『岡山商大論叢』（岡山商科大学）
第49巻第１号　2013年７月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol.49　No.１　July 2013

《論　　説》

ピットマンモデルを利用した寸法指標のノンパラメトリック推定

佐　井　至　道

Nonparametric Estimation of Size Indices Using Pitman Model

Shido Sai

１　はじめに

官庁における統計調査の多くは標本調査によって行われる。調査された情報がそのままレコードとして並べられた個票データの公開に関して，秘匿方法，リスク評価方法，有用性の評価方法に関する研究が行われてきた。

このうち，リスク評価方法についてはいくつかの指標が提案されているが，特に国内においては寸法指標を用いることが主流となっている。一般に第三者が既に情報を持っていて，個票データが公開された場合に，その値を利用して個体の識別を行える変数をキー変数と呼ぶが，寸法指標とは，キー変数の値の同じ組み合わせに対する個体数をそれぞれ求めた場合の，個体数別の頻度分布である。個票データが標本から作成されている場合には，リスク評価には標本寸法指標だけでなく，標本寸法指標を基にして推定された母集団寸法指標が必要になる。

母集団寸法指標の推定には，この分野の初期の研究ではポアソンガンマモデル（Bethlehem et al.（1990））が用いられていたが，実データへの当てはまりの悪さが指摘され，現在ではピットマンモデル（Pitman（1995））のような，より柔軟なモデルが用いられることが多い。

一方でノンパラメトリックな方法が Engen（1978）によって提案されたが，実際の寸法指標からかけ離れた値が推定されるとともに，標本寸法指標のわずかな変化に対して母集団寸法指標の推定値が大きく変化することから，長い間利用されることはなかった。佐井（2002, 2003）では，標本寸法指標が得られているときの母集団寸法指標の尤度関数をポアソン分布の確率関数の積で近似するとともに，母集団寸法指標の推定値に数種類の簡単な制約を課すことによって，推定を安定させる方法を提案した。また，佐井（2006）では，制約条件をペナルティー関数で表現することにより，計算時間の短縮を図った。これにより，官庁統計などの比較的大きい母集団のデータに対してもノンパラメトリックな方法が適用可能となった。

本論文では，制約条件を表すペナルティー関数の代わりに，ピットマンモデルの確率関数を利用する方法を新たに提案する。

まず２節では，これまでに提案されている標本寸法指標に基づく母集団寸法指標のノンパラメトリック推定法について紹介する。

2.1 節では，佐井（2002, 2003）で提案した母集団寸法指標の尤度関数の近似方法と，母集団寸法指標の推定値に課すべき制約条件，最尤推定値の探索方法について述べる。本論文で提案する方法と同様に，この方法では母集団寸法指標の推定値を非負整数に限定している。

2.2 節では，佐井（2006）で提案した母集団寸法指標の推定値に対する制約条件のペナルティー関数化での書き換えと，尤度関数とペナルティー関数を結合した新たな関数の最大化について例も用いて説明する。この方法では母集団寸法指標の推定値は非負の実数に拡張されている。

３節で新たな推定方法を提案するが，まず3.1 節では，ピットマンモデルを用いた母集団寸法指標の推定について説明する。

3.2 節において，2.2 節のペナルティー関数の代わりにピットマンモデルの確率関数を用いる方法を提案する。また，母集団寸法指標の推定値を非負整数に限定する新たな探索法についても述べる。これは2.1 節の方法と比較して，推定値の移動に柔軟性がある。

3.3 節では推定値の具体的な探索方法について例を用いて説明する。提案する方法では，母集団寸法指標の尤度関数とピットマンモデルの確率関数の重みのバランスの決定が難しいが，ピットマンモデルの重みを段階的に小さくすることで対処する。

3.4 節では，実際の個票データについて適用を行い，有用性と問題点を探ることにする。

２　母集団寸法指標のノンパラメトリック最尤推定

2.1　母集団寸法指標の制約付き最尤推定

この節では，佐井（2002, 2003）で提案した方法を簡略に紹介する。

非復元単純無作為抽出された標本の大きさ（個体数）を n とする。標本から個票データが作成されているとき，すべてのキー変数の値の組み合わせの一つ一つを分割表の呼び方に倣ってセルと呼ぶ。そのうちサイズ l のセル数，すなわち l 個の個体が入っているセル数を s_l と表して標本寸法指標と呼ぶ。また母集団の大きさを N，抽出率を $\lambda = n/N$ として，母集団寸法指標を S_l とする。

佐井（2002, 2003）では，標本寸法指標に基づく母集団寸法指標の制約付きノンパラメトリック推定法を提案した。その際，サイズの最大値として頻度が 0 でないサイズの最大値である L を用いた。母集団寸法指標が未知であれば，L もまた未知であるが，推定結果に影響を与えない範囲で比較的小さい値を選んだ。本論文でも実データを扱う場合には同様の設定を行うが, 数式では頻度が 0 でない可能性のあるサイズの最大値を用いて，標本寸法指標は $(s_1, s_2, \ldots, s_n)$，母集団寸法指標は $(S_1, S_2, \ldots, S_N)$ と表す。

標本寸法指標 $(s_1, s_2, \ldots, s_n)$ が得られたときに，尤度を最大にするような母集団寸法指標 $(S_1, S_2, \ldots, S_N)$ を求める必要があるが，尤度関数には膨大な組み合わせについての和が含まれるため計算には時間を要し，母集団の大きさ N が100程度でも実用性を持たない。

幸い，抽出率 λ が十分小さいときに，この尤度関数はポアソン分布の確率関数の積を用いて

$$\begin{aligned} & \mathrm{L_{ap}}(S_1, S_2, \ldots, S_N \mid s_1, s_2, \ldots, s_n) \\ = \quad & \frac{1}{{}_N C_n \cdot \lambda^n (1-\lambda)^{N-n}} \prod_{l'=1}^{L} \frac{e^{-\mu_{l'}} {\mu_{l'}}^{s_{l'}}}{s_{l'}!} \end{aligned} \tag{1}$$

と近似することが可能で，計算時間を大幅に短縮することができる。ただし

$$\mu_{l'} = \sum_{l=l'}^{L} S_l \cdot {}_l C_{l'} \lambda^{l'} (1-\lambda)^{l-l'} \tag{2}$$

である。

(1) 式の近似された尤度関数を最大にする母集団寸法指標は，特に抽出率 λ が小さい場合には多くの値が負となることが多く，推定値が非負の条件の下で尤度の最大値を探索しても，0が飛び飛びに現れるような母集団寸法指標の推定値が得られることが多い。

例えば，$N = 3000$, $n = 600$ の簡単な例を考える。母集団寸法指標を $(S_1, S_2, S_3) = (2100, 300, 100)$ として，標本を非復元単純無作為抽出して $(s_1, s_2, s_3) = (548, 23, 2)$ が得られている。なお，サイズが4以上の頻度はすべて0であるため省略する。以後，この例を例1と呼ぶ。

この標本寸法指標を基に，サイズの最大値を3として，非負整数という条件の下で (1) 式を最大にする母集団寸法指標の推定値を求めると $(\hat{S}_1, \hat{S}_2, \hat{S}_3) = (2277, 0, 241)$ となってしまう。

そこで，佐井（2003）では，母集団寸法指標の推定値 $\hat{S}_l$ にいくつかの制約を置く場合について，実データを基に比較を行った。ここでは本論文

で使用する制約条件のみを列挙する。

(a) $\hat{S}_l \geq 0 \quad (l = 1, 2, \ldots, N)$

(b) $\sum_{l=1}^{N} l \cdot \hat{S}_l = N$

(c) $\hat{S}_{l-1} \geq \hat{S}_l \quad (l = 2, 3, \ldots, N)$

(d) $2 \cdot \log \hat{S}_l \leq \log \hat{S}_{l-1} + \log \hat{S}_{l+1} \quad (l = 2, 3, \ldots, N-1)$

(a), (b)は自然な制約である。(c)は母集団寸法指標の推定値が単調減少，(d) は母集団寸法指標の推定値の対数が下に凸という条件であり，実データの多くの寸法指標がこれらの条件を満たすという経験から設定されたものである。

図１に，N = 3000でサイズの最大値を３とした場合に上記四つの制約条件を満たす領域を斜線で示している。底面の二つの軸は S_1，S_2 を，底面に垂直な軸は S_3 を表している。立方体の三つの頂点を結ぶ三角形の周および内部は条件 (a)，(b) を表し，その中を通る２本の線分と曲線はそれぞれ条件 (c) と (d) を規定するものである。実際のデータでは，サイズの最大値は数十，あるいは100を超える場合も多いが，本論文では視覚的

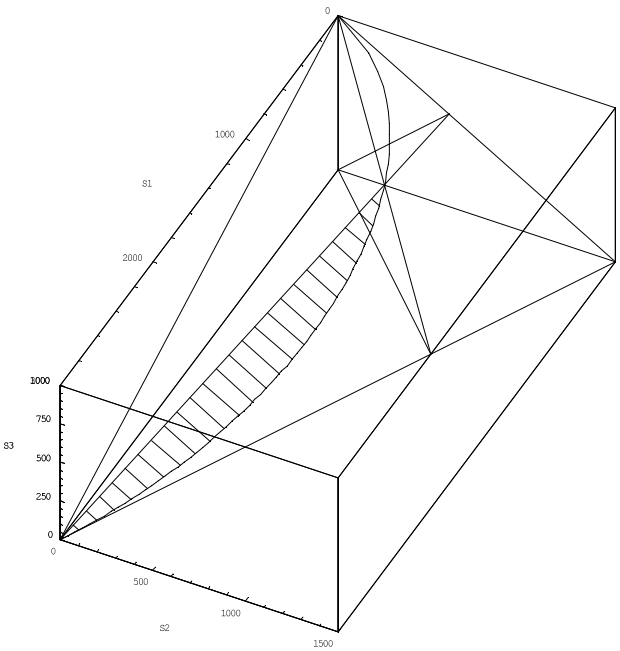

図１　制約条件 (a) ～ (d) を満たす領域 (N=3000, サイズの最大値が３の場合)

に表現するために最大値が３の簡単な例を適宜用いることにする。

すべての母集団寸法指標について網羅的に尤度を計算することは困難であるため, 佐井 (2003) では数種類の最大尤度の探索法について比較を行った。その中では，サイズ１以外の任意の一つのサイズの頻度を ± １の範囲で増減させ，個体数の過不足をサイズ１の頻度で調整する方法が推定結果と計算時間を勘案して最も優れていた。

これは探索の p 番目のステップで暫定的な推定値

$$(\hat{S}_{1(p)}, \ldots, \hat{S}_{l-1(p)}, \hat{S}_{l(p)}, \hat{S}_{l+1(p)}, \ldots, \hat{S}_{N(p)}) \tag{3}$$

における (1) 式の値が得られているとき，$2 \leq l \leq N$，$\hat{S}_{1(p)} \geq l$ であるすべての l について

$$(\hat{S}_{1(p)} - l, \ldots, \hat{S}_{l-1(p)}, \hat{S}_{l(p)} + 1, \hat{S}_{l+1(p)}, \ldots, \hat{S}_{N(p)}) \tag{4}$$

における (1) 式の値と，$2 \leq l \leq N$，$\hat{S}_{l(p)} > 0$ であるすべての l について

$$(\hat{S}_{1(p)} + l, \ldots, \hat{S}_{l-1(p)}, \hat{S}_{l(p)} - 1, \hat{S}_{l+1(p)}, \ldots, \hat{S}_{N(p)}) \tag{5}$$

における (1) 式の値を求め，その中で最大値をとる母集団寸法指標を p ＋１番目のステップの推定値

$$(\hat{S}_{1(p+1)}, \ldots, \hat{S}_{l-1(p+1)}, \hat{S}_{l(p+1)}, \hat{S}_{l+1(p+1)}, \ldots, \hat{S}_{N(p+1)}) \tag{6}$$

とする方法である。先に述べたように，実際には最大サイズは N の代わりに，ある程度小さな L を用いた。

2.2　ペナルティー関数の利用

この節では，佐井（2006）で提案した方法を簡略に紹介する。

渋谷（2005）の提案した滑らかなペナルティー関数

$$\mathrm{Pnlt}(x;\varepsilon) = \varepsilon \log(e^{\frac{x}{\varepsilon}} + 1) \tag{7}$$

を利用する。ただし，ε は正のパラメータである。

$y = \mathrm{Pnlt}(x;\varepsilon)$ は，x が負であれば $y = 0$，x が正であれば $y = x$ に近い値をとるが，$x = 0$ 付近は滑らかに結ばれており微分可能である。

前節で挙げた四つの制約条件のうち，(b) を除いた三つの条件をこの滑らかなペナルティー関数で表して，その和をペナルティー関数 $\mathrm{G}(S_1, S_2, \ldots, S_N)$ と表す。

$$
\begin{aligned}
& \mathrm{G}(S_1, S_2, \ldots, S_N) \\
= \quad & -c_1 \sum_{l=1}^{N} \mathrm{Pnlt}(-S_l;\ \varepsilon_1) - c_2 \sum_{l=2}^{N} \mathrm{Pnlt}(-(S_{l-1} - S_l);\ \varepsilon_2) \\
& -c_3 \sum_{l=2}^{N-1} \mathrm{Pnlt}(-(\log S_{l-1} + \log S_{l+1} - 2\log S_l);\ \varepsilon_3) \qquad (8)
\end{aligned}
$$

ただし，c_1，c_2，c_3 は，それぞれの制約条件の影響力を表す正のパラメータである。

近似された対数尤度関数にペナルティー関数を

$$
\begin{aligned}
& \mathrm{F}_1(S_1, S_2, \ldots, S_N \mid s_1, s_2, \ldots, s_n) \\
= \quad & \log \mathrm{L}_{\mathrm{ap}}(S_1, S_2, \ldots, S_N \mid s_1, s_2, \ldots, s_n) + \mathrm{G}(S_1, S_2, \ldots, S_N) \qquad (9)
\end{aligned}
$$

と加えることにより，関数 $\mathrm{F}_1(S_1, S_2, \ldots, S_N \mid s_1, s_2, \ldots, s_n)$ の最大値を求めるという単純化を行うことができる。

近似された尤度関数では，母集団寸法指標 S_l は (2) 式にのみ現れ，非負整数だけでなく，負も含めて実数値をとることが可能で，S_l で偏微分することが可能である。またペナルティー関数 (8) 式でも S_l は実数値をとり，S_l で偏微分することが可能で，(9) 式も S_l で偏微分可能である。

そこで，ピットマンモデルで推定された母集団寸法指標の推定値を初期値として，ステップごとに，$\mathrm{F}_1(\cdot)$ が増加する方向に各 $\hat{S}_l$ を独立に移動させる勾配法によって最大値を探索した。なお，ピットマンモデルを用いた

母集団寸法指標の推定については次節で説明する。

$\hat{S}_{l(p)}$ を p 番目のステップにおける母集団寸法指標の推定値とするとき，

$$\hat{S}_{l(p+1)} = \hat{S}_{l(p)} + t_{l(p)} \cdot \mathrm{g}\left(\frac{\partial \mathrm{F}_1}{\partial S_l}\right) \tag{10}$$

によって，$p+1$ 番目のステップの推定値 $\hat{S}_{l(p+1)}$ を求める。ただし，$\mathrm{g}(\cdot)$ は微係数を調整する関数である。また $t_{l(p)}$ は移動する距離をコントロールする係数で，サイズ l，そのステップでの母集団寸法指標の推定値 $\hat{S}_{l(p)}$ などを基に各ステップで変化させたが，$\mathrm{F}_1(\cdot)$ が増加しない場合には小さくし，基準値を下回ったところで探索を終了する。

簡単な例をいくつか示す。

まず，2.1 節で用いた例 1 を再び用いる。$N = 3000$, $(S_1, S_2, S_3) = (2100, 300, 100)$ の母集団から大きさ $n = 600$ の標本を 1 組抽出する。サイズの最大値は 3 であるが，ピットマンモデルのサイズを打ち切るとともに補正すれば，(2100, 300, 100) に近い寸法指標を表現することは可能である。標本から計算された寸法指標 $(s_1, s_2, s_3) = (548, 23, 2)$ を基に上記の方法で推定した場合の，推定値と探索の過程を図 2 に示す。図中の点は実際の母集団寸法指標，初期値，最終的な推定値で，右にある拡大図には探索の過程の向きを矢印で示している。推定値は $(\hat{S}_1, \hat{S}_2, \hat{S}_3) = (2131.5229, 174.0126, 173.4839)$ である。

次に，上と同じ母集団から大きさ $n = 600$の標本を10組抽出する。この例を例 1'と呼ぶ。それぞれの標本寸法指標を基にした母集団寸法指標の推定値と探索の過程を図 3 に示す。点や矢印は省略するが，すべて下から上へ向かって移動している。

この場合，実際の母集団寸法指標は制約条件を満たす領域の内部にあり，ピットマンモデルの推定値である初期値もすべて領域の内部にあるが，推定値の探索は領域の端付近で終了している。

別の母集団を考える。

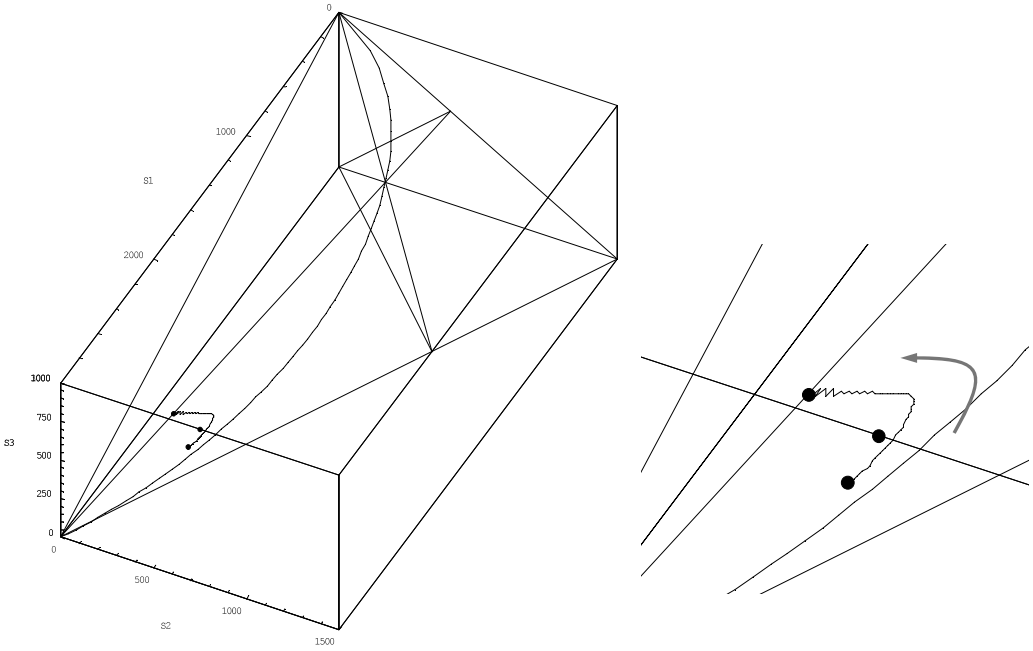

図２　ペナルティー関数を用いた推定値の探索の過程（例１）

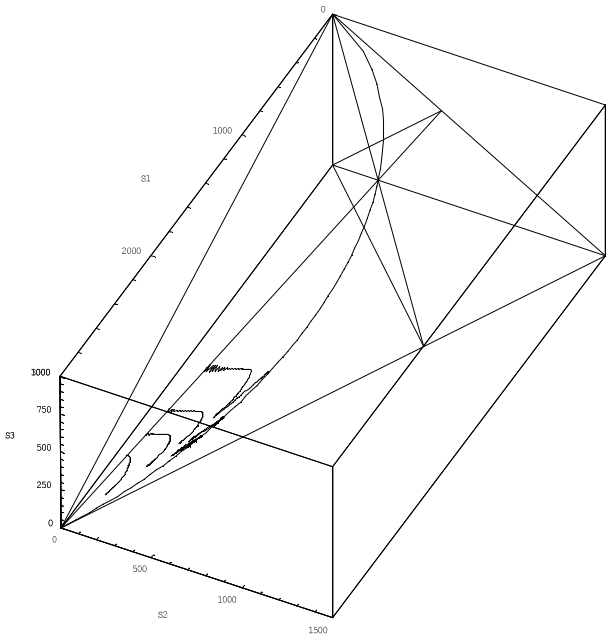

図３　ペナルティー関数を用いた推定値の探索の過程（例 1’）

$N = 3000$, $(S_1, S_2, S_3) = (1500, 300, 300)$ の母集団から大きさ $n = 600$ の標本を１組抽出し，その標本寸法指標 $(s_1, s_2, s_3) = (516, 39, 2)$ が得られている。この例を例２と呼ぶ。この標本寸法指標を基にした母集団寸法指標の推定値と探索の過程を図４に示す。推定値は $(\hat{S}_1, \hat{S}_2, \hat{S}_3) = (1500.0000, 375.0000, 250.0000)$ である。

ピットマンモデルでは，サイズの打ち切りや補正を行っても，この母集団寸法指標への当てはまりは悪い。この場合には，ピットマンモデルによる初期値よりもノンパラメトリック法による推定値の方が実際の母集団寸法指標に近い。

次に，上と同じ母集団から大きさ $n = 600$の標本を10組抽出する。この

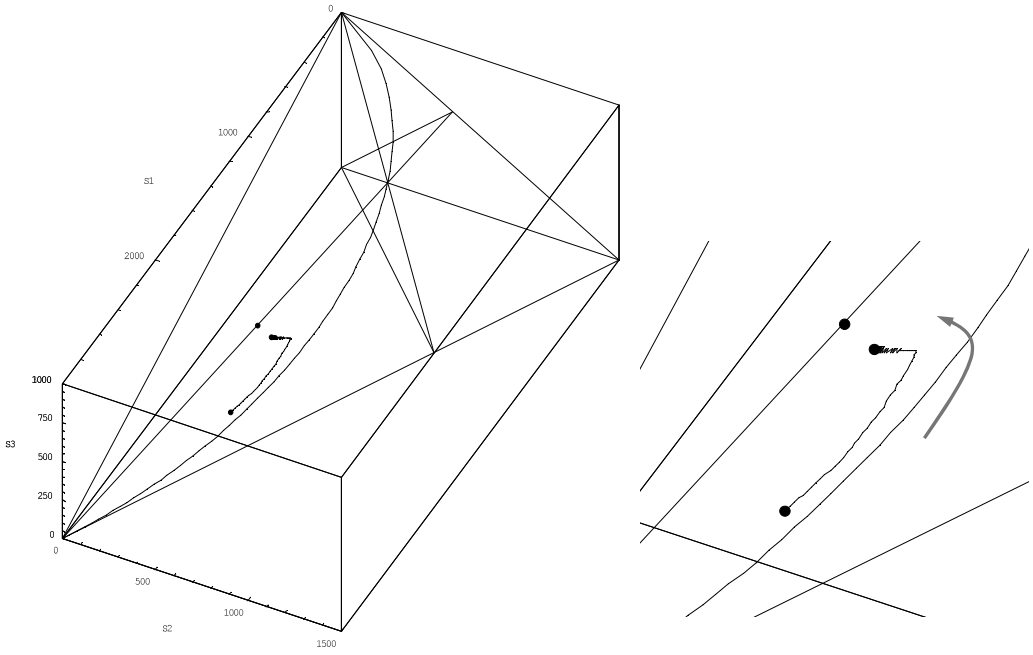

図４　ペナルティー関数を用いた推定値の探索の過程（例２）

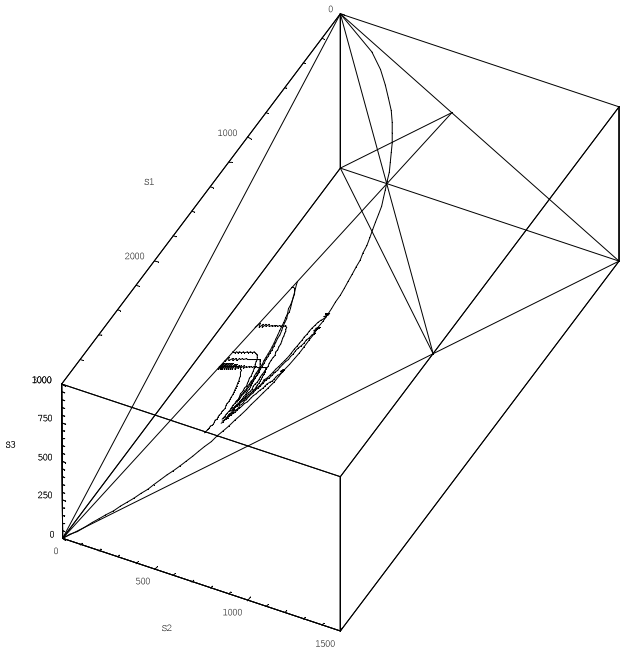

図５　ペナルティー関数を用いた推定値の探索の過程（例２’）

例を例2'と呼ぶ。それぞれの標本寸法指標を基にした母集団寸法指標の推定値と探索の過程を図5に示す。移動はすべて下から上へ向かっている。

3　ピットマンモデルを利用したノンパラメトリック推定

3.1　ピットマンモデルによる推定

2節では，制約条件を満たす領域内で最大尤度を探索する佐井（2002, 2003）の方法と，その制約条件をペナルティー関数で表現した佐井（2006）の方法を紹介した。

前者の探索では，領域内において尤度関数をより大きくする方向に推定値を移動し，後者の探索でも，領域の端に近づかない限りはペナルティー関数の影響をほとんど受けないため，前者と同様に，単に尤度関数をより大きくする方向に移動することになる。したがって，探索の初期値としてピットマンモデルによる推定値を用いているものの，ノンパラメトリック法による推定値への影響は限定的である。

Pitman (1995) によって提案されたピットマンモデルはEwens-Pitman sampling formula (EPSF) とも呼ばれる。2個のパラメータ α，θ $(0 \leq \alpha < 1, \theta > -\alpha)$ を持ち，現在，個票データのリスク評価に用いられているモデルの中では当てはまりの良いことが知られている。

ピットマンモデルについては様々な説明が可能であるが，しばしば壺を使って説明される。区別のつかない壺が十分多くあり，玉を1個ずつ壺に入れていく。最初の玉は任意の壺に入れるが，2個目以降の玉は，確率 $(\theta + u\alpha)/(\theta + n)$ で空の壺に入れ，確率 $(l - \alpha)/(\theta + n)$ で l 個入っている壺に入れる。ただし n はそれまでに入れた玉の数で，u はそれまでに玉が入っている壺の数である。

個票データのリスク評価では，壺がキー変数の値の組み合わせによって作られるセルに対応し，玉は個体に対応している。

ピットマンモデルは超母集団モデルの一つでもある。超母集団モデルを想定した場合，母集団は超母集団からサンプリングされたと考えられ，大

きさ N の母集団からサンプリングされた大きさ n の標本は，超母集団モデルから直接サンプリングされたと見なすことができる。そのため，標本寸法指標を基に超母集団モデルであるピットマンモデルのパラメータ α , θ を推定すると，

$$\mathrm{E}(S_l) = \frac{N}{l} \cdot {}_{N-1}C_{l-1} \cdot \frac{(1-\alpha)^{[l-1]}(\theta+\alpha)^{[N-l]}}{(\theta+1)^{[N-1]}} \tag{11}$$

を用いることによって，母集団寸法指標を推定することができる。ただし

$$x^{[k]} = x(x+1)\cdots(x+k-1) \tag{12}$$

である。

パラメータの推定にはモーメント法や最尤法があるが，本論文では最尤法を用いる。

3.2　ピットマンモデルの確率関数を利用したノンパラメトリック法の改善

ピットマンモデルにおける母集団寸法指標の確率関数は

$$\mathrm{P}(S_1, S_2, \ldots, S_N) = N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{l=1}^{N} \left\{ \frac{(1-\alpha)^{[l-1]}}{l!} \right\}^{S_l} \frac{1}{S_l!} \tag{13}$$

である。ただし, U は母集団において個体が入っているセル数であり，

$$x^{[k:a]} = x(x+a)\cdots\{x+(k-1)a\} \tag{14}$$

である。

佐井 (2002, 2003, 2006) で提案したノンパラメトリック最尤推定法では，母集団寸法指標の推定値を探索する際の初期値の計算にのみピットマンモデルを用いたが，この節では，探索の過程においてもピットマンモデルを用いる方法を考える。

(13) 式のピットマンモデルの確率関数をペナルティー関数の代わりに

用いて

$$\begin{aligned} & \mathrm{F}_2(S_1, S_2, \ldots, S_N \mid s_1, s_2, \ldots, s_n) \\ = \; & \log \mathrm{L}_{\mathrm{ap}}(S_1, S_2, \ldots, S_N \mid s_1, s_2, \ldots, s_n) \\ & + c \cdot \log \mathrm{P}(S_1, S_2, \ldots, S_N) \end{aligned} \tag{15}$$

を最大化する方法を提案する。ただし c はピットマンモデルの影響力を表す定数である。

(15) 式の母集団寸法指標の近似的な尤度関数 $\mathrm{L}_{\mathrm{ap}}(\cdot)$ とピットマンモデルの確率関数 $\mathrm{P}(\cdot)$ について，バランスをとりながら最大値を探索することになるが，ピットマンモデルについては，得られている標本寸法指標 $(s_1, s_2, \ldots, s_n)$ からパラメータの推定値 $\hat{\alpha}$, $\hat{\theta}$ を求め，その値を用いて探索の各ステップにおける母集団寸法指標の確率関数 (13) 式を計算することになる。

これとは別に，探索の各ステップにおける母集団寸法指標の推定値 $(\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N)$ からパラメータの推定値 $\hat{\alpha}$, $\hat{\theta}$ を求め，その値を用いてピットマンモデルの標本寸法指標の確率関数 $\mathrm{P}(s_1, s_2, \ldots, s_n)$ を $\mathrm{P}(S_1, S_2, \ldots, S_N)$ の代わりに計算する方法も考えられる。

尤度関数 $\mathrm{L}_{\mathrm{ap}}(S_1, S_2, \ldots, S_N \mid s_1, s_2, \ldots, s_n)$ は，母集団寸法指標から標本寸法指標が得られる確率であるため，それに合わせて，標本寸法指標の確率関数を用いる後者の方が妥当かもしれないが，その場合，探索の各ステップにおいてピットマンモデルのパラメータを推定する必要があり，計算時間の観点から採用を見送った。

(15) 式を最大化する母集団寸法指標の探索では，(13) 式に $S_l!$ が含まれるが，近似を行うことによって連続変量として母集団寸法指標を扱うことは可能である。母集団寸法指標を連続型と考えると，探索の際の移動距離を自由に変えられるメリットもあるが，最終的に母集団寸法指標の推定値として整数値を得ようとすると，整数化の仕方によって推定値が大きく異なってしまうディメリットもある。

本論文では非負整数に限定して母集団寸法指標を探索するが，同様の設定を行った2.1 節の探索の方法では，例えばサイズ10のセル数を１個減らす場合，サイズ１のセル数を10個増やさなければならなかった。この方法は柔軟性に欠けており，尤度関数がより大きくなるような母集団寸法指標の推定値の移動ができなくなる可能性がある。そこで，新たに次のような方法を提案する。

探索の p 番目のステップで暫定的な推定値

$$(\hat{S}_{1(p)},\ldots,\hat{S}_{l-1(p)},\hat{S}_{l(p)},\ldots,\hat{S}_{l'(p)},\hat{S}_{l'+1(p)},\ldots,\hat{S}_{N(p)}) \quad (16)$$

における (15) 式の値が得られているとき，l ，l' $(1 \leq l \leq N,\ 0 \leq l' \leq N-1,\ \hat{S}_l > 0,\ \hat{S}_{l'} > 0)$ のすべての組み合わせについて

$$(\hat{S}_{1(p)},\ldots,\hat{S}_{l-1(p)}+1,\hat{S}_{l(p)}-1,\ldots,\hat{S}_{l'(p)}-1,\hat{S}_{l'+1(p)}+1,\ldots,\hat{S}_{N(p)}) \quad (17)$$

における (15) 式の値を求め，その中で最大値をとる母集団寸法指標を p ＋１番目のステップの推定値

$$(\hat{S}_{1(p+1)},\ldots,\hat{S}_{l-1(p+1)},\hat{S}_{l(p+1)},\ldots,\hat{S}_{l'(p+1)},\hat{S}_{l'+1(p+1)},\ldots,\hat{S}_{N(p+1)}) \quad (18)$$

としながら，推定値を移動していく方法である。

これは，サイズ l の一つのセルの個体１個を，サイズ l' のセルに移動させることに等しい。図６は個体の移動のイメージで，四角一つずつがセルを，その中の点が個体を表している。この図の場合，サイズ２の一つのセルの個体１個を，サイズ４のセルに移動させており，あるステップの推定値

$$(\hat{S}_{1(p)},\hat{S}_{2(p)},\hat{S}_{3(p)},\hat{S}_{4(p)},\hat{S}_{5(p)}) = (3,2,1,1,0)$$

を

$$\begin{aligned} & (\hat{S}_{1(p+1)},\hat{S}_{2(p+1)},\hat{S}_{3(p+1)},\hat{S}_{4(p+1)},\hat{S}_{5(p+1)}) \\ = \ & (\hat{S}_{1(p)}+1,\hat{S}_{2(p)}-1,\hat{S}_{3(p)},\hat{S}_{4(p)}-1,\hat{S}_{5(p)}+1) \\ = \ & (4,1,1,0,1) \end{aligned}$$

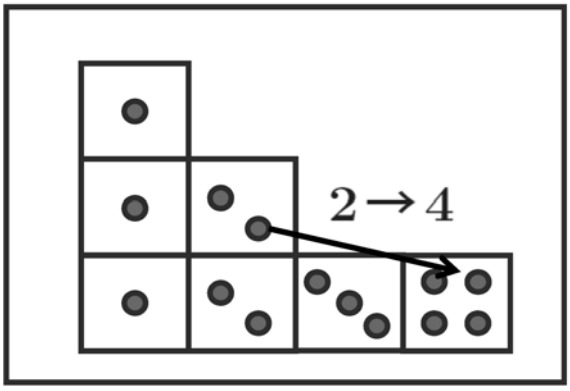

図６　推定値の探索における個体の移動の例

と変化させることになる。この移動を２→４と書くことにすると，(16) 式から (17) 式への移動は $l \to l'$ と書ける。なお，$l' = 0$ の場合，$(\hat{S}_{0(p)}, \hat{S}_{1(p)}, \dots)$ が $(\hat{S}_{0(p)} - 1, \hat{S}_{1(p)} + 1, \dots)$ と変化するが，$\hat{S}_0$ は推定の対象とせず，探索においてその値は常に十分大きいものとする。

3.3　推定値の探索法の詳細

従来の方法と同様に，探索の初期値にはピットマンモデルで得られた推定値を用いる。ピットマンモデルによる推定値は非負の実数値であるため，初期値として使用するためには整数化する必要がある。ここでは，各サイズについて得られた寸法指標を整数に切り捨て，それによる総個体数の不足分についてはサイズ１の寸法指標を増加させることにより調整する。これを，母集団寸法指標の推定値を探索する際の初期値 $(\hat{S}_{1(0)}, \hat{S}_{2(0)}, \dots, \hat{S}_{N(0)})$ とする。

推定値の探索には前節で述べた方法を用いるが，(15) 式の定数 c を適切に選ぶことは難しい。c を大きくすると，ピットマンモデルで得られた初期値からほとんど動くことなく探索は終了してしまい，c を小さくすると，ピットマンモデルの影響はほとんどなくなり，単に $\mathrm{L}_{\mathrm{ap}}(\cdot)$ を最大にすることと等しくなってしまう。

そこで，定数 c を段階的に小さくしながら，すなわちピットマンモデル

の影響を徐々に小さくしながら，(15) 式の $\mathrm{F}_2(\cdot)$ の値を増加させていく方法を用いる。$c = 1.0$として探索を始め，個体をどのように移動させても (15) 式の値が増加しなくなった場合には，c の値を0.1倍して再度探索を行う。$c = 10^{-10}$ となっても，2.1 節で挙げた四つの制約条件（a）～（d）を満たす領域で $\mathrm{F}_2(\cdot)$ の値を増加できなくなった時点で探索を終了し，それを最終的な推定値とする。

2.2 節で用いた簡単な例を再度取り上げる。

例 1 （$N = 3000$, $(S_1, S_2, S_3) = (2100, 300, 100)$, $n = 600$, $(s_1, s_2, s_3) = (548, 23, 2)$）における推定値と探索の過程を図 7 に示す。探索の両端の点は初期値と最終的な推定値で，探索途中の点は c の値を0.1倍した点である。この場合，$c = 10^{-3}$となった後の160ステップ目で，$c = 10^{-10}$となっても制約条件 $\hat{S}_2 \geq \hat{S}_3$ を満たす範囲で $\mathrm{F}_2(\cdot)$ の値を増加させることができず，探索を終了している。推定値は $(\hat{S}_1, \hat{S}_2, \hat{S}_3) = (2125, 175, 175)$ である。

また図 8 は，例 1’（$N = 3000$, $(S_1, S_2, S_3) = (2100, 300, 100)$, $n = 600$）における10組の標本寸法指標を基にした推定値と探索の過程である。ステップ数は37 ～ 241で，すべて下から上へ向かって移動している。

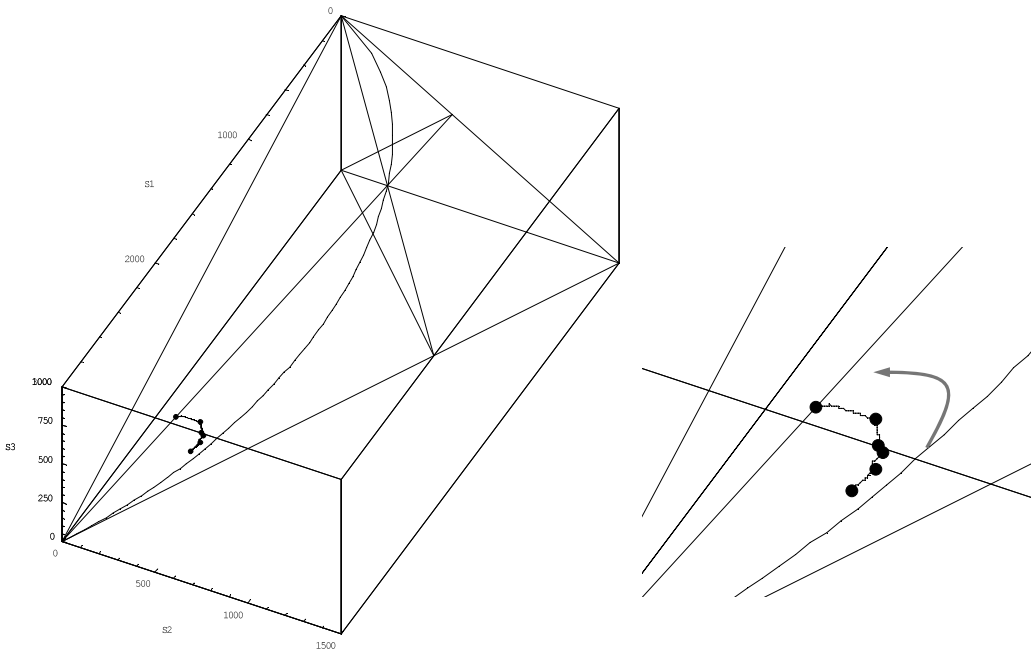

図７　提案した方法による推定値の探索の過程（例１）

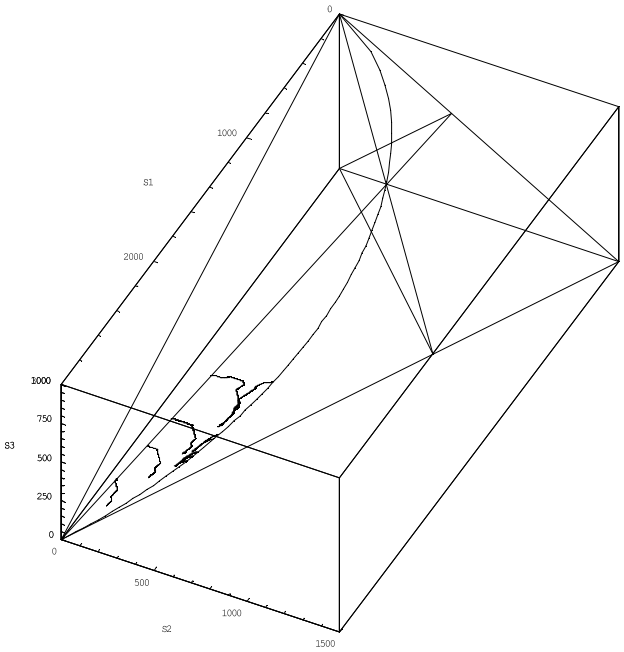

図８　提案した方法による推定値の探索の過程（例 1’）

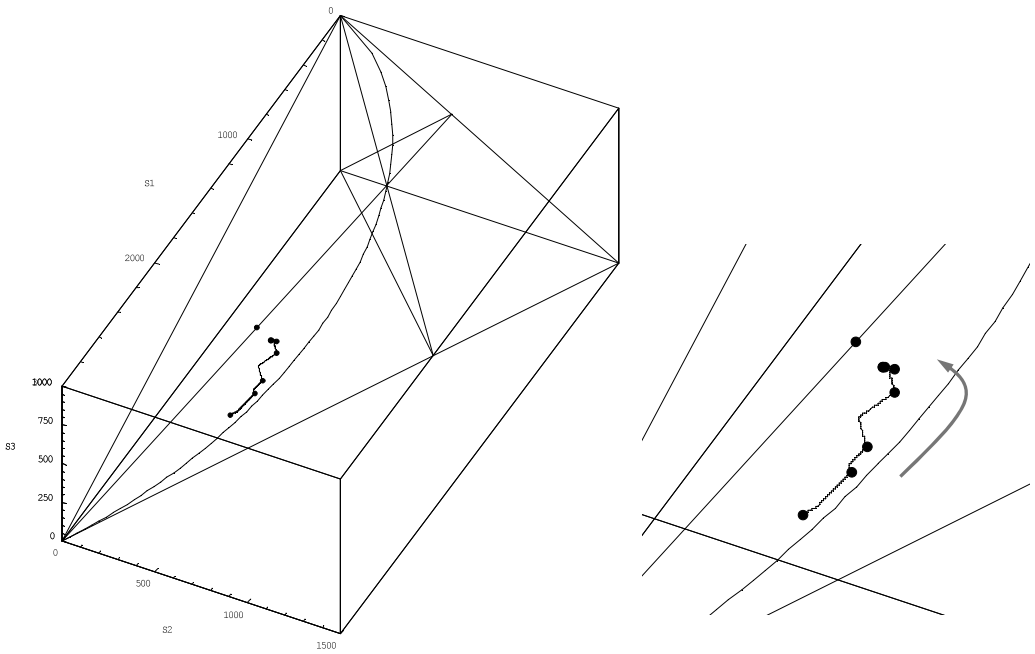

図９　提案した方法による推定値の探索の過程（例２）

図 9 には，例 2（ $N = 3000$, $(S_1, S_2, S_3) = (1500, 300, 300)$, $n = 600$, $(s_1, s_2, s_3) = (516, 39, 2)$）における推定値と探索の過程を示す。この場合，302ステップ目で $c = 10^{-5}$ となった点から $\mathrm{F}_2(\cdot)$ の値を増加させることができず，探索を終了している。推定値は $(\hat{S}_1, \hat{S}_2, \hat{S}_3) = (1500, 375, 250)$ である。

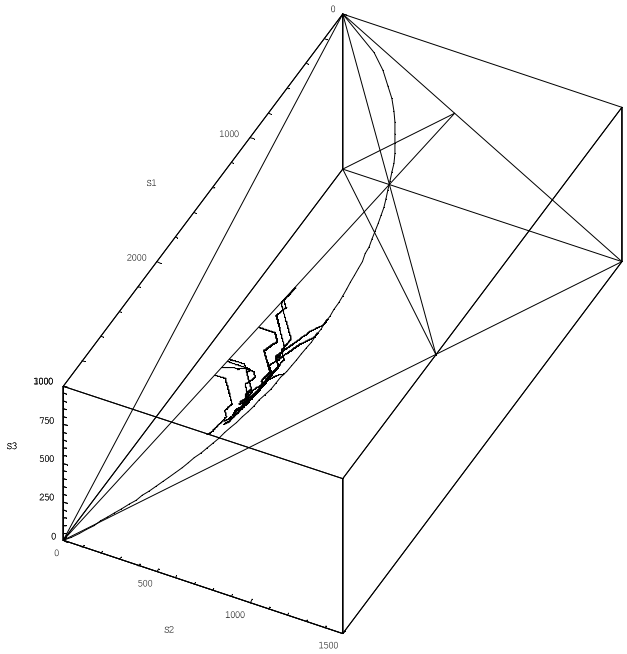

図10　提案した方法による推定値の探索の過程（例２’）

さらに図10には，例２’（$N = 3000$, $(S_1, S_2, S_3) = (1500, 300, 300)$，$n = 600$）における10組の標本寸法指標を基にした推定値と探索の過程を示す。ステップ数は224 ～ 429で，すべて下から上へ向かって移動している。

図７と図９の最終的な母集団寸法指標の推定値を，2.2 節の (9) 式を最大にする推定値と比較すると，後者は一致し，前者も極めて近いものの，探索の過程は異なっている。(9) 式では，制約条件の境界に近づかない限り，単に尤度関数をより大きくするように推定値が移動するのに対して，本論文で提案した方法では，探索の途中でもピットマンモデルの影響をある程度受けながら移動する。そのため，さらに厳しい制約を課した場合には，推定値が大きく異なることも想定される。

図８と図３，図10と図５をそれぞれ比較しても同様のことが言える。

3.4　適用例

3.3 節で用いた例は，母集団寸法指標の推定値を探索する過程を図で表現するために，最大サイズが３の小さいデータに対するものであった。

この節では実際のデータを用いて，提案した方法について更に検討を行う。

2007年12月に行われた労働力調査で得られた個票データのうち，就業者46441人分と非就業者37268人分のデータを用いる。元の個票データに含まれる34変数のうち15変数をキー変数と見なし，そのうち６変数には多少の秘匿を施す。詳細については佐井 (2012) を参照していただきたい。なお，この秘匿方法が公開に適していることを主張するものではない。

就業者の個票データから得られる標本寸法指標は

$$(s_1, s_2, s_3, \ldots, s_{12}) = (45266, 454, 50, 15, 4, 3, 1, 0, 0, 0, 0, 1)$$

となる。同じ寸法指標のサイズ10までの部分を表１のactualの行にも示す。

実験のため，この個票データから抽出率約１/５で大きさ9288のサブサンプルを非復元単純無作為抽出する。寸法指標は（9225, 27, 3）である。

表１に，標本寸法指標の推定値の探索の過程を示す。サイズ11以上については省略するが，すべてのステップにおいて推定値は０である。Pitman（1）はピットマンモデルを用いた推定値を整数値に四捨五入したもので，Pitman（2）はPitman（1）の総個体数を調整した探索の初期値を示す。サブサンプルから推定されたパラメータの値は $\hat{\alpha} = 0.97558323$, $\hat{\theta} = 2.88862512 \cdot 10^4$ である。c の値を0.1倍していく中で，各 c の値における最後のステップの推定値と通算のステップ数を表示している。ここでは $c = 10^{-3}$ とした後の73ステップ目で，$c = 10^{-10}$ となっても $\mathrm{F}_2(\cdot)$ の値を増加させることができず，探索を終了した。

表１　標本寸法指標の推定値の探索の過程（労働力調査，就業者）

c	$\hat{s}_1$	$\hat{s}_2$	$\hat{s}_3$	$\hat{s}_4$	$\hat{s}_5$	$\hat{s}_6$	$\hat{s}_7$	$\hat{s}_8$	$\hat{s}_9$	$\hat{s}_{10}$	step
Pitman(1)	45222	417	78	22	7	3	1	0	0	0	–
Pitman(2)	45235	417	78	21	7	2	1	0	0	0	0
1.0	45237	416	78	21	7	2	1	0	0	0	1
10^{-1}	45261	405	76	22	7	2	1	0	0	0	14
10^{-2}	45332	354	81	25	8	3	0	0	0	0	57
10^{-3}	45351	326	92	26	8	3	0	0	0	0	73
actual	45266	454	50	15	4	3	1	0	0	0	–

表２には非就業者の個票データから抽出率約１/５抽出したサブサンプルを基にした標本寸法指標の推定値の探索結果を示す。ここでは，$c = 10^{-4}$とした後の3200ステップ目で，$c = 10^{-10}$としても$\mathrm{F}_2(\cdot)$の値を増加させることができず，探索を終了した。なお，実際の標本寸法指標の最大サイズは175である。

また，表３と表４には，アメリカで2000年に実施されたセンサスの１％抽出個票データ（U.S. Census Bureau (2003)）を用いた結果を示す。ワシ

表２　標本寸法指標の推定値の探索の過程（労働力調査，非就業者）

c	$\hat{s}_1$	$\hat{s}_2$	$\hat{s}_3$	$\hat{s}_4$	$\hat{s}_5$	$\hat{s}_6$	$\hat{s}_7$	$\hat{s}_8$	$\hat{s}_9$	$\hat{s}_{10}$	step
Pitman(1)	6566	1471	699	420	284	206	157	123	100	83	–
Pitman(2)	10114	1470	698	419	283	206	157	123	100	82	0
1.0	7433	1694	823	507	353	264	199	151	116	91	2016
10^{-1}	7043	1612	808	528	387	285	211	158	120	93	2414
10^{-2}	6612	1481	788	573	417	305	225	168	127	97	2888
10^{-3}	6815	1194	811	590	430	315	232	172	129	98	3161
10^{-4}	6857	1134	822	596	433	316	232	172	129	98	3200
actual	6158	1603	729	427	278	198	170	146	115	82	–

表３　標本寸法指標の推定値の探索の過程（センサス，就業者）

c	$\hat{s}_1$	$\hat{s}_2$	$\hat{s}_3$	$\hat{s}_4$	$\hat{s}_5$	$\hat{s}_6$	$\hat{s}_7$	$\hat{s}_8$	$\hat{s}_9$	$\hat{s}_{10}$	step
Pitman(1)	30184	25	0	0	0	0	0	0	0	0	–
Pitman(2)	30186	24	0	0	0	0	0	0	0	0	0
1.0	30186	24	0	0	0	0	0	0	0	0	0
10^{-1}	30184	25	0	0	0	0	0	0	0	0	1
actual	30099	66	1	0	0	0	0	0	0	0	–

表４　標本寸法指標の推定値の探索の過程（センサス，非就業者）

c	$\hat{s}_1$	$\hat{s}_2$	$\hat{s}_3$	$\hat{s}_4$	$\hat{s}_5$	$\hat{s}_6$	$\hat{s}_7$	$\hat{s}_8$	$\hat{s}_9$	$\hat{s}_{10}$	step
Pitman(1)	9255	259	90	45	27	18	13	10	7	6	–
Pitman(2)	9922	259	89	45	27	18	12	9	7	5	0
1.0	9845	276	96	48	29	18	12	9	7	5	48
10^{-1}	9839	277	96	49	29	18	12	9	7	5	52
10^{-2}	9834	276	94	51	30	18	12	9	7	5	57
10^{-3}	9833	279	91	52	30	18	12	9	7	5	59
actual	9228	392	72	30	12	7	6	6	4	8	–

ントン州在住の20歳以上のうち，就業者30234人分の個票データから抽出率約1/5でサブサンプルを抽出し，その寸法指標を基にした標本寸法指標の推定値の探索結果を表3に示す。表4は非就業者11725人分の個票データに関する同様の結果である。キー変数などの詳細については佐井(2012)を参照していただきたい。

これらの例では，ピットマンモデルのデータへの当てはまりがよく，提案した方法の最後のステップにおいて得られた推定値は，初期値と比べて，実際の寸法指標から遠ざかっているものが多い。ただし，例えば表1では $\hat{s}_1$，$\hat{s}_3$ は $c=10^{-1}$ の最後のステップの推定値の方がピットマンモデルによる推定値よりも実際の値に近い。本報告では，佐井（2003）で提案した制約条件をそのまま用いているが，やや領域が広すぎることも考えられる。そのため，領域を縮小するか，例えば c の値がある程度大きいうちに探索を終了するなどの措置をとる必要があるかもしれない。

次に，表3と表4で用いたセンサスの個票データの標本寸法指標を基に，母集団寸法指標を推定した探索結果を表5，表6に示す。1％抽出データであることから，例えば就業者の母集団の大きさは，簡単に $N=30234\cdot 10^2$ としている。また，$c<10^{-2}$ となった時点で探索を止めている。

表5　母集団寸法指標の推定値の探索の過程（センサス，就業者）

c	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$	$\hat{S}_6$	$\hat{S}_7$	$\hat{S}_8$	$\hat{S}_9$	$\hat{S}_{10}$	step
Pitman(1)	2520208	113410	33754	14310	7151	3935	2307	1415	898	585	–
Pitman(2)	2520530	113410	33754	14309	7151	3934	2307	1415	897	584	0
1.0	2520464	113424	33759	14312	7152	3935	2307	1415	897	584	42
10^{-1}	2520462	113425	33759	14312	7152	3935	2307	1415	897	584	43
10^{-2}	2520381	113423	33760	14314	7154	3937	2309	1416	898	585	113

表6　母集団寸法指標の推定値の探索の過程（センサス，非就業者）

c	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$	$\hat{S}_4$	$\hat{S}_5$	$\hat{S}_6$	$\hat{S}_7$	$\hat{S}_8$	$\hat{S}_9$	$\hat{S}_{10}$	step
Pitman(1)	683412	22473	7982	4121	2526	1711	1238	939	737	594	–
Pitman(2)	867464	22473	7981	4120	2526	1711	1238	938	736	593	0
1.0	831478	27378	9737	5034	3090	2096	1519	1153	906	732	26254
10^{-1}	829718	27334	9734	5044	3105	2114	1538	1173	927	753	27868
10^{-2}	815170	26993	9737	5149	3259	2298	1744	1744	1398	1134	41094

探索のための計算時間はWindows7，CPU: Intel(R)Core(TM)i7-2860QM 2.50GHzのパーソナルコンピュータで，表５の推定値を得る113ステップに約0.5秒かかり，１ステップあたり平均で0.0043秒程度必要である。また表６の推定値を求める41094ステップには約22分19秒かかり，１ステップあたり平均で0.033秒程度必要である。

計算時間を左右する要素は様々ある。標本と母集団の大きさ，標本寸法指標と母集団寸法指標の最大サイズ，ピットマンモデルによる推定法と提案した方法による推定値の乖離などが主なものである。(15)式の $\mathrm{F}_2(\cdot)$ を，１ステップにおいて最大サイズの２乗程度の回数計算する必要があるため，母集団寸法指標の最大サイズは小さく押さえたいところであるが，小さくしすぎると推定結果へ影響があるため慎重に選ぶ必要がある。

表５，表６の検討では最大サイズをそれぞれ100，200としたが，最大サイズの推定値はそれぞれ $\hat{S}_{100}=0$，$\hat{S}_{200}=1$ であった。計算時間を短縮するために，後者の最大サイズはもう少し小さくしてもよいかもしれない。例えば最大サイズを100とすると，推定値を得る探索52413ステップに約４分９秒かかり，最大サイズを150とすると，推定値を得る探索44847ステップに約10分34秒かかる。最大サイズの推定値はそれぞれ $\hat{S}_{100}=6$，$\hat{S}_{150}=2$ となり，裾を打ち切られているが，サイズの小さい部分の推定値には大きな違いはない。

４　おわりに

本論文では，制約付きノンパラメトリック最尤推定法において，佐井(2006）で用いた滑らかなペナルティー関数の代わりに，ピットマンモデルの確率関数を利用する方法を提案した。

ペナルティー関数を用いる方法と比較すると，母集団寸法指標の推定値の探索の過程が異なり，その意味では優位性があると判断されるが，推定値を整数に限定しているために，母集団が大きい場合，あるいは母集団寸法指標の最大サイズが大きい場合には計算時間を要することがある。

また実データを用いた検討も行ったが，ピットマンモデルの当てはまりがよいデータであるため，提案した方法の有効性を十分実証できたとは言えない。

ピットマンモデルによって良い推定が可能かどうかは，標本寸法指標に対する当てはめを行った時点である程度予測することができるため，提案した方法を用いるか否かは，その時点で判断することも可能である。

また，仮にピットマンモデルによる推定がある程度良いと予想される場合にも，推定値がピットマンモデルで推定可能な領域からどちらへ出ようとしているかなどを知るためにも利用する意味は大きいと思われる。

謝　　辞

ピットマンモデルの確率関数をペナルティー関数の代わりに用いる方法について，東京大学大学院の竹村彰通先生には有益な助言をいただきました。また，慶應義塾大学名誉教授の渋谷政昭先生をはじめ，共同研究をしている諸先生方からは数々の助言をいただきました。ここに感謝いたします。

本論文は科学研究費補助金（課題番号22300097），統計数理研究所共同利用研究プログラム（23－共研－2028，24－共研－2030，25－共研－2030）の研究成果に基づくものである。

参考文献

［1］Bethlehem, J. G., Keller, W. J. and Pannekoek, J.（1990）. Disclosure control of microdata, *Journal of the American Statistical Asociation,* **85**, 38–45.
［2］Engen, S.（1978）. *Stochastic Abundance Models,* Chapman and Hall, London.
［3］Pitman, J.（1995）. Exchangeable and partially exchangeable random partitions, *Probability Theory and Related Fields,* **102**, 145–148.
［4］佐井至道（2002）．サイズインデックスの制約付き最尤推定，岡山商大論叢，**37**（3），61–79.
［5］佐井至道（2003）．母集団寸法指標のノンパラメトリック推定，統計数理，**51**（2），183–197.
［6］佐井至道（2006）．ペナルティ関数を利用した母集団寸法指標の制約付きノンパラメトリック推定，岡山商大論叢，**42**（1），1–21.
［7］佐井至道（2012）．事後層化による個票データのリスク評価の改善，岡山商大論叢，

48 (1), 1-25.
[8] U.S. Census Bureau (2003). Census 2000 Public Use Microdata Sample (microdata), Washington, D. C.