主成分分析における固有値・固有ベクトルの信頼区間
～漸近理論およびリサンプリング法に基づく信頼区間の性能評価～

M2007MM016 松下剛士
指導教員：田中豊

1　はじめに

主成分分析において固有値や固有ベクトルは、点推定だけが利用される事が多い。しかし、主成分分析における信頼区間・信頼領域と求める事が出来れば、主成分分析に使われている固有値固有ベクトルの信頼度を知ることができ、主成分をどこまで取り上げるか、また、主成分の解析をどこまで細かく行うかの参考にすることができる。

主成分分析における固有値・固有ベクトルの信頼区間については、多変量正規分布を仮定したときの漸近理論に基づく方法のほか、ブートストラップなどのリサンプリング法に基づく方法が知られている。本研究はそれらの方法による信頼区間の性能をシミュレーションを用いて評価することを目的とする。

実験に用いるデータは、[1] の P84 表 3 のデータから取られる平均と分散共分散行列を用いる。そこで得られた平均と分散共分散行列を母集団の平均と分散共分散行列として、mvrnorm でサンプルデータを作成し、漸近信頼区間やブートストラップ信頼区間を求める。その信頼区間を用いて被覆率を出して、性能評価を行う。

漸近信頼区間には [5] を、ブートストラップ信頼区間には [2][3][6]、t 法のジャックナイフ分散推定量は [4] を参考にした。

2　漸近信頼区間

この節における $\boldsymbol{l}$ は多変量正規分布に従う母集団から抽出されたサンプルの分散共分散行列 $\boldsymbol{S}$ の固有値であり、$\boldsymbol{a}$ は $\boldsymbol{S}$ の固有ベクトルである。また、λ_k と $\boldsymbol{\alpha}_k$ はそれぞれ、$\lambda_k = E(l_k)$ と $\boldsymbol{\alpha}_k = E(\boldsymbol{a}_k)$ であり、z_α は、標準正規分布の上側 $100\alpha\%$ 点である。

2.1　漸近理論による固有値の信頼区間

固有値の漸近分散共分散は、

$$cov(l_k, l_{k'}) = \begin{cases} \dfrac{2\lambda_k^2}{n-1} & k = k' \\ 0 & k \neq k' \end{cases}$$

であり、l_k に対する周辺分布は、

$$l_k \sim N(\lambda_k, \frac{2\lambda_k^2}{n-1}) \tag{1}$$

と近似することができる。これより、

$$\frac{l_k - \lambda_k}{\lambda_k[2/(n-1)]^{1/2}} \sim N(0,1)$$

を導く事が出来る。$\tau^2 = 2/(n-1)$ とし、ここから λ_k に対する信頼係数 $(1-2\alpha)$ の信頼区間を導くと、

$$\frac{l_k}{1+\tau z_\alpha} \leq \lambda_k \leq \frac{l_k}{1-\tau z_\alpha} \tag{2}$$

と表わされる。この信頼区間を引き出す時に、n が十分に大きいと仮定し、$\tau z_\alpha \leq 1$ となりたつものとする。

別の方法で近似信頼区間を出すには、$\log l_k$ の分布を用いる方法がある。式 (1) に対してデルタ法を適用すると、

$$\log l_k \sim N(\log \lambda_k, \frac{2}{n-1}) \tag{3}$$

となる。これによって、分散が未知のパラメータ λ_k に依存する性質を取り除くことになる。信頼係数 $1-\alpha$ を持つ $\log \lambda_k$ のための近似信頼限界は、$\log l_k \pm \tau z_\alpha$ となり、λ_k への変換によって、

$$l_k e^{-\tau z_\alpha} \leq \lambda_k \leq l_k e^{\tau z_\alpha} \tag{4}$$

という近似信頼区間を得る事ができる。

l_k は、漸近的に独立である。よって、λ_k のいくつかに対する同時信頼領域は、全体的に一定の信頼水準を得るために個々の信頼係数を選び、式 (2) や式 (4) の様な信頼区間を、単純に組み合わせる事によって得る事ができる。

2.2　漸近理論による固有ベクトルの信頼区間

α_{kj} のそれぞれの近似信頼区間は、漸近分散共分散が

$$cov(a_{kh}, a_{k'j'}) = \begin{cases} \dfrac{\lambda_k}{n-1} \displaystyle\sum_{l=1, l\neq k}^{p} \frac{\lambda_l \alpha_{lj} \alpha_{lj'}}{(\lambda_l - \lambda_k)^2} & k = k' \\ -\dfrac{\lambda_k \lambda_{k'} \alpha_{kj} \alpha_{k'j'}}{(n-1)(\lambda_k - \lambda_{k'})^2} & k \neq k' \end{cases}$$

となることを利用して a_{kj} の周辺分布から作られる。区間は λ_k の時と同様の方法で作られるが、個々の a_{kj} は互いに独立でないため、いくつかの α_{kj} のために同時信頼領域を見つける必要があり、式はより難くなる。$\boldsymbol{a}_k$ の周辺分布は、

$$\boldsymbol{a}_k \sim N(\boldsymbol{\alpha}_k, \boldsymbol{T}_k)$$

と近似することができる。ここで、

$$\boldsymbol{T}_k = \frac{\lambda_k}{n-1} \sum_{l=1, l\neq k}^{p} \frac{\lambda_l}{(\lambda_l - \lambda_k)^2} \boldsymbol{\alpha}_l \boldsymbol{\alpha}_l^{'}$$

である。行列 $\boldsymbol{T}_k$ は、固有ベクトル $\boldsymbol{a}_k$ に対応する 1 つのゼロ固有値を持つので、ランクは $(p-1)$ である。最終的には一般逆行列を用いて、近似的に

$$(n-1)\boldsymbol{\alpha}_k^{'}(l_k \boldsymbol{S}^{-1} + l_k^{-1}\boldsymbol{S} - 2\boldsymbol{I}_p)\boldsymbol{\alpha}_k \leq \chi^2_{(p-1);\alpha} \tag{5}$$

となる。ただし、今回は母集団の分散共分散がわかるため、$a_{ki} \sim N(\alpha_{ki}, \boldsymbol{T}_{k(i,i)})$ と式 (2) より

$$a_{ki} - \sqrt{T_{k(i,i)}} z_\alpha \leq \alpha_{ki} \leq a_{ki} + \sqrt{T_{k(i,i)}} z_\alpha \qquad (6)$$

を用いる。$T_{k(i,i)}$ は $\boldsymbol{a}_k$ からなる分散共分散行列 $\boldsymbol{T}_k$ の i 番目の対角要素である。

3 ブートストラップ信頼区間

この節では信頼区間を構成するためのブートストラップ法について述べる。ブートストラップ法を用いた信頼区間のもっとも簡単な方法は、パーセンタイル法である。次にブートストラップ t 法があるが、この方法では推定量の分散の推定が必要になる。分散があまり信頼できない場合には、この方法を用いるのに注意を要する。最後の方法は、BCa 法と呼ばれるもので、パーセンタイル法を改良した方法で、推定量の偏りとその分布の歪みを同時に補正する。

以下のブートストラップ信頼区間はノンパラメトリック・ブートストラップ法のもとで得られる信頼区間である。

3.1 信頼区間について

信頼区間の正確度について参考文献 [3] より、パーセンタイル法の正確度は、

$$Pr\{\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma} \leq t\} = \Phi(t) + O(n^{-1/2}) \qquad (7)$$

で表わされ、t 法や BCa 法の正確度は、

$$Pr\{\theta \in I_L\} - (1 - \alpha) = O(n^{-1}) \qquad (8)$$

と表わされている。

式 (7) では、1 次の正確度を持っているといい、式 (8) では、2 次の正確度を持っているという。

3.2 パーセンタイル法

パーセントタイル法による信頼区間は、ブートストラップサンプルを小さい順に並び変えて、求めたい点にある値を信頼限界とするものである。実際には、任意の α に対し、ブートストラップ反復回数を B とすると、ブートストラップサンプルの $B\alpha$ 番目の点を求めればいい。

3.2.1 アルゴリズム

(i). 図 1 の様に、もとの標本 $y_1, \ldots, y_n$ から重複を許して無作為に大きさ n の標本 $y_1^*, \ldots, y_n^*$ を抽出をする事で作成する。

(ii). ブートストラップ推定量 $\hat{\theta}^* = \theta(Y_1^*, \ldots, Y_n^*)$ の値の計算を B 回繰り返し、$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ を求める。

(iii). $\hat{\theta}_\alpha = \hat{\theta}^*_{(B\alpha)}$ とする。$B\alpha$ は自然数とし、$\hat{\theta}^*_{(B\alpha)}$ は $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ に対する $B\alpha$ 番目の順序統計量とする。このとき $(\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha})$ は、θ に対する名目上の被覆率 $1 - 2\alpha$ をもつ両側信頼区間とする。

3.3 ブートストラップ t 法

上で述べたパーセンタイル法は、ブートストラップ信頼区間を構成するために用いる推定量 $\hat{\theta}$ の分散推定値を求める必要がないので簡単な方法だが、信頼区間の被覆率を見ると満足できるものだというわけではない。一般的には本節で述べるブートストラップ t 法のほうが精度がいいと言われている。この節では $\hat{\theta}$ の標準誤差に対する推定値 $\hat{\sigma}/\sqrt{n}$ が得られるとする。式 (7) より、スチューデント化された量 $T = \sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma}$ の分布に対しての正規近似は、1 次の正確度しか持っていないが、ブートストラップ標本に基づく T^* の分布は、T の分布に対し 2 次の正確度を持っている。よって

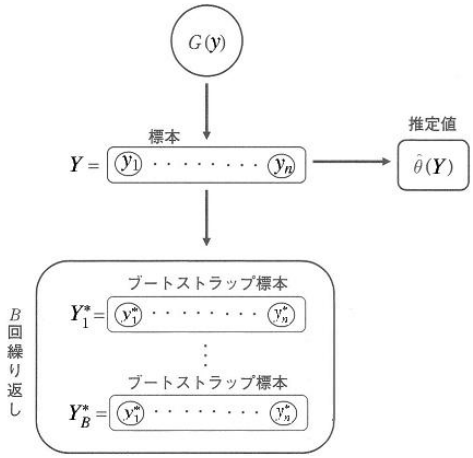

図 1 ブートストラップ標本の作成

$$(\hat{\theta} - n^{-1/2}\hat{\sigma} w_{1-\alpha}, \hat{\theta} - n^{-1/2}\hat{\sigma} w_\alpha)$$

は 2 次の正確度を持つ信頼度 $1 - 2\alpha$ の信頼区間になる。

3.3.1 ジャックナイフ分散推定

上の t 法で固有値と固有ベクトルの分散を求める時、通常の方法では分散を求める事が出来ない。そこで、ここでは [4] にあるジャックナイフ分散推定量を用いる。

$$\tilde{\theta}_{n,i} = n\hat{\theta}_n - (n-1)\hat{\theta}_{n-1}^{(i)}, \quad i = 1, \ldots, n$$

をジャックナイフ疑似量と言う。$\hat{\theta}_{n-1}^{(i)}$ は、元のデータから i 番目のデータを抜いて計算した $\hat{\theta}$ である。疑似量 $\tilde{\theta}_{n,i}, i = 1, \ldots, n$ は互いに独立で同じ分布に従う確率変数のように扱える、$\tilde{\theta}_{n,i}$ の分散は $\sqrt{n}\hat{\theta}_n$ の分散と近似的に等しい、という事が正しいとして提案された分散 $V(\sqrt{n}\hat{\theta}_n)$ の推定量

$$\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{\theta}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\tilde{\theta}_{n,j})^2$$

より、$V(\hat{\theta}_n) = V(\sqrt{n}\hat{\theta}_n)/n$ の推定量は

$$V_{JACK} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\theta}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\tilde{\theta}_{n,j})^2 \qquad (9)$$

$$= \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{n-1}^{(i)} - \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{n-1}^{(j)})^2 \qquad (10)$$

3.3.2 アルゴリズム

(i). 図 1 の様に、もとの標本 $y_1, \ldots, y_n$ から重複を許して無作為に大きさ n の標本 $y_1^*, \ldots, y_n^*$ を抽出をする事で作成する。

(ii). (i) で得られたブートストラップ標本を使い、$\hat{\theta}$ と $\hat{\sigma}$ に対応するブートストラップ推定値 $\hat{\theta}^*$、$\hat{\sigma}^*$ を計算する。

(iii). ブートストラップ t 値　$T^* = \sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ を計算し、B 回繰り返し、$T_1^*, \ldots, T_B^*$ を求める。

(iv). $w_\alpha = T^*_{(B\alpha)}$ とする。$B\alpha$ は自然数とし、$T^*_{(B\alpha)}$ はパーセンタイル法の時と同様の順序統計量とする。このとき $(\hat{\theta} - n^{-1/2}\hat{\sigma}w_{1-\alpha}, \hat{\theta} - n^{-1/2}\hat{\sigma}w_\alpha)$ は、名目上の被覆率 $1-2\alpha$ をもつ両側信頼区間である。

3.4 BCa 法

推定量 $\hat{\theta}$ は、偏りを持っているかもしれないし、$\hat{\theta}$ の分布は歪みを持っているかもしれない。パーセントタイル法では $\hat{\theta}$ に対するブートストラップ推定量 $\hat{\theta}^*$ の分布を修正せずにパーセント点を求めているため、推定量の偏りや分布の歪みの影響を受けてしまう。そのため、偏りや分布の歪みの可能性がある時は、パーセントタイル法で信頼区間を構成するのはあまり良くない。

BCa 法は、推定量の偏りとその分布の歪みを同時に修正する方法である。この方法を適用するためには、偏り修正量 z_0 と加速定数と呼ばれる量 a を推定しなければならない。

$\hat{\theta}^*$ をパーセンタイル法で定義した量とする。この時、元のデータから計算される $\hat{\theta}$ の値は、$\hat{\theta}^*$ の分布の中央に位置するものもあれば、そうでないものもある。この食い違いを表す量 $\hat{\theta}^*$ の分布の中心と $\hat{\theta}$ との偏差 $z_0 = \Phi^{-1}(Pr[\hat{\theta}^* \leq \hat{\theta}])$ は、推定値 $\hat{\theta}$ の θ に対する偏りの程度を表す尺度と考えられる。Φ は標準正規分布の分布関数である。ブートストラップ推定値 $\hat{\theta}^*$ のちょうど半分が $\hat{\theta}$ より小さい時、$Pr[\hat{\theta}^* \leq \hat{\theta}] = 0.5$ となり、$z_0 = 0$ となる。z_0 は、ブートストラップ推定値 $\hat{\theta}^* = \theta(Y_1^*, \ldots, Y_n^*)$ を用いて

$$\hat{z}_0 = \Phi^{-1}(\#\{\hat{\theta}_b^* \leq \hat{\theta}\}/B) \tag{11}$$

と表わされる。

この偏り修正量の推定値 $\hat{z}_0$ が信用できるものであるためには、式 (11) における B をパーセンタイル法の B より、倍程度大きくとる必要がある。

加速定数 a は、ブートストラップ標本に基づいて推定を行える。しかし、計算量を減らすために、正則な推定量 $\hat{\theta}$ に対して次のような処理を行う。

$\hat{\theta}_{(i)}$ を、元のデータから i 番目のデータ y_i を取り除いたものから計算された $\hat{\theta}$ の値とする。ここで、$\hat{\theta}_{(\cdot)} = n^{-1}\sum_{i=1}^n \hat{\theta}_{(i)}$ と定義すると、加速定数 $\hat{a}$ は

$$\hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}} \tag{12}$$

となる。

3.4.1 アルゴリズム

(i). 図 1 の様に、もとの標本 $y_1, \ldots, y_n$ から重複を許して無作為に大きさ n の標本 $y_1^*, \ldots, y_n^*$ を抽出をする事で作成する。

(ii). ブートストラップ推定量 $\hat{\theta}^* = \theta(Y_1^*, \ldots, Y_n^*)$ の値の計算を B 回繰り返し、$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ を求める。

(iii). 偏り修正推定量 $\hat{z}_0$ と、加速定数推定量 $\hat{a}$ を、式 (11)、(12) から計算する。

(iv). $(\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha}) = (\hat{\theta}^*_{(B\hat{\alpha})}, \hat{\theta}^*_{(B\widehat{1-\alpha})})$ を θ に対する信頼度 $1-2\alpha$ の両側信頼区間とする。ここで、任意の $\alpha (0 \leq \alpha \leq 1)$ に対して、$\hat{\alpha}$ は次のように求める。

$$\hat{\alpha} = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_\alpha}{1 - \hat{a}(\hat{z}_0 + z_\alpha)}) \tag{13}$$

$\widehat{1-\alpha}$ は、式 (13) の z_α を $z_{1-\alpha}$ で置き換えたものである。

4 データについて

[1] の P84 表 3 のデータから取られる平均と分散共分散行列を用いる。

[1] から得られた平均と分散共分散行列を母集団の分散共分散行列と平均として、mvrnorm でサンプルデータを作成し、各信頼区間を求める。サンプルサイズは、50、200、500 の 3 種とする。また、平均と分散共分散行列、固有値と固有ベクトルは下の表のものである。

平均			
149	38.7	72.23333	79.36667
分散共分散行列			
53.51724	40.79310	27.58621	28.75862
40.79310	41.73448	29.83103	24.35517
27.58621	29.83103	26.52989	17.22184
28.75862	24.35517	17.22184	18.24023

表 1　母集団の平均と分散共分散行列

固有値			
124.814224	10.848728	2.495491	1.863396
固有ベクトル			
-0.6240226	0.6455638	0.22363789	-0.37924835
-0.5591912	-0.3456392	-0.74576837	-0.10801960
-0.4083340	-0.6604700	0.62447014	-0.08414179
-0.3621661	0.1660133	0.06206998	0.91510798

表 2　母集団の固有値と固有ベクトル

5 シミュレーションについて

シミュレーションによる性能の評価基準として、被覆率を使う。

5.1 被覆率について

図 2 の縦軸は頻度を、横軸は第 1 固有値、(1,1) 固有ベクトルの大きさを示している。分布の両側にある縦線は両側信頼限界である。母集団の第 1 固有値は 124.814224 であり、母集団の (1,1) 固有ベクトル −0.6240226 である。母集団の固有値、固有ベクトルが両側信頼限界の間に入る時、それぞれ被覆回数を+1 する。被覆率は、図 2 の様なことを 1000 回繰り返して、母集団の値が含まれる割合を調べるものである。

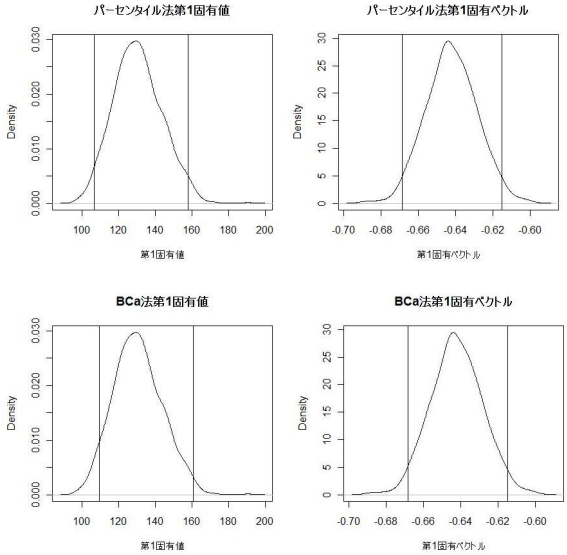

図 2 両側信頼区間と第 1 固有値、(1,1) 固有ベクトル

	下側信頼限界	上側信頼限界
漸近固有値	112.6412	157.1251
漸近対数固有値	111.2681	154.7389
パーセンタイル法	103.6404	158.4347
t 法	110.3003	174.1243
BCa 法	106.2877	160.8743

表 3 第 1 固有値両側信頼限界

表 3 は、各手法の第 1 固有値の両側信頼点の 1 例を示している。サンプルサイズ n は 200 としている。また、信頼区間は 95% 、パーセンタイル法、t 法、BCa 法は、ブートストラップ反復回数を 1000 回としている。

5.2 プログラムについて

ブートストラップ法と漸近法の固有ベクトルは向きの制限をかけるため、母集団の固有ベクトルとサンプルの固有ベクトルの内積が正になるように取っている。

t 法ではジャックナイフ分散推定をブートストラップ反復回数分行っており、被覆率を出すのに非常に時間がかかるため、t 法のシミュレーションを行わない事にした。

6 各手法の被覆率

この章の各表は被覆率を計算する回数を 1000 回、サンプルサイズ n=200、ブートストラップ反復回数 1000 回の被覆率の結果である。

固有値				
漸近法	.945	.964	.969	.949
対数を用いた漸近法	.943	.956	.973	.940

表 4 漸近法固有値

漸近法固有ベクトル			
.946	.959	.919	.967
.952	.941	.820	.915
.954	.949	.857	.913
.960	.954	.908	.981

表 5 漸近法固有ベクトル

固有値			
.957	.938	.956	.877
固有ベクトル			
.941	.947	.948	.953
.944	.939	.905	.944
.934	.949	.917	.945
.935	.949	.948	.827

表 6 パーセンタイル法

固有値			
.952	.940	.912	.903
固有ベクトル			
.940	.946	.957	.939
.943	.938	.885	.949
.936	.947	.938	.937
.934	.946	.930	.858

表 7 BCa 法

7 考察

シミュレーションの結果から、サンプルが正規分布の時、固有値は、n が大きいと全ての手法の差異が小さくなり、固有ベクトルは、特定の要素を除いて、n の大きさにかかわらず全ての手法が似たような結果を出すという事が分かった。

8 おわりに

今後の課題としては、サンプルデータを正規分布以外のもので行うことが考えられる。これらを行う事によって、BCa 法と t 法、パーセンタイル法の性能の比較を行い、計算時間を考慮した利便性と正確性の比較も考えられる。

参考文献

[1] 田中豊、脇本和昌：多変量統計解析法, 現代数学社 (1983).

[2] 北川源四郎、竹村彰通：21 世紀の統計科学 3:数理・計算の統計科学, 東京大学出版会 (2008).

[3] 伊庭幸人、汪金芳、田栗正章、手塚集、樺島祥介、上田修功：統計科学のフロンティア 11-計算統計 , 岩波書店 (2003).

[4] 前園宜彦：統計的推測の漸近理論, 九州大学出版会 (2001).

[5] I.T. Jolliffe：Principal Component Analysis (Springer Series in Statistics) , Springer(2002).

[6] Efron.B,Tibshirani,R.J.：An Introduction to the Bootstrap. Chapman & Hall,New York(1993).