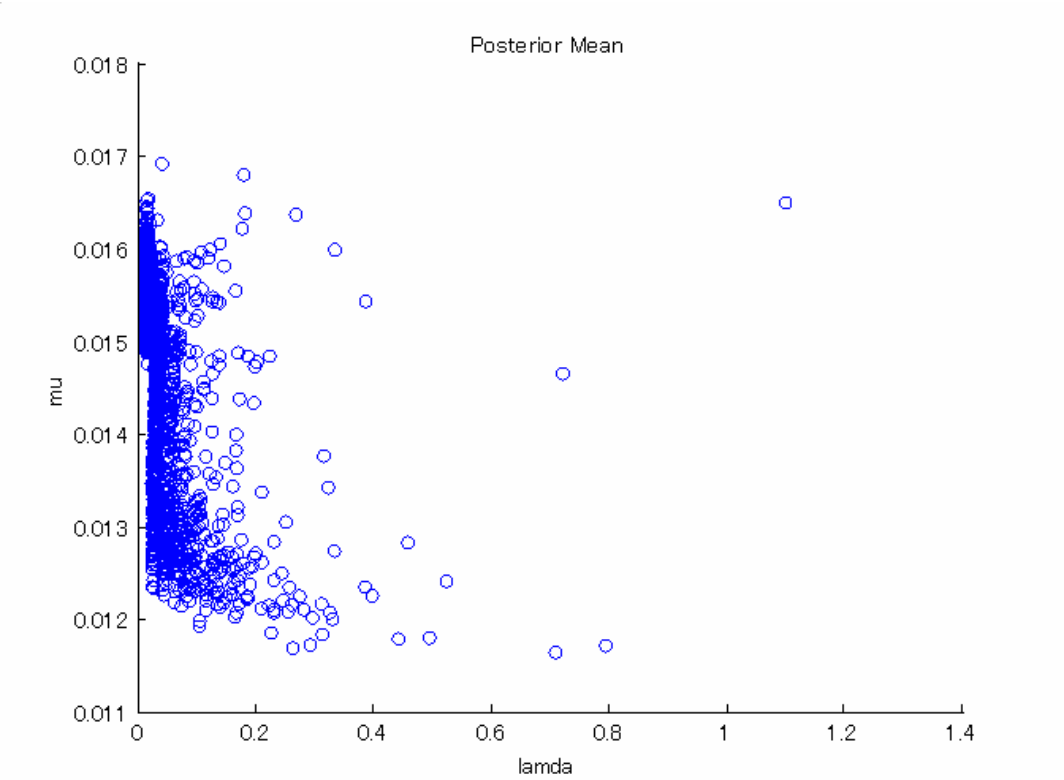


表４は、λ_i と μ_i の事後平均、最終購買以降の期待生存時間、１年後の維持率、観測終了時点での生存確率、検証期間中の期待購買回数という６つの顧客別統計値を、期待購買回数に関してベスト 10 とワースト 10 の 20 人の顧客に対して示したものである。また最後の３行は、それぞれ 2357 人の６つの統計値の平均値、最小値、最大値を表す。たとえば、観測終了時点(9/30/97)の生存確率の平均は 0.628 であるが、顧客によって 0.007 から 1.000 と違い、検証期間 39 週間中の期待購買回数も、平均は 0.75 回であるが、最小 0.03 回から最大 25.01 回と大きく異なる。これらの統計値は、顧客をランク付けするなど実務で有用である。HB モデルでは MCMC ステップの副産物として顧客ごとに λ_i と μ_i の分布が得られるので、λ_i と μ_i の関数で表される統計値 (例えば(2)式) であればその分布も簡単に求められる。表４の最後の４列の統計値も、そのようにして求めた。これに対して SMC は、最後の２列の統計値を複雑な積分から導き、それが論文の主結果であると提唱としている (彼らの論文中の(11)~(13)式と(22)式)。MCMC 法では点推定値ではなくパラメータの分布自体が得られるため(例えば図 4)、統計的仮説検定を適用することも容易である。

表 4. ベスト 10 顧客とワースト 10 顧客の統計値

顧客ランク	事後平均 (λ)	事後平均 (μ)	期待生存期間の平均 (年)	1 年後の維持率	観測終了時点での生存確率	検証期間中の期待購買回数
1	0.793	0.0117	1.88	0.573	0.998	25.01
2	0.708	0.0117	1.91	0.572	0.996	22.28
3	0.523	0.0124	1.78	0.554	0.994	16.21
4	0.494	0.0118	1.84	0.570	0.998	15.59
5	0.442	0.0118	1.84	0.569	0.990	13.80
6	0.397	0.0123	1.81	0.558	0.985	12.27
7	0.386	0.0124	1.79	0.557	0.980	11.82
8	0.458	0.0128	1.69	0.546	0.757	10.61
9	0.330	0.0120	1.78	0.562	0.994	10.30
10	0.325	0.0121	1.85	0.563	0.992	10.15
…	**…**	**…**	**…**	**…**	**…**	**…**
2348	0.015	0.0157	1.44	0.486	0.503	0.18
2349	0.016	0.0162	1.38	0.476	0.493	0.18
2350	0.015	0.0158	1.40	0.479	0.496	0.18
2351	0.015	0.0155	1.48	0.493	0.512	0.18
2352	0.014	0.0159	1.37	0.481	0.512	0.18
2353	0.015	0.0159	1.45	0.484	0.505	0.18
2354	0.015	0.0164	1.41	0.477	0.495	0.18
2355	0.015	0.0159	1.41	0.482	0.502	0.18
2356	0.015	0.0161	1.39	0.478	0.500	0.18
2357	1.097	0.0165	1.36	0.471	0.007	0.03
平均	0.038	0.0149	1.51	0.502	0.628	0.75
最小	0.014	0.0117	1.33	0.463	0.007	0.03
最大	1.097	0.0169	1.91	0.573	1.000	25.01

4.2. デパートの FSP データ

このデータは日本の某デパートにおけるFSPメンバーの購買履歴である。１店舗からのデータであるが、１０フロア以上にわたって衣料、家具、内装、家電、玩具、グルメ食品などさまざまな商品の購買が記録されている。観測期間は 2000 年 7 月 1 日から 2001 年 6 月 29 日までの 52 週間である。計算上の配慮から、2000 年 7 月中にFSPのメンバーになった顧客の中から 400 人をランダムに抽出し、分析の対象とした。最初と最後の 26 週間でデータを、それぞれ推定用と検証用に分けた。同日の複数レシートは１回の購買（店舗訪問）として統合

し、負の金額（返品など）は購買回数としてカウントしなかった[3]。推定用データの記述統計が表5に示されている。リピート購買回数x_iは0回が17人いるが101回という顧客もおり、１日おきぐらいに購買している顧客も多数いる。購買間隔日数の分布を顧客別に調べると、概ね指数分布の形状をしていることから、この購買プロセスはポアソン仮定を満たしていることが確認できる。

表 5.　デパート：データの記述統計

	平均	標準偏差	最小	最大
リピート購買数	16.02	16.79	0	101
観測期間 T (日数)	171.24	8.81	151	181
リーセンシー (T-t) (日数)	24.94	42.82	0	181
平均購買金額 ($\times 10^5$ yen)	0.067	0.120	0.0022	1.830
FOOD	0.79	0.273	0	1
AGE	52.7	14.6	22	87
FEMALE	0.93	0.25	0	1

データベースに含まれる顧客特性に関する情報は、性別、年令と住所である。顧客の多くが通勤・通学途中の乗り換えの際にこのデパートに寄るため、住所と店舗との地理的な距離は必ずしも店へのアクセスの容易さと関係していない。ここでは総訪問回数の中で食品を購入した訪問回数の割合を変数 FOOD と定義し、これを店舗へのアクセスのしやすさを表す説明変数としてモデルに組み込んだ。したがって FOOD は 0 から 1 の値、もし顧客が全ての店舗訪問で食品を購買していれば 1、２回の訪問に対して食品の購買が１回の割合であれば 0.5、全ての店舗訪問で食品を一度も購買していなければ 0、となる。もう一つの顧客特性変数として、データベースから１回当たりの平均購買金額を作った。説明変数のスケールをそろえるために、平均購買金額は 10^{-4} 円、年令は 10^{-2} 才の単位となっている。

表 6 は、推定用と検証用データにおける Pareto/NBD と HB モデルの集計レベルと非集計レベルのフィットを示したものである。両方のモデルとも似たような精度であるが、HB モデルの方が若干優れている。推定期間の購買回数ごとに顧客をグループ化し検証期間での平均購買回数をプロットした図８でも、２つのモデルは類似した結果となった。

[3] 本論文では、「訪問」と「購買」を同意に使う。

表 6.　デパート：モデルのフィット

精度指標		Pareto/NBD	HB M3
非集計レベルの指標			
相関係数	検証データ	0.90	0. 90
	推定データ	1.00	1.00
MSE	検証データ	58.2	58.6
	推定データ	1.22	1.16
集計レベルの指標			
時系列 RMS	検証データ	222.1	213.6
	推定データ	374.5	326.1
	全データ	307.9	275.6

図 8.　デパート：推定期間の購買回数別にみた検証期間の期待購買回数

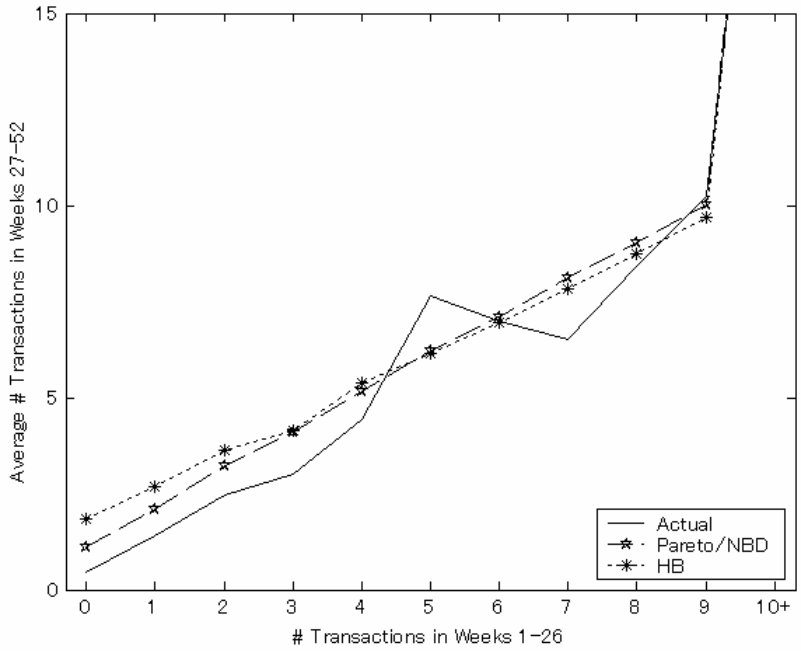
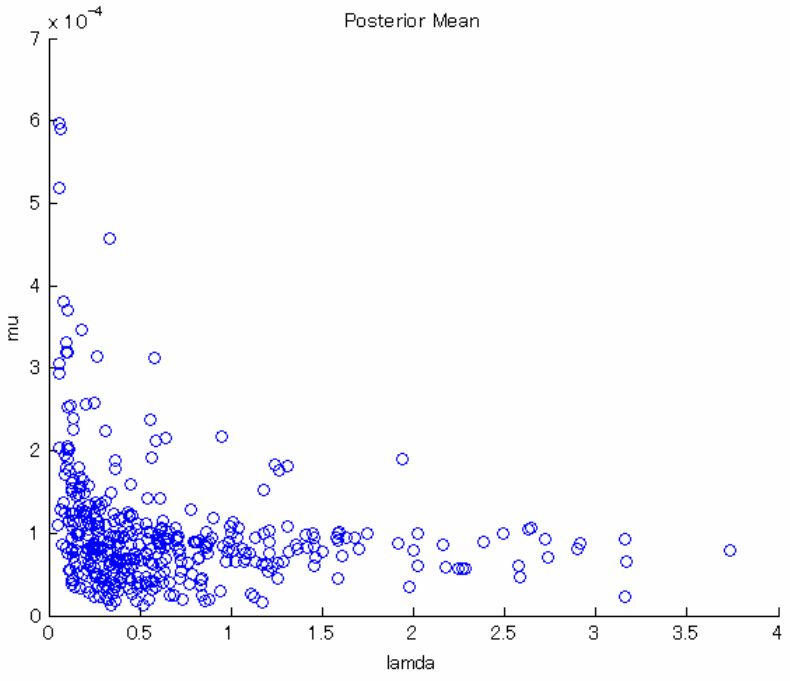


表 7 は、この HB モデル(M3)の他に、切片のみと２つの説明変数のみを組み込んだ２つのサブモデル(M1、M2)の推定結果を示したものである。説明変数を加えてもパラメータの推定値は安定している。周辺対数尤度によるとベストモデルは M3 となっている。購買頻度と離脱率のパラメータの相関は-0.12 で有意ではない。このことは、顧客別λ_iとμ_iをプロットした図 9 の散布図からも確認できる。

表 7.　デパート：モデルの推定結果
(カッコ内の数字は 2.5%と 97.5%の領域を表す)
* は５％で有意を表す

		HB M1	HB M2	HB M3
購買頻度 λ	切片	-0.89 (-1.00, -0.79)	-1.92 (-2.27, -1.56)	-2.09 (-2.62, -1.56)
	平均 購買金額	---	-0.18* (-0.32, -0.05)	-0.19* (-0.33, -0.05)
	FOOD	---	1.44* (1.05, 1.84)	1.43* (1.02, 1.84)
	AGE	---	---	0.08 (-0.58, 0.72)
	FEMALE	---	---	0.15 (-0.22, 0.52)
離脱率 μ	切片	-8.74 (-10.60, -7.50)	-8.87 (-11.53, -6.70)	-8.75 (-10.79, -6.77)
	平均 購買金額	---	-0.33 (-2.39, 0.78)	-1.24 (-2.53, 0.72)
	FOOD	---	-0.73 (-2.77, 1.40)	-0.59 (-3.14, 1.56)
	AGE	---	---	-0.39 (-2.73, 1.92)
	FEMALE	---	---	-0.58 (-2.34, 1.33)
相関係数 (log(λ), log(μ))		-0.14 (-0.55, 0.29)	-0.19 (-0.57, 0.20)	-0.12 (-0.51, 0.33)
周辺対数尤度		-1695	-1682	-1650

図 9.　デパート：λとμの顧客別事後平均の散布図



購買頻度に対して有意な説明変数は、平均購買金額と FOOD である。年令と性別は購買頻度に影響していない。また離脱率に対しては、有意な説明変数は無かった。つまり、顧客生存時間は平均購買金額、FOOD、年令、性別で異ならない。

ここでの示唆は、食品購買の割合が高い顧客と１回あたりの平均購買金額が低い顧客ほど店舗を頻繁に訪れるということである。これは、このデパートのマネージャーに対するインタビューでの発言、「食品購入者は１回当たりの購買金額は低いが、頻繁に店舗を訪問するため重要な顧客と認識している。」とも一致する。食品はファッション、アクセサリー、家具・内装品と比較するとマージンも低いため、それだけでは特に利益にならない。しかしこのデパートは、キー顧客層を引きつけるために、近年、高価なグルメや輸入食材を取り揃えて食料品フロアを大々的に改装した。

5. 結論

顧客の離脱は直接には観測されない。企業はリーセンシーを使った経験則(例えば 3 ヶ月購買がなければ離脱)に頼って、この判断を下しているのが現状である。本論文では、既存の RF 分析に消費者行動理論に基づいたモデルを組み込むことによって、消費者の異質性を考慮し、この観測されない離脱の確率を推定できることを示した。

この研究で提案された HB モデルは、過去のマーケティング研究で十分に検証された Pareto/NBD モデルでも用いられている消費者行動の仮定：(1)ポアソン購買プロセス、(2)無記憶的離脱プロセス（定数ハザードモデル)、(3)両プロセスにおける顧客の異質性、を置いている。しかし、購買と離脱プロセスのパラメータが独立に分布しているという Pareto/NBD モデルの制約を課さないため、より柔軟なモデルとなっている。顧客の異質性は、Pareto/NBD モデルにおける混合分布の代わりに、階層ベイズの枠組みから事前分布として組み込まれるため、MCMC 法を用いることによって総計にまつわる複雑な積分が不要となる。メリットとしては、(1)概念、推定、計算の単純化、(2)モデルの柔軟性、(3)潜在変数の個人別推定、(4)正確な誤差の推定、(5)階層モデルへの発展性、(6)正式なベイズのパラダイム、などが挙げられる。

HB モデルは Pareto/NBD モデルと同様にデータにフィットすることが２つの実データから検証された。モデルのアウトプットとして、λ_i と μ_i の事後平均、最終購買以降の期待生存時間、

１年後の維持率、観測終了時点での生存確率、検証期間中の期待購買回数など、実際のマーケティングなどで有益な顧客指標が MCMC 推定法の副産物として得られる。これらの指標を Pareto/NBD モデルから計算するには、顧客ごと、指標ごとに複雑な数値積分を行わなければならない。

λとμを顧客特性の関数とした階層モデルからは、購買頻度が高い顧客や生存時間の長い顧客の特徴、たとえば人口統計的要因や１回当たりの平均購買金額の高低のような購買行動的要因など、ワン・トゥー・ワン・マーケティングで有用な知見が得られる。また HB モデルでは、購買と離脱プロセスのパラメータが独立に分布するという Pareto/NBD モデルの仮定を統計的に検定したり、λ_i と μ_i を散布図としてプロットすることによって視覚的な診断を行ったりすることが可能である。今回は、この仮定が両方のデータで満たされていることが確認された。

この研究の限界としては、第１に、現実の離脱率は観測されないため、推定された離脱率の妥当性を外的には評価できないことが挙げられる。これを克服するために、本論文ではモデルの検証としては第 4 節に記述されている３つの指標を用いた。非集計レベルでの精度指標として、顧客別に推定された購買回数と実際に観測された回数の相関係数と平均二乗誤差(MSE)を用い、さらに推定期間の購買回数ごとに顧客をグループ化して検証期間での平均購買回数を比較した。また集計レベルでの精度指標としては、週別の累積購買回数を RMS で評価した。

第２の限界は、消費者行動の仮定 1 と 2 が当てはまらない状況では、Pareto/NBD モデルや HB モデルが機能しないことが挙げられる(Chatfield and Goodhardt 1973)。したがって、適切な業界やカテゴリーを選択し、記述統計などで仮定が満たされているかをチェックすることが重要である。

付録：　生存確率と尤度関数の導出

まず、観測された購買履歴から顧客が生存している確率をベイズの定理に基づいて(A1)式のように表す。

$$P(\tau > T \mid \lambda, \mu, x, t, T) = P(\text{生存} \mid \text{履歴})$$

$$= \frac{P(\text{生存} \,\&\, \text{履歴})}{P(\text{履歴})} \tag{A1}$$

$$= \frac{P(\text{履歴} \mid \text{生存}) P(\text{生存})}{P(\text{履歴} \,\&\, \text{生存}) + P(\text{履歴} \,\&\, \text{死亡})}$$

すると、生存時間が指数分布なので

$$P(\text{生存}) = P(\tau > T) = e^{-\mu T}$$

となる。さらに、

$$P(\text{履歴} \mid \text{生存}) = P(x\text{回目の購買が}t\text{に起きる}\,\&\,[t,T]\text{に購買が起きない})$$
$$= \frac{\lambda^x t^{x-1}}{\Gamma(x)} e^{-\lambda t} \times e^{-\lambda(T-t)}$$
$$= \frac{\lambda^x t^{x-1}}{\Gamma(x)} e^{-\lambda T}$$

$$P(\text{履歴} \,\&\, \text{死亡}) = \int_t^T P(x\text{回目の購買が}t\text{に起きる}\,\&\,[t,y]\text{に購買が起きない}\,\&\, y \in [t,T]\text{に死亡}) dy$$
$$= \int_t^T \frac{\lambda^x t^{x-1}}{\Gamma(x)} e^{-\lambda t} \times e^{-\lambda(y-t)} \times \mu e^{-\mu y} dy$$
$$= \frac{\lambda^x t^{x-1}}{\Gamma(x)} \mu \int_t^T e^{-(\lambda+\mu) y} dy$$
$$= \frac{\lambda^x t^{x-1}}{\Gamma(x)} \frac{\mu}{\lambda+\mu} \left\{ e^{-(\lambda+\mu)t} - e^{-(\lambda+\mu)T} \right\}$$

を(A1)式に代入すると、下の式が得られる。

$$P(\tau > T \mid \lambda, \mu, x, t, T) = \frac{\frac{\lambda^x t^{x-1}}{\Gamma(x)} e^{-\lambda T} \times e^{-\mu T}}{\frac{\lambda^x t^{x-1}}{\Gamma(x)} e^{-\lambda T} \times e^{-\mu T} + \frac{\lambda^x t^{x-1}}{\Gamma(x)} \frac{\mu}{\lambda+\mu} \left\{ e^{-(\lambda+\mu)t} - e^{-(\lambda+\mu)T} \right\}}$$

$$= \frac{1}{1 + \frac{\mu}{\lambda+\mu} \left\{ e^{(\lambda+\mu)(T-t)} - 1 \right\}}$$

また、尤度関数は（A1）式の分母のパラメータに依存する部分なので、下の式のように表せる。

$$\begin{aligned} L(x, t, T \mid \lambda, \mu) &\propto P(\text{履歴}) \\ &= \frac{\lambda^x t^{x-1}}{\Gamma(x)} e^{-\lambda T} \times e^{-\mu T} + \frac{\lambda^x t^{x-1}}{\Gamma(x)} \frac{\mu}{\lambda+\mu} \left\{ e^{-(\lambda+\mu)t} - e^{-(\lambda+\mu)T} \right\} \\ &= \frac{\lambda^x t^{x-1}}{\Gamma(x)} \left\{ \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)T} + \frac{\mu}{\lambda+\mu} e^{-(\lambda+\mu)t} \right\} \end{aligned}$$

期間 t における期待購買回数は以下で表せる。

$$E[X(t) \mid \lambda, \mu] = \lambda E[\eta] = \frac{\lambda}{\mu} \left(1 - e^{-\mu t}\right) \qquad \text{where } \eta = \min(\tau, t). \tag{5}$$

その他の有用な統計は、

最終購買以後の期待生存時間 $= \frac{1}{\mu}$

1 年後の生存確率$= \exp(-52\mu)$、　ここでは時間の単位は週である。

参考文献

Abe, Makoto (2006), "Counting Your Customers One by One: An Individual Level RF Analysis Based on Consumer Behavior Theory", Working Paper, CIRJE-F-408, The University of Tokyo.

----- (2008), "Counting Your Customers One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model", Marketing Science, forthcoming.

Bass, Frank M., Moshe M. Givon, Manohar U. Kalwani, David Reibstein, Gordon P. Wright (1984), "An investigation into the order of the brand choice process," *Marketing Science*, 2 (4), 267-187.

Blattberg, Robert C. and John Deighton (1996), "Manage marketing by the customer equity test," *Harvard Business Review*, 74 (4), 136-144.

Chatfield, C. and G. J. Goodhardt (1973), "A consumer purchasing model with Erlang inter-purchase times," *Journal of the American Statstical Association* (December), 828-835.

Congdon, Peter (2001), *Bayesian Statistical Modelling*, London, UK: Wiley.

Ehrenberg, A. S. C. (1972), *Repeat-Buying: Theory and Applications*, Amsterdam; North-Holland.

Ehrenberg, A. S. C. (1988), *Repeat-Buying: Facts, Theory and Data*, 2nd Ed. New York; Oxford University Press.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005a), "'Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science*, 24 (2), 275-284.

-----, ----- and ----- (2005b), "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," *Journal of Marketing Research*, 42 (4), 415-430.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (1995), *Bayesian Data Analysis*, Boca Raton, Florida: Chapman & Hall.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based approaches to the Calculation of Posterior Moments," in J. M. Bernardo, J. M. Berger, A. P. Dawid and A. F. M. Smith, (eds.), *Bayesian Statistics 4*, 169-193, Oxford: Oxford University Press.

Park, Young-Hoon and Peter S. Fader (2004), "Modeling Browsing Behavior at Multiple Websites," *Marketing Science*, 23 (3), 280-303.

Reinartz, Werner J. and V. Kumar (2000), "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of*

Marketing, 64 (4), 17-35.

----- and ----- (2003), "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," *Journal of Marketing*, 67 (1), 77-99.

Rossi, Peter E., Greg Allenby and Rob McCulloch (2005), *Bayesian Statistics and Marketing*, London, UK: Wiley.

Schmittlein, David C., Donald G. Morrison, and Richard Colombo (1987), "Counting your customers: Who are they and what will they do next?" *Management Science*, 33 (1), 1-24.

----- and Robert A. Peterson (1994), "Customer Base Analysis: An Industrial Purchase Process Application," *Marketing Science*, 13 (1), 41-67.

Tanner, Martin A. and Wing Hung Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82 (398), Theory and Methods, 528-540.