潜在変数を含む統計モデルにおける効率的なパラメータ推定

中村　永友[1]　　土屋　高宏[2]

要　旨

本報告は各データ点に対してある種の重みが潜在変数として扱われる統計モデルのパラメータの推定誤差が小さくなるような効果的なブートストラップ法の提案を行う．これは観測データから推定された統計モデル $f(\hat{\theta})$ の構造を十分反映するようなリサンプリングの方法でもある．このような統計モデルの例としては，t-分布モデル，M-推定量，有限混合正規分布モデル等があげられる．数値実験を通してその有効性を検証する．

キーワード：混合正規分布モデル，t-分布モデル，ブートストラップ，信頼区間

1　はじめに：問題の所在

データに統計モデルをあてはめた後に，ノンパラメトリック・ブートストラップ法などのリサンプリング手法によって推定したパラメータの挙動を調べることは，その統計モデルが複雑であったり，パラメータの推定量が陽に書き下すことができない場合によく使われる手法である．ここで対象となるのは，データ点へのある種の重みが潜在変数となる統計モデルで，尤度原理に基づく t-分布モデルや混合分布モデルが挙げられる．前者はデータの各点の重みが潜在変数であり，後者は各成分分布に所属する確率がそれにあたる．このようなモデルでパラメータの推定値の挙動（信頼区間など）を調べる際には，ブートストラップ法を用いることができる．

ブートストラップ法を適用する際には，ブートストラップ標本の生成，仮定するモデルのあてはめ（パラメータ推定），これらを決められた回数くり返す，という手順を一般的に踏む．このブートストラップ標本に基づく統計的推測は，通常のデータに対するものと通常は同じことを行うが，そのデータの状況によってはある種の推定が不安定になるなどの問題がある．本論文はブートストラップ法によるパラメータの推定の安定化法，すなわち分散減少の方法を提案する．これはモデル評価を安定的に行うための改良，分散安定化法として位置づけられる．

本研究の動機は，シミュレーション実験を数多く行っていると，ブートストラップ標本がオリジナル標本の構造を反映しないということがしばしば生じ，正確なパラメータ推定の評価ができないことがあったことによる．つまり，$\hat{\theta}$ をオリジナルデータから推定されたパラメータ，$\hat{\theta}^*$ をブートストラップ標本から推定されたパラメータとするとき，通常はある程度（確率分布の意味で）$\hat{\theta} \approx \hat{\theta}^*$ であるはずであるが，そうならない場合がある．それは，リサンプリングされたデータが元のデータ構造をうまく反映していないことが大きな要因である．そのような事例として，２つの有限混合分布モデルに関する簡単な例と，１つの一般的なリサンプリングの例を示す．

例１　２つの母集団からのラベルのないデータがあり，混合分布モデルによってモデル推定されている．第１母集団からのデータ数が大多数の９割以上で，第２母集団からのデータが少数であったとする．このデータに対してリサンプリングを行ったとき，最悪の状況として第２母集団からのデータが１つも抽出されない，あるいはデータが１つのみ抽出されるということが想定される．このようにして得られたリサンプリ

[1] 札幌学院大学経済学部; nagatomo@sgu.ac.jp.
[2] 城西大学理学部; takahiro@math.josai.ac.jp.

ングデータに対して，混合分布モデルをあてはめたとき，前者のリサンプリングデータでは第2母集団の情報が得られない，後者では第2母集団の分散が0で推定されるか，はずれ値として扱われて，その分布に関する情報が得られない等のことがある．

例 2　2成分の有限混合分布モデルで成分分布のラベル1と2が入れ替わって推定された場合，識別ができないという問題が生じる．この場合最初に推定されたモデルと異なるモデルとなる．

例 3　ブートストラップ法によって推定されたパラメータの推定値の分布が切断される，あるいは外れ値が存在する．つまり，推定値の分布がきれいに裾を引かず，ある一定以上の値に対して，ブートストラップ標本から正常に推定できないということがある．また外れ値の存在によって期待値や分散が偏る．この場合，目的のパラメータの推定値や信頼区間が偏ってしまう．

これらの例のように，リサンプリングによって，本来のデータ構造を反映しないデータが得られ，本来得たい情報が得られないのは大きな問題と考える．

改めて要点をまとめると次のようになる．オリジナルの標本 $\boldsymbol{X}$ からリサンプリングされたブートストラップ標本 $\boldsymbol{X}^*$ は，必ずしも $\boldsymbol{X}$ の構造を反映しないことがある．つまり，統計モデルをあてはめて $f(\hat{\boldsymbol{\theta}}|\boldsymbol{X})$ が推定され，$\boldsymbol{X}^*$ から推定された統計モデル $f(\hat{\boldsymbol{\theta}}^*|\boldsymbol{X}^*)$ で，$\hat{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}^*$ であることが前提で様々な統計的推測を行うことになるが，$\hat{\boldsymbol{\theta}}$ と $\hat{\boldsymbol{\theta}}^*$ が著しく異なるとき（例えば，$|\hat{\boldsymbol{\theta}}| \not\approx |\hat{\boldsymbol{\theta}}^*|$ や $\|\hat{\boldsymbol{\theta}}\| \neq \|\hat{\boldsymbol{\theta}}^*\|$ 等），この $\hat{\boldsymbol{\theta}}^*$ をどう処理すればよいのか？ということが問題と考える．許容外の推定値に対しての処理をするのではなく，そうならないような推定上の提案をすることが本報告の目的である．

2　潜在変数を持つモデル

潜在変数を持つ統計モデルは，前述のように t-分布モデル（平均や分散の最尤推定量）や有限混合分布モデルの他，他の典型的なものとしてはロバスト統計の M-推定量（Huber and Ronchetti, 1987, 2009）が挙げられる．本報告で対象としたモデルを表1に示す．

t-分布モデルを例として，モデルの推定手順，誤差推定の手順，提案法の手順，そしてブートストラップ法の手順を次に示す．

表 1 ：潜在変数を持つ統計モデル

Models	Weight Function $w(x)$
t-Distribution	$\dfrac{\nu+p}{\nu+x^2}$ ν：自由度，p：データの次元
Huber Type	$\begin{cases} 1 & \text{if} \quad \lvert x\rvert \le k \\ \dfrac{k}{\lvert x\rvert} & \text{if} \quad \lvert x\rvert > k \end{cases}$
Tukey Type	$\begin{cases} \left[1-\left(\dfrac{x}{c}\right)^2\right]^2 & \text{if} \quad \lvert x\rvert \le c \\ 0 & \text{if} \quad \lvert x\rvert > c \end{cases}$
“Fair” Type	$\dfrac{1}{1+\lvert x\rvert/c}$
Cauchy Type	$\dfrac{1}{1+(x/c)^2}$
Welch Type	$\exp\left[-\left(\dfrac{x}{c}\right)^2\right]$

t-分布モデルの位置パラメータ（平均）の信頼区間をブートストラップ法で推定する．位置パラメータ（平均）を μ，尺度パラメータ（擬分散）を v^2，自由度 ν とする t-分布を $t(\mu, v^2, \nu)$ で表す．これらのパラメータの最尤推定値は，EM 法を用いて以下のように求められる（Lange, Little and Taylor, 1989; 中村 他, 1993）．

$$x_i \overset{iid}{\sim} t(\mu, v^2, \nu),\ (i=1, 2, \ldots, n)$$

と仮定して，

$$\hat{\mu} = \left(\sum_{i=1}^{n} \hat{w}_i\right)^{-1} \sum_{i=1}^{n} \hat{w}_i x_i, \tag{1}$$

$$\hat{v}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i (x_i - \hat{\mu})^2. \tag{2}$$

ここで，

$$\hat{w}_i = \frac{\nu+1}{\nu+\hat{\delta}_i^2},\quad \hat{\delta}_i^2 = \frac{(x_i-\hat{\mu})^2}{v}.$$

$\hat{w}_i$ や $\hat{\delta}_i^2$ は EM 法を経て，位置パラメータや尺度パラメータと同時に推定される．ここでは自由度 ν は推定するパラメータとせず，定数として扱う．これを推定することも可能であるが，適度なデータ数がないと，推定値が非常に不安定である．このモデルの中では w_i が潜在変数となる．n はデータ数，w_i はデータの各点への重みとなり，δ_i^2 は平均からデータ点への多変量の意味でマハラノビス距離である．

EM 法でこれらのパラメータを推定するときには次

の手続きを踏む．

Step 1　μ と v^2 の初期値を与える．
Step 2　$\hat{\delta}_i^2$ と $\hat{w}_i$ を推定する．(E-Step)
Step 3　$\hat{\mu}$ と $\hat{v}^2$ を計算する（近似値の推定）．(M-Step)
Step 4　Step 2 と Step 3 を収束するまで繰り返す．

収束の基準は，目的関数としての尤度関数と推定すべきパラメータ値の直前の値との差分が一定以下となったときとする．

これを踏まえて，パラメータ値の誤差推定のための手順は以下の通りとなる．

Step 1　μ と v^2 の初期値を与える．
Step 2　$\hat{\delta}_i^2$ と $\hat{w}_i$ を推定する．(E-Step)
Step 3　$\hat{\mu}$ と $\hat{v}^2$ を計算する（近似値の推定）．(M-Step)
Step 4　Step 2 と Step 3 を収束するまで繰り返す．
Step 5　ブートストラップ標本を X^* 作る．
Step 6　μ^* と v^{2*} の初期値として，$\hat{\mu}$ と $\hat{v}^2$ を与える．
Step 7　$\hat{\delta}_i^2$ と $\hat{w}_i^*$ を推定する．(E-Step)
Step 8　$\hat{\mu}^*$ と $\hat{v}^{2*}$ を計算する．(M-Step)
Step 9　Step 7 と Step 8 を収束するまで繰り返す．
Step 10　Step 5 から Step 9 を必要な回数（ブートストラップ反復）を繰り返す．
Step 11　誤差分散を計算する．

つまりこの手順は，(1)オリジナルのデータの EM 法によるパラメータ推定，(2)ブートストラップ標本の生成と **EM 法による**パラメータの推定，(3)ブートストラップ反復を行う，(4)誤差分散の推定，となる．

このアルゴリズムで推定されたのちに，ブートストラップ法でパラメータ値の誤差推定を行う際に，$\hat{w}_i^*$ を推定せずに，$\hat{w}_i$ を代用し，さらに $\hat{\delta}_i^{2*}$ も $\hat{\delta}_i^2$ で代用することを提案する．オリジナルのデータ $X=\{x_1, \ldots, x_n\}$ からのパラメータ推定は，この一連のアルゴリズムを用いるが，通常のブートストラップ法を用いる場合も同様に，このアルゴリズムを適用することになる．

しかし，提案方法では，オリジナルのデータ X から推定された $\hat{W}=\{\hat{w}_1, \ldots, \hat{w}_n\}$ を固定してブートストラップ標本 $X^*=\{x_1^*, \ldots, x_n^*\}$ を使って $\hat{\mu}^*$ と $\hat{v}^{2*}$ を推定するのである．次の手順で推定する．

Step 1　μ と v^2 の初期値を与える．
Step 2　$\hat{\delta}_i^2$ と $\hat{w}_i$ を推定する．(E-Step)
Step 3　$\hat{\mu}$ と $\hat{v}^2$ を計算する（近似値の推定）．(M-Step)
Step 4　Step 2 と Step 3 を収束するまで繰り返す．
Step 5　ブートストラップ標本を X^* を作る．
Step 6　$\hat{\delta}_i^{2*}$ と $\hat{w}_i^*$ を，$\hat{\delta}_i^2$ と $\hat{w}_i$ の代用値を用いる．
Step 7　$\hat{\mu}^*$ と $\hat{v}^{2*}$ を計算する．
Step 8　Step 5 から Step 7 を必要な回数（ブートストラップ反復）を繰り返す．
Step 9　誤差分散を計算する．

つまり，(1)オリジナルのデータの EM 法によるパラメータ推定，(2)ブートストラップ標本の抽出と（**EM 法適用しない**）パラメータの推定，(3)ブートストラップ反復を行う，(4)誤差分散の推定，となる．この方法では，ブートストラップ標本に対する EM 法の反復計算が必要ないのである．

ブートストラップ法による推定値の経験分布の作成手続きは以下の通りとなる．

Step 1　オリジナルデータ $\boldsymbol{X}=\{x_1, x_2, \ldots, x_n\}$ に統計モデル $f(X|\theta)$ をあてはめ，$f(\boldsymbol{X}|\hat{\theta})$ が推定される．その際に統計モデル内にある潜在変数 $\boldsymbol{Z}=\{z_1, z_2, \ldots, z_n\}$ が $\hat{\boldsymbol{Z}}=\{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n\}$ として推定される．
Step 2　オリジナルデータ $\boldsymbol{X}$ からブートストラップ標本 $\boldsymbol{X}^{*(b)}$ を作成する．同時に $x_i^{*(b)}$ に対応する潜在変数 $z_i^{*(b)}$ も抽出する．これを $\boldsymbol{Z}^{*(b)}$ とおく．
Step 3　$\boldsymbol{X}^{*(b)}$ に対して統計モデル $f(\boldsymbol{X}|\theta)$ をあてはめるときに，潜在変数は推定せずに $\boldsymbol{Z}^{*(b)}$ を用いてパラメータ $\hat{\theta}^{*(b)}$ を推定する．すなわち $f(\boldsymbol{X}|\hat{\theta}^{*(b)})$ が推定される．
Step 4　手順 1 と手順 2 を B 回繰り返して（$b=1, \ldots, B$），$\hat{\theta}^*$ の経験分布を作り，適宜パラ

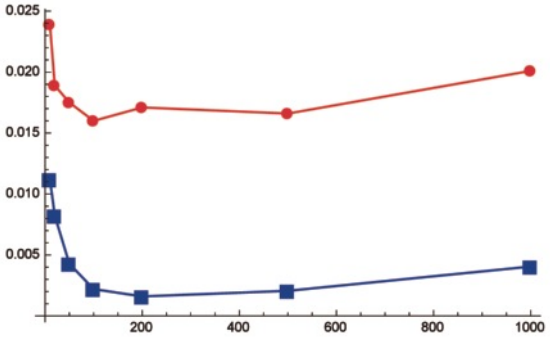

図１：ブートストラップ抽出回数による推定値の分散の変化（t-分布モデル）

●：通常の方法，■：提案手法

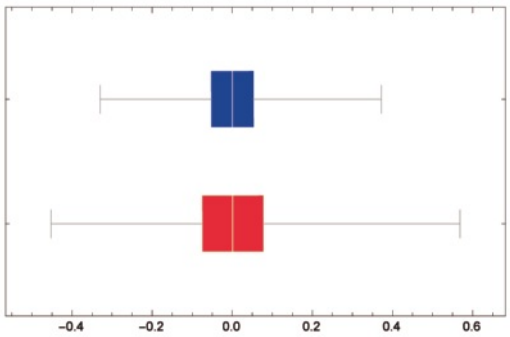

図２：提案手法とブートストラップ法による推定値の分布の比較（t(4)）

メータ θ に関する推測（区間推定等）を行う．

ここで提案した方法は次のように解釈できる．オリジナルデータにあてはめた統計モデルの構造を保持して，リサンプリングを行う．これは見方を変えれば，あてはめた統計モデルを事前情報としたサンプリングを行うようなものである．また，一種の加重ブートストラップ法（weighted bootstrap method）とも見ることができる．

3　数 値 実 験

数値実験で示す主たることは，(1)目的パラメータの期待値が一致すること，(2)推定値の分散が減少していること，(3)ブートストラップ抽出回数の減少数である．これらは，分散の値での比較，分散減少法による信頼区間との関係を見ることになる（中村 他，2014）．

また，表１に示した潜在変数を含む統計モデルに対しても同様の数値実験を行う．

3.1　ブートストラップ法の抽出回数と分散の比較

図１に自由度４の t-分布モデル $t(4)$ に対して，ブートストラップ法の回数を変えて分散の推移を示す．通常の方法に対して推定値の分散が絶対的に小さいことが示されている．また，200回程度を底としてそれ以上のリサンプリング回数では，分散が大きくなっている．通常の方法（従来法）と提案手法のブートストラップ抽出回数の比較を行う目的であったが，提案手法の方がはるかに分散が小さいため，抽出回数については従来法との比較対象とはならないようだ．

3.2　信頼区間の推定

自由度４の t-分布モデル $t(4)$ からデータ数 $n=100$ を抽出し，(1)単純にその平均を求め，これを1,000,000回繰り返して信頼区間を求める，(2)ブートストラップ抽出を100,000回繰り返して最尤法で平均を求め信頼区間を求める，(3)提案手法，の３つの95%信頼区間を求めた結果を表２に，そのときの分布の様子（ボックスプロット）を図２に示す．

表２：t(4)の平均の95%信頼区間

理論値	(−0.278，0.278)
(1)　数値的	(−0.278，0.278)
(2)　MLE＋Bootstrap	(−0.217，0.208)
(3)　提案手法	(−0.114，0.109)

ここでは具体的な数値は示さないが，平均(期待値)は有効４桁の範囲で0.0となった．また信頼区間も理論値と通常の方法((1)数値的)は一致し，MLE のモデルとブートストラップ法による推定値((2)MLE＋Bootstrap)はかなりその幅が小さくなっている．(3)提案手法はそれに比べてはるかに狭く推定された．

3.3　種々の統計モデルによる提案手法の比較

データ発生の統計モデルを自由度４の t-分布の他，表１で示したロバスト推定で用いられる種々の方法を試した．標本の大きさを $n=100$，500，ブートストラップの回数を $B=100$，20，シミュレーションの回数を $S=10,000$ とした．その結果を表３と表４に示す．

推定値の分散や標準誤差の比を見てわかるように，一部に効率の悪いモデルもあるが，t-分布モデルでは分散が半分以下になるなど，効果的なものがほとんどである．

4　お わ り に

数値実験で検証したほとんどの統計モデルで確実な

表 3 ：数値実験 1

統計モデル	提案方法		通常の方法		分散の比	標準誤差の比
	平均の推定値	推定値の分散	平均の推定値	推定値の分散		
$t(4)$	−0.0001	0.0073	−0.0002	0.0154	0.47	0.68
Huber	−0.0014	0.0144	−0.0015	0.0185	0.78	0.88
Tukey	−0.0039	0.0128	−0.0041	0.0174	0.74	0.85
Fair	−0.0032	0.0089	−0.0029	0.0130	0.68	0.83
Cauchy	−0.0009	0.0118	−0.0008	0.0167	0.70	0.84
Welsch	−0.0048	0.0012	−0.0047	0.0017	0.71	0.85

データ数：$n=100$，ブートストラップ反復回数：$B=100$，シミュレーション回数：$S=10,000$.

表 4 ：数値実験 2

統計モデル	提案方法		通常の方法		分散の比	標準誤差の比
	平均の推定値	推定値の分散	平均の推定値	推定値の分散		
$t(4)$	−0.0017	0.0014	−0.0016	0.0028	0.50	0.71
Huber	−0.0023	0.0029	−0.0024	0.0037	0.77	0.88
Tukey	−0.0014	0.0026	−0.0013	0.0035	0.72	0.85
Fair	−0.0014	0.0018	−0.0013	0.0026	0.68	0.82
Cauchy	−0.0014	0.0024	−0.0013	0.0034	0.70	0.83
Welsch	−0.0053	0.0025	−0.0055	0.0034	0.71	0.84

データ数：$n=500$，ブートストラップ反復回数：$B=20$，シミュレーション回数：$S=10,000$.

分散減少が認められた．今後の研究課題は理論的な背景を示すことで分散減少の原理，新たな分散と元の分散の関係，重みを固定して推定された推定値の性質，MLE か否か（漸近的に MLE）等が明らかになるであろう．

本研究で示したことが他の何に役に立つであろうか．例えば，混合分布モデルの成分数推定でブートストラップ法により対数尤度のバイアス推定を行う方法に対して（Nakamura and Konishi, 2016; 中村・小西, 1998），安定的な情報量規準の計算が可能となる．

参考文献

［1］ Huber, P.J. and Ronchetti, E.M. (1987). *Robust Statistical Procedures*, 2nd ed., Society for Industrial and Applied Mathematics.

［2］ Huber, P.J. and Ronchetti, E.M. (2009). *Robust Statistics*, 2nd ed., John Wiley & Sons Inc. ISBN978-0-470-12990-6.

［3］ Lange, K.L., Little, R.J.A., and Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of American Statistical Association*, **84**, 881–896.

［4］ Nakamura, N. and Konishi, S. (2016). Estimating the number of components for multivariate normal mixture models via bootstrap information criteria, preparing to submit.

［5］ 中村永友・小西貞則・大隅昇（1993）. 混合分布モデルを用いた分類法とデータ構造の色彩表示 ── LANDSAT 画像データの解析 ──, 統計数理, Vol.41, 149–167.

［6］ 中村永友・土屋高宏・小西貞則（2014）. 潜在変数を含む統計モデルにおけるブートストラップ分散減少法, 2014年度統計関連学会連合大会（日本統計学会第83回大会, 応用統計学会年次大会, 日本計量生物学会年次大会）, 東京大学, 東京, 2014.09. 14–16, 予稿集, 162.

The Proceedings of the Research Institute of Sapporo Gakuin University Vol.3, 17-22 (2016)

An Efficient Parameter Estimation for Statistical Models Associated with Latent Variables

Nagatomo NAKAMURA[1]
and
Takahiro TSUCHIYA[2]

Abstract

We propose an efficient bootstrap method for statistical models which have the latent variables as a weight for each data point. The proposed method, can also be seen as the resampling of a way to fully reflect the structure of the estimated statistical model $f(\hat{\theta})$ from the observed data. Examples of such statistical models are the t-distribution model, the M-estimator, and the finite normal mixture model, etc. The effectiveness of the proposed method is verified through numerical experiments.

Keywords: Normal Mixture Model, t-Distribution Model, Bootstrapping, Confidence Interval.

[1]Department of Economics, Sapporo Gakuiun University; nagatomo@sgu.ac.jp.
[2]Department of Mathematics, Josai University; takahiro@math.josai.ac.jp.