1　T.W.Anderson(2003) で与えられる漸近的な受容域の適用

1.1　扱うデータの仮定と構成した漸近的な受容域の確認

以下は T.W.Anderson(2003)[*1]の **Theorem 13.5.1.**(pp.545-547) と同様の定義や仮定と，その仮定の下 SUBSECTION 11.6.1.(pp.473-474) で導出された母固有値 λ_i の有意水準 α の下での漸近的な受容域である．

母集団のデータベクトルは p 次元正規分布 $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ に独立同分布に従うとして，そこから抽出される標本 $\boldsymbol{x}_1,...,\boldsymbol{x}_N$ の標本共分散行列は $\boldsymbol{S}=\frac{1}{n}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})'$ である．ただし，$n=N-1$ である．

母共分散行列 $\boldsymbol{\Sigma}$ の固有値を $\lambda_1>\cdots>\lambda_p>0$ とする．

標本共分散行列 $\boldsymbol{S}$ の固有値を $l_1>\cdots>l_p>0$ とする．[*2]

母固有値 λ_i についての有意水準 α の仮説検定において，帰無仮説 $\lambda_i=\lambda_i^0$ を考えるとき，検定統計量を $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ と作る．

この検定統計量は漸近的に標準正規分布 $N(0,1)$ に従う．

$\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ は漸近的に $N(0,1)$ に従うため，帰無仮説 $H:\lambda_i=\lambda_i^0$ における有意水準 α の両側検定の（漸近的な）受容域は

(4)　$$-z(\alpha)\le\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\le z(\alpha),$$

である．ただし，$z(\alpha)$ は $N(0,1)$ の上側 $100(\alpha/2)\%$ 点[*3]，すなわち $P_Z^H\{Z>z(\alpha)\}=\alpha/2$ を満たす点である（ただし $Z\sim N(0,1)$）．

区間 **(4)** は，信頼係数 $1-\alpha$ を持つ λ_i の以下の信頼区間[*4]を与える．

(5)　$$\frac{l_i}{1+\sqrt{2/n}z(\alpha)}\le\lambda_i\le\frac{l_i}{1-\sqrt{2/n}z(\alpha)}.$$

実現値 l_i について，区間 **(5)** は λ_i の $100(1-\alpha)\%$ 信頼区間である．

1.2　適用する目的と方法

T.W.Anderson で与えられた帰無仮説 $H:\lambda_i=\lambda_i^0$ における有意水準 α の両側検定の（漸近的な）受容域は

(4)　$$-z(\alpha)\le\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\le z(\alpha),$$

であった．

ここでは，主成分分析において第何番目までの主成分を用いたのか根拠を持って説明することが目的である．

例として library(kernlab) にある spam データ（データ数 4601，58 次元）や library(scar) にある decathlon データ（データ数 614，10 次元），iris データ（データ数 150，5 次元）を標本として用いる．[*5]

方法は，第 i 主成分に対応する未知な母固有値 λ_i に対して，帰無仮説 $H:\lambda_i=0$ という帰無仮説を設定し，標本であるデータセットから標本固有値 l_i を導出し[*6]，これらを用いて受容域

(4)　$$-z(\alpha)\le\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\le z(\alpha),$$

[*1] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis Third Edition.* Wiley, New York.

[*2] 固有値については，テキストでは「>0」，つまり共分散行列の正定値性を明示的に仮定していないが，後の説明のためそれを仮定する．

[*3] 一般に，確率変数 X の確率密度関数 f_X を持つ分布について，$0<\alpha<1$ に対して $\int_z^\infty f_X(x)dx=\alpha$ となる z を上側 $100\alpha\%$ 点という．

[*4] 漸近分布により区間が求められるから，正確には近似的な信頼区間である．

[*5] ちなみに，spam データはベイズ統計や機械学習の文脈でよく見られる．どのようなメールがスパムなのか，メールによって学習するため．

[*6] 漸近的な受容域を十分に構成出来るだけの標本数かはわからないが，検定統計量を導出してみる．

による有意水準 α の両側検定を行う．

ただし，コードでは各標本固有値 l_i における信頼区間が各母固有値の帰無仮説 $H : \lambda_i^0 = 0$ を含むのか図示するため，信頼係数 $1-\alpha$ を持つ λ_i の以下の信頼区間を用いる．

(5) $\frac{l_i}{1+\sqrt{2/nz(\alpha)}} \leq \lambda_i \leq \frac{l_i}{1-\sqrt{2/nz(\alpha)}}$.

$i = 1,2,\ldots$ に対して，この仮説検定を帰無仮説が棄却されなくなるまで行っていくことで，帰無仮説が棄却されるような固有値に対応する主成分が，有意水準 α の下では有意に 0 でない固有値を持つ，つまり元のデータの分散を説明するに有用な主成分であると主張できる[*7]．

例えば，spam データにおいて 57 個[*8]の主成分を構成出来るが，この仮説検定によって帰無仮説 $\lambda_5 = 0$ が棄却出来ないのであれば，設定している有意水準 α の下では第 1，第 2，第 3，第 4 主成分が主成分分析として利用するのに有効で，第 5 主成分以降は用いても意味がないことがわかる．

1.3 T.W.Anderson で与えられる検定統計量の仮説検定への適用における注意点と扱うデータセットに対する確認

T.W.Anderson で与えられる検定統計量は母集団に正規性を仮定している．従って，上記の例としてあげた spam データ（データ数 4601，58 次元）のような正規分布に従っていない母集団に対してはそのまま適用することは出来ないはずである．

ちなみに，spam データ（の各変数）が正規分布に従っていないことは shapiro.test() 関数[*9]やヒストグラム（図 1），qq プロット（qqnorm() 関数と qqline() 関数）（図 2）から確認出来る．

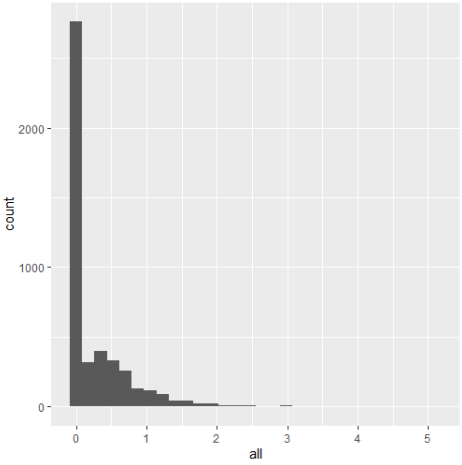

図 1　spam データの all 変数のヒストグラム，正規分布のようであれば良い

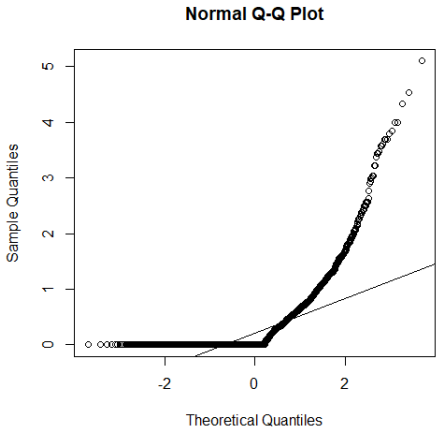

図 2　spam データの all 変数の qq プロット，直線に近ければ良い

図 1 は正規分布のようには見えないし，図 2 は直線に並んでいるようには見えない．

しかし，この母集団に対する正規性が満たされていなかったとしても，上手く検定統計量が適用できるのかどうかもこ

[*7] ただし，「棄却されない」ことが意味することは「（材料が不十分で）棄却されない」ということであり，そのことを主張として使うことは本来注意が必要である．ここではその結果を一つの指針として適用してみる程度で考えれば良い．

[*8] 58 次元目はラベルであるから除く．

[*9] spam データセットの all 変数は shapiro.test() 関数により，p-value<2.2e-16 であり，正規分布していないと考えられる．（「正規分布している」という帰無仮説が立てられている．）

の検証では確認したい．図 3 以降がその確認である．ただし，これらのコードは前回の資料において母集団を spam データのデータセットとした場合である．図 3 は検定統計量の値でヒストグラムを描画し，図 4 は spam データセットの第 1 主成分に対応する固有値を母固有値として，データセットから標本として取る個数を増やしていき，それぞれの標本固有値で信頼区間を構成し，満たしたい有意水準 A/100 に対して 100(1 − A/100)% 信頼区間を最初に構成出来た標本数を確認しているコード（リスト 1）の出力結果である．

リスト 1　構成した漸近的な信頼区間 **(5)** $\frac{l_i}{1+\sqrt{2/nz(\alpha)}} \leq \lambda_i \leq \frac{l_i}{1-\sqrt{2/nz(\alpha)}}$ の確認を行う関数のプログラムコードとその実行

```
Anderson <- function(alpha, SAMPLE_NUM, SAMPLE_EV, POP_EV){

  level <- 1 - (alpha / 2) / 100
  z <- qnorm(level, 0, 1)
  T_Anderson <- abs(sqrt((SAMPLE_NUM - 1) / 2) * (SAMPLE_EV - POP_EV) / POP_EV)

 return(if(T_Anderson <= z) {1} else {0})

}

ConfiPlot <- function(alpha, Tri, Max, df){

  POP_NUM <- Max
  SAMPLE_NUM <- seq(5, POP_NUM, by = 5)
  Percents <- numeric(length(SAMPLE_NUM))

  POP <- as.matrix(df)
  POP_EV <- (prcomp(POP)$sdev[1]) ^ 2

  decide <- 0

  for(i in 1:length(SAMPLE_NUM)){

    Percent <- 0

    for(j in 1:(100*Tri)){
    SAMPLE <- POP[sample(1:nrow(df), SAMPLE_NUM[i]), ]
    SAMPLE_EV <- (prcomp(SAMPLE)$sdev[1]) ^ 2
    Percent <- Percent + Anderson(alpha, SAMPLE_NUM[i], SAMPLE_EV, POP_EV)
    }

    Percents[i] <- (Percent / Tri)
    if((Percents[i] >= 100 - alpha) && (decide == 0))
      decide <- SAMPLE_NUM[i]
  }

  plot(SAMPLE_NUM, Percents, xlim = c(0, POP_NUM), ylim = c(0, 100)
       , xaxt = "n", yaxt = "n", xlab = "抽出した標本の数(SAMPLE_NUM)"
       , ylab = "構成した信頼区間に母固有値が入った回数(Percents)", pch = 1)
  abline(h = 100 - alpha, col = 'red')

  axis(side = 2, at = c(0, 50, 100), labels = c(0, 50, 100), cex.axis=0.6)
  axis(side = 2, at = 100 - alpha, labels = 100 - alpha, col.ticks ='red', col.axis = "red")
  axis(side = 1, at = c(0, POP_NUM / 2, POP_NUM), labels = c(0, POP_NUM / 2, POP_NUM), cex.axis = 0.6)

  if(decide != 0){
     abline(v = decide, col = 'black')
     axis(side = 1, at = decide,  labels = decide, col.ticks ='black', col.axis = "black")
  }
}

```

```
52 library(kernlab)
53 data(spam)
54
55 ALLSPAM <- spam[, 1:57]
56
57 ConfiPlot(5, 5, nrow(ALLSPAM), ALLSPAM)
```

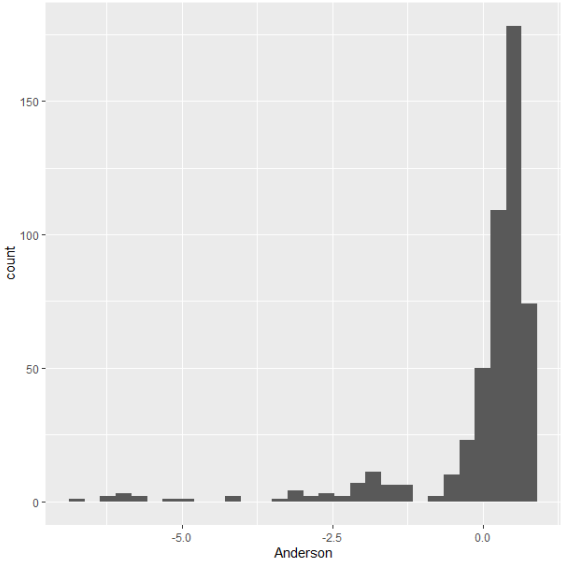

図 3　spam データを T.W.Anderson の検定統計量に変換したヒストグラム (試行回数 Tri：5，標本数 N：4500)，標準正規分布のようであれば良い

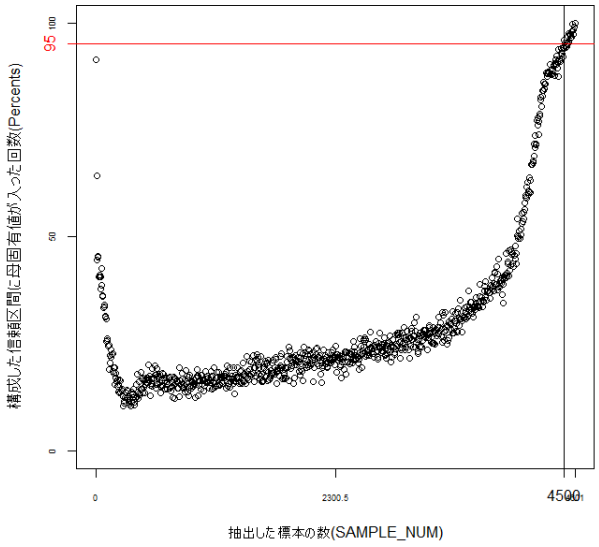

図 4　spam データの母固有値の 95% 信頼区間の構成に必要な標本数（Tri：5）

図 3 は標準正規分布には見えない．分布の対称性も表れない．図 4 については，ここではデータセットを母集団としているため傾きが不自然に大きくなる様子が見られるが，母集団のデータ数が無限大と考えれば標本数を増やせば徐々に信頼区間を構成することはできるのではないかと思う．ただ，母集団のデータ数が増えれば増えるほど傾きが小さくなるのであればあまりその期待は出来ない．また，「傾きが不自然に大きくなる」というのは，グラフとして不自然なのであって，標本数が母集団のデータ数に近づけば母固有値と標本固有値が近い値となるのは自然である．従って，母集団のデータ数に対して大きい標本数を取っていれば，検定統計量が上手く適用できたために信頼区間が構成出来たとは考えられない．

ちなみに，以下の図 5，6 は spam データの内でもラベル”spam” のみと”nonspam” のみをデータセットとして図 4 と同じコードを適用したものである．それぞれのデータ数はラベル”spam” が 1813 個，ラベル”nonspam” が 2788 個と異なることに注意が必要である．

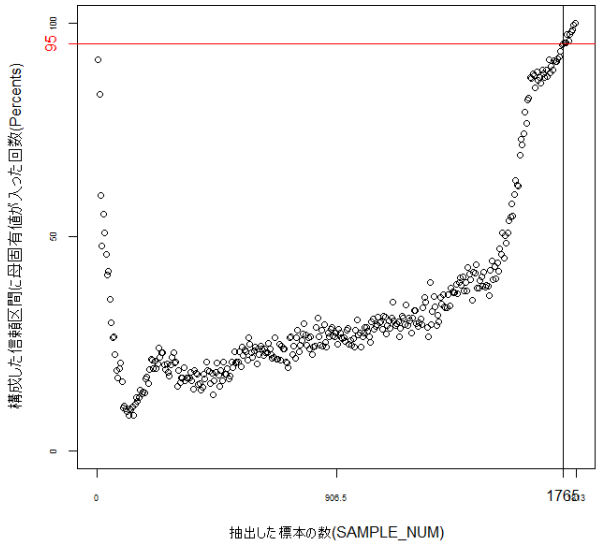

図 5　spam データ（ラベル”spam”）の母固有値の 95% 信頼区間の構成に必要な標本数（Tri：5）

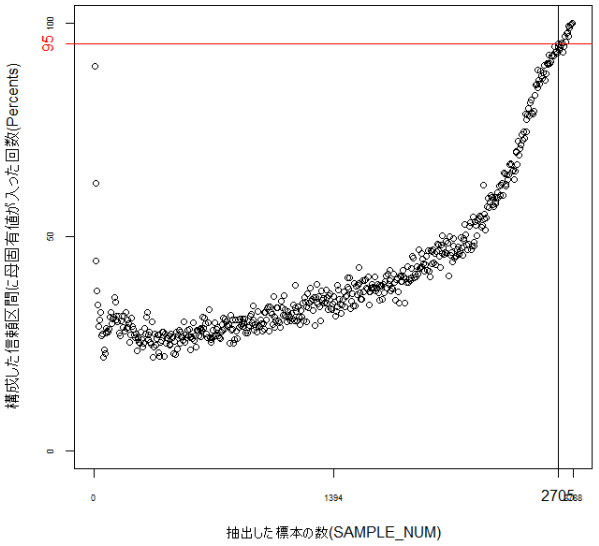

図 6　spam データ（ラベル”nonspam”）の母固有値の 95% 信頼区間の構成に必要な標本数（Tri：5）

データ数の違いから横軸のメモリが異なることを考慮すれば，図 5 と 6 に大きな違いは見られない．横軸を母集団のデータ数に対して標本数が占める割合（%）にすれば良いかもしれない．

多少だが，”nonspam” の方が滑らかに見える．これは，spam データがスパムメールの様々な特徴（メールの長さ，money や all といった単語等）を変数として持つことから，spam として判断されるもの中でも違いが表れるためだと考えられる．”nonspam” はメールとしてある程度の普遍性を持ったメールなのではないだろうか．

ちなみに，library(scar) にある decathlon データ（データ数 614，10 次元）については以下のようになる．

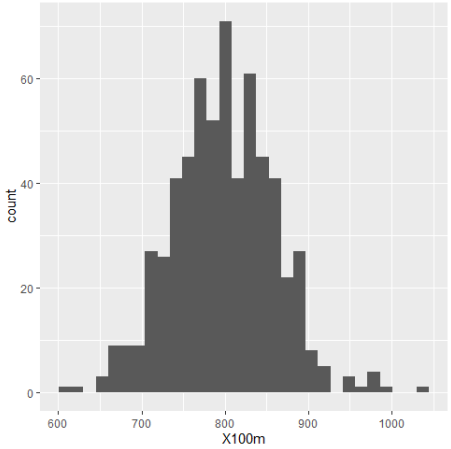

図 7　decathlon データの X100m 変数のヒストグラム，正規分布のようであれば良い

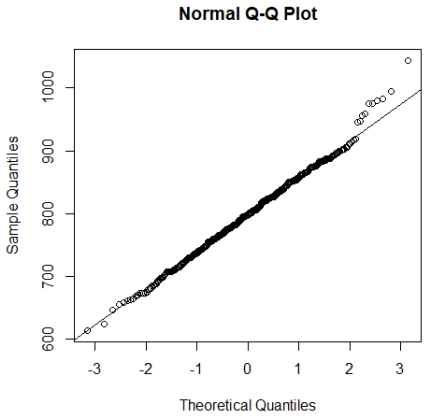

図 8　decathlon データの X100m 変数の qq プロット，直線に近ければ良い

図 7 と 8 より，decathlon データは spam データに比べて正規分布に従っているように見える[*10]．従って，spam デー

[*10] decathlon データの X100m 変数は shapiro.test() 関数により，p-value = 0.051 であり，spam データよりは $p-$ 値が大きく，有意水準によっては

タよりは検定統計量も上手く適用出来そうである．decathlon データは十種競技の点数であるから，平均的な点数があってその周辺に記録が集まっているのは納得できる．

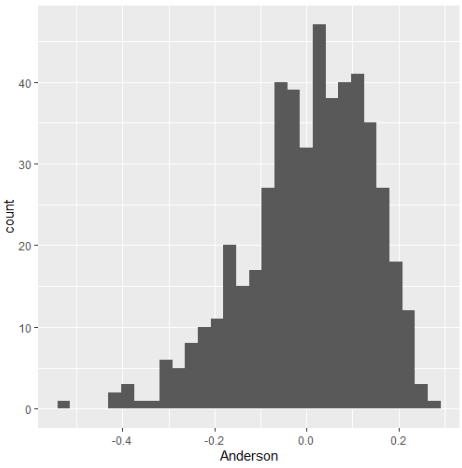

図 9　decathlon データを T.W.Anderson の検定統計量に変換したヒストグラム (Tri：5，N：600，標準正規分布のようであれば良い

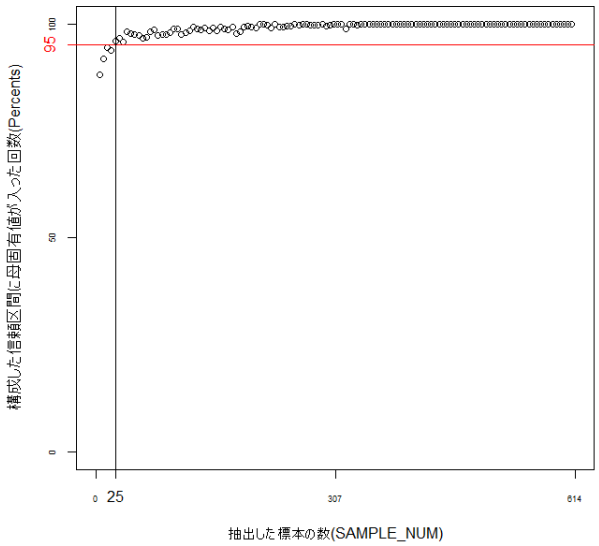

図 10　decathlon データの母固有値の 95% 信頼区間の構成に必要な標本数 (Tri：5)

図 9 は図 3 に比べれば標準正規分布に近い．Tri（試行回数）を大きくすればより近づくだろう．（データセットと同じ標本数を取ると，検定統計量の式 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ より与えられる値が全て 0 となるはずだが，実際にプロットしてみると非常に小さい誤差があるようだった．）

図 10 は図 4 に比べれば傾きに大きな変化もなく標本数が増えれば信頼区間が構成されている．さらには，母集団のデータ数に対して少ない標本数で満たしたい有意水準に対する信頼区間が構成出来ている．このグラフは次に見る図 11 のような，母集団を多変量正規分布から生成した時のグラフに近いこともわかる．以上の結果から，decathlon データは多変量正規分布に従うと考えて良いのではないだろうか．またこれらの結果は，正規分布に従う母集団に対して検定統計量が上手く信頼区間を構成できているということでもある．

以下は spam データ，decathlon データの期待値ベクトルと共分散行列を導出し，それをパラメータとして持つ多変量正規分布から 100 個と 1000 個のデータを生成して母集団を構成し，リスト 1 と同様のコード（リスト 2）を実行した結果である．

リスト 2　構成した漸近的な信頼区間 **(5)** $\frac{l_i}{1+\sqrt{2/n}z(\alpha)} \leq \lambda_i \leq \frac{l_i}{1-\sqrt{2/n}z(\alpha)}$ の確認を行う関数のプログラムコードとその実行

```
1  Anderson <- function(alpha, SAMPLE_NUM, SAMPLE_EV, POP_EV){
2
3    level <- 1 - (alpha / 2) / 100
4    z <- qnorm(level, 0, 1)
5    T_Anderson <- abs(sqrt((SAMPLE_NUM - 1) / 2) * (SAMPLE_EV - POP_EV) / POP_EV)
6
7   return(if(T_Anderson <= z) {1} else {0})
8
9  }
```

正規分布していると考えられる．

```
10
11 ConfiPlot <- function(alpha, Tri, Max, df){
12
13   POP_NUM <- Max
14   SAMPLE_NUM <- 2:POP_NUM
15   Percents <- numeric(length(SAMPLE_NUM))
16
17   POP <- as.matrix(df)
18   POP_EV <- (prcomp(POP)$sdev[1]) ^ 2
19
20   decide <- 0
21
22   for(i in 1:length(SAMPLE_NUM)){
23
24     Percent <- 0
25
26     for(j in 1:(100*Tri)){
27     SAMPLE <- POP[sample(1:nrow(df), SAMPLE_NUM[i]), ]
28     SAMPLE_EV <- (prcomp(SAMPLE)$sdev[1]) ^ 2
29     Percent <- Percent + Anderson(alpha, SAMPLE_NUM[i], SAMPLE_EV, POP_EV)
30     }
31
32     Percents[i] <- (Percent / Tri)
33     if((Percents[i] >= 100 - alpha) && (decide == 0))
34       decide <- SAMPLE_NUM[i]
35   }
36
37   plot(SAMPLE_NUM, Percents, xlim = c(0, POP_NUM), ylim = c(0, 100)
38        , xaxt = "n", yaxt = "n", xlab = "抽出した標本の数(SAMPLE_NUM)"
39        , ylab = "構成した信頼区間に母固有値が入った回数(Percents)", pch = 1)
40   abline(h = 100 - alpha, col = 'red')
41
42   axis(side = 2, at = c(0, 50, 100), labels = c(0, 50, 100), cex.axis=0.6)
43   axis(side = 2, at = 100 - alpha,  labels = 100 - alpha, col.ticks ='red', col.axis = "red")
44   axis(side = 1, at = c(0, POP_NUM / 2, POP_NUM), labels = c(0, POP_NUM / 2, POP_NUM), cex.axis=0.6)
45
46   if(decide != 0){
47      abline(v = decide, col = 'black')
48      axis(side = 1, at = decide,  labels = decide, col.ticks ='black', col.axis = "black")
49   }
50 }
51
52 library(kernlab)
53 data(spam)
54 DF <- spam[, 1:57]
55
56 Mu <- prcomp(DF)$center
57 Sigma <- cov(DF)
58
59 a <- 100
60
61 library(MASS)
62 POP <- mvrnorm(a, Mu, Sigma)
63
64 ConfiPlot(5, 5, a, POP)
```

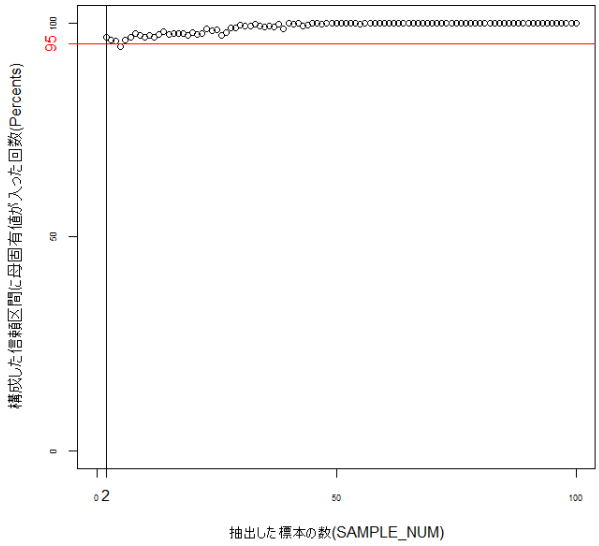

図 11　spam データの期待値ベクトルと共分散行列に従う多変量正規分布より母集団を構成（Tri：5，データ数：100）

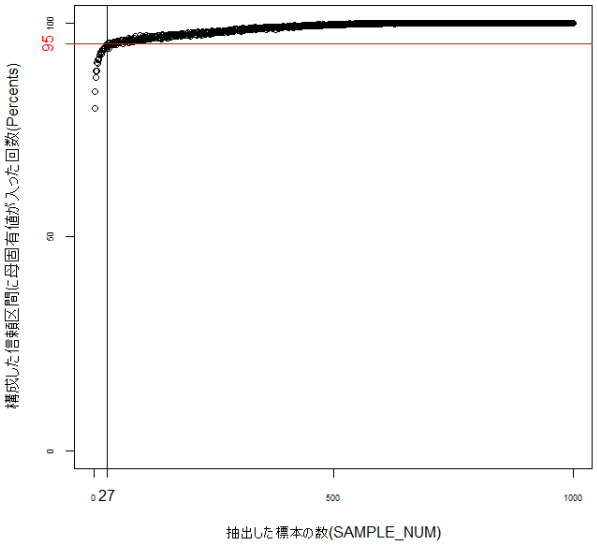

図 12　decathlon データの期待値ベクトルと共分散行列に従う多変量正規分布より母集団を構成（Tri：50，データ数：1000）

図 11 と 12 のどちらも，当然だが前回掲載したようなグラフになる．ただし，図 11 は次元数により処理時間が長くなるため Tri が少なく，必要な標本数があまりにも少なくなっている．前回と同じ結論だが，これらより，多変量正規分布に従う母集団のデータに対して，検定統計量によって少ない標本数で信頼区間が構成出来ていることがわかる．また，図 12 を図 10 と比較すると，必要な標本数が近いことがわかる．このことからも decathlon データが正規分布に近いとわかる．

1.4　T.W.Anderson(2003) で与えられる漸近的な受容域の適用を行うコード 1

以下では，データセットを標本として考え，母集団の未知な母固有値 λ_i について有意水準 α の仮説検定（両側検定）を行う．

ただし，帰無仮説 $H:\lambda_i=\lambda_i^0$，対立仮説 $K:\lambda_i\neq\lambda_i^0$ である．

データセットから標本固有値を導出し，T.W.Anderson で与えられる信頼係数 $1-\alpha$ を持つ λ_i の以下の信頼区間[*11]

(5)　$\dfrac{l_i}{1+\sqrt{2/n}z(\alpha)}\leq\lambda_i\leq\dfrac{l_i}{1-\sqrt{2/n}z(\alpha)}$,

の下限（左辺）と上限（右辺）を求める．これを各主成分で求めて信頼区間を図示し，区間が帰無仮説 $H:\lambda_i=\lambda_i^0$ を含むのか確認する．

リスト 3　構成した漸近的な受容域 **(4)** $-z(\alpha)\leq\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\leq z(\alpha)$ の適用を行う関数のプログラムコードとその実行

```
1 #Andersonの(5)に対応する.
2 #lambda_iの信頼区間の下限と上限を持つベクトルを返す関数
3 Anderson <- function(A, N, L){ #仮説検定の有意水準A/100のA%,
```

[*11] 漸近分布により区間が求められるから，正確には近似的な信頼区間である．

また，信頼区間 **(5)** は受容域 **(4)** $-z(\alpha)\leq\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\leq z(\alpha)$ より導出していた．受容域の中央にある検定統計量は，**(3)** $\sqrt{n}\frac{l_i-\lambda_i}{\sqrt{2}l_i}$ から分母の分布収束を用いて求めていた．**(3)** からは受容域が $l_i\left(1-\frac{z}{\sqrt{n/2}}\right)\leq\lambda_i\leq l_i\left(1+\frac{z}{\sqrt{n/2}}\right)$ と求まる．**(3)** を使って仮説検定や推定を行うことに問題は無さそうである．わざわざ分布収束を考えたのは，検定統計量に含まれる変数 l_i を出来るだけ減らすための操作を行ったと考えれば納得できる．

```
                                #標本数N，標本固有値Lを受け取る．

  ci <- c(0, 0) #lambda_iの信頼区間の下限と上限を入れるベクトル
  z <- qnorm(1 - (A / 2) / 100, 0, 1) # 100(A/2/100)%点z

  ci[1] <- L / (1 + sqrt(2 / (N - 1)) * z) #lambda_iの信頼区間の下限，(5)の左辺
  ci[2] <- L / (1 - sqrt(2 / (N - 1)) * z) #lambda_iの信頼区間の上限，(5)の右辺

 return(ci) #lambda_iの信頼区間の下限と上限を持つベクトルを返す

}

#データセット全てを標本とし，それより第一主成分から順に対応する標本固有値を導出する．
#それぞれの標本固有値でAnderson()関数を用いて信頼区間の上限と下限を導出する．
#構成した信頼区間が帰無仮説H：lambda_i=lambda_i^0を含むかグラフで確認するため，
#各主成分に対応する標本固有値における信頼区間と帰無仮説の位置関係をプロットする関数.
AndersonConInt <- function(A, df, Lam0, y.lower, y.upper){
                          #仮説検定の有意水準A/100のA%，標本として受け取るデータセットdf，
                          #帰無仮説H：lambda_i=Lam0，グラフ自体のy軸の上限(y.upper)と下限(y.lower)を受け取る．
  M <- ncol(df) #データセットの次元数
  plot(NULL, xlab="主成分", ylab="Confidence Interval", xlim=c(1, M),
       ylim=c(y.lower, y.upper)) #各主成分に対応するように信頼区間を表示する．

  for(i in 1:M){
   ci <- Anderson(A, nrow(df), (prcomp(df)$sdev[i]) ^ 2) #第i主成分に対応する標本固有値で
   ci.lower <- ci[1]                                     #Anderson()関数を呼び出し，
   ci.upper <- ci[2]                                     #信頼区間の下限と上限をci.lower, ci.upperに
   if( ci.lower > Lam0 | ci.upper < Lam0){               #代入する．
    points(x=c(i,i), y=ci, pch=15, cex=0.5, col=1)       #信頼区間に帰無仮説H：lambda_i=Lam0が含まれていなければ，
    segments(x0=i, y0=ci.lower, x1=i, y1=ci.upper, col=1)#その信頼区間を黒色で描画する．
   }
   else{
    points(x=c(i,i), y=ci, pch=15, cex=0.5, col=2)       #信頼区間に帰無仮説H：lambda_i=Lam0が含まれれば，
    segments(x0=i, y0=ci.lower, x1=i, y1=ci.upper, col=2)#その信頼区間を赤色で描画する．
   }
  }
  abline(h=Lam0, lty=2)   #帰無仮説H：lambda_i=Lam0の位置に点線を引く．

}

library(kernlab)
data(spam)
library(scar)
data(decathlon)

df1 <- spam[,1:57]            #spamデータ全てのデータセット
df2 <- spam[1:1813, 1:57]     #spamデータのラベル"spam"のみのデータセット
df3 <- spam[1814:4601, 1:57] #spamデータのラベル"nonspam"のみのデータセット
df4 <- decathlon              #decathlonデータのデータセット
df5 <- iris[, 1:4]            #irisデータのデータセット

Mu <- prcomp(df2)$center      #df2のデータセットの期待値ベクトル
Sigma <- cov(df2)             #df2のデータセットの共分散行列
a <- 2000                     #標本のデータセットとして作りたい標本の数
library(MASS)
df6 <- mvrnorm(a, Mu, Sigma) #標本として考えたいデータセットを生成する．

AndersonConInt(1, df1, 0, 0, 35000)  #仮説検定の有意水準A/100を1/100=0.01として，
                                     #標本として受け取るデータセットdfをdf1，
                                     #ここではspamデータ全てのデータセットとして，
                                     #帰無仮説H：lambda_i=Lam0をLam0=0，
```

```
65                                #グラフ自体のy軸の上限を35000，下限を0としてAndersonConInt()関数を呼び出す．
```

1.5　リスト 3 の実行結果と考察

図 13 は AndersonContInt(1, df1, 0, 0, 35000) の結果である．ただし，これは仮説検定の有意水準 A/100 を 1/100=0.01 として，標本として受け取るデータセット df を df1，ここでは spam データ全てのデータセットとして，帰無仮説 H：λ_i =Lam0 を Lam0=0，グラフ自体の y 軸の上限を 35000，下限を 0 として AndersonConInt() 関数を呼び出している．

図 13 から 16 までは df1（spam データ全て）を標本として使用している．

図 13 から 15 にかけてプロットする上限を落としている．

図 13 から 15 までは有意水準が 1.0 だが図 16 は有意水準が 10.0 となっている．

帰無仮説が「棄却されている」のが「黒色」で，『棄却されていない』のが『赤色』の信頼区間として描画されている．横軸に平行な点線が上限と下限の間になければ棄却され，間にあれば棄却されない．

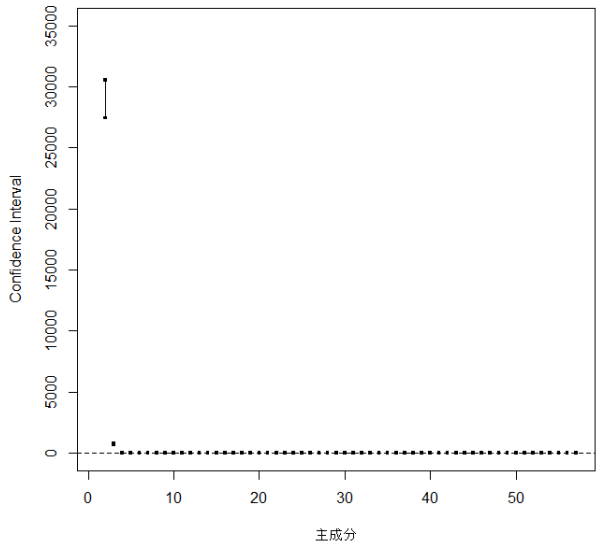

図 13　リスト 3 によるプロット (AndersonContInt(1, df1, 0, 0, 35000))

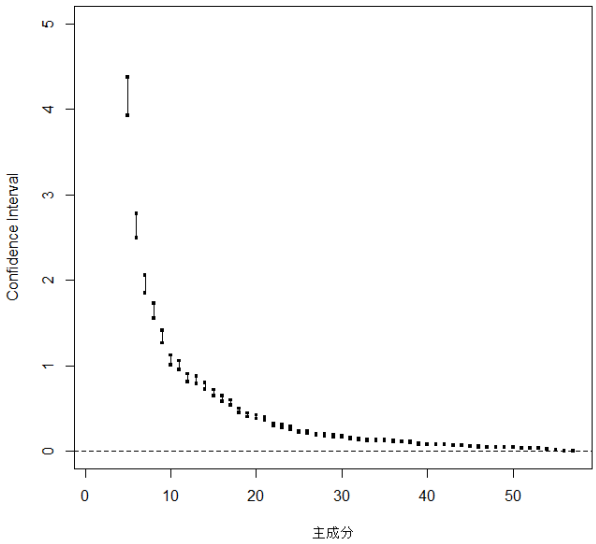

図 14　上限 (y.upper) が 5

図 13 で上限を大きく設定しているのは，第一主成分に対応する標本固有値で構成される 99% 信頼区間の上限と下限が，他の主成分に対するそれらに比べて非常に大きくなっているためである．また，図 4 を思い出すと，spam データは信頼区間を構成するのに必要な標本数はデータセットを母集団としても大きかった（4601 個に対して 4500 個必要）．従って，データセット全てを使わずにデータセットから少ない標本を抽出し，それをデータセットとしてコードに入れることで計算時間は短縮するというようなことは出来そうにない．

図 13 だと第三主成分以降の標本固有値で構成される 99% 信頼区間がわからないため，図 14 では上限を小さく 5 と設定している．ここでは第 5 主成分以降に対応する標本固有値による 99% 信頼区間が確認出来る．主成分の番号が進むにつれて 99% 信頼区間の上限と下限が 0 に近づいている．

これは帰無仮説 $H:\lambda_i=0$ の仮説検定だが，帰無仮説 $H:\lambda_i=\lambda_i^0$ における有意水準 α の両側検定の（漸近的な）受容域

(4)　$-z(\alpha)\leq\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\leq z(\alpha)$,

を見ると，λ_i^0 は 0 を取れないため，信頼区間が帰無仮説 $H:\lambda_i=0$ を含むことはない．グラフで言えば赤色の信頼区間が表れることはない．従って，帰無仮説を設定するのであれば，Lam0 には 0 でない値を設定しなければならない．また，グラフを見てもわかる通り，主成分の番号が進むにつれて信頼区間が狭くなる．つまり，両側検定だと各主成分において棄却されないような母固有値を含む信頼区間が狭くなり，棄却されるが信頼区間の上限も下限も λ_i^0 を下回るような信頼区間が表れる．特にこの分析では，固有値がどれぐらい小さいのかを見たいのであるから，この場合を棄却されないとして扱いたい．従って，両側検定として特定の値で帰無仮説 $H:\lambda_i=\lambda_i^0$ を設定するよりは，帰無仮説が複合仮説の（右）片側検定として母固有値をある値以下とする帰無仮説 $H:\lambda_i\le\lambda_i^0$ を設定するのが良い[*12][*13]．

ちなみに，以下 $m=1,2,...$ に対して，テキストの Section 11.7.1.(pp.478-479) では特定の量 γ を設定し，片側検定として第 $m+1$ 主成分以降の主成分が無視できるかどうか（つまり m 番目までの主成分が母集団に対する上手い推定を与えるかどうか）の仮説検定を帰無仮説 $H:\lambda_{m+1}+\cdots+\lambda_p\ge\gamma$ で考えている．つまり，第 $m+1$ 主成分以降の主成分を無視する上で根拠を得たいという仮説検定である．また，Section 11.7.2.(pp.479-480) では，特定の量 δ を設定し，同様のアイデアで帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\ge\delta$ を考えている，さらには，Section 11.7.3.(pp.480-482) では両側検定として帰無仮説 $H:\lambda_{m+1}=\cdots=\lambda_p$ を考えている．

特に Section 11.7.2.(pp.479-480) で考えられている帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\ge\delta$ を用いた仮説検定については，信頼区間の厳密な導出は与えずに，最後にその適用を載せている．

帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\ge\delta$ は，両辺から 1 を引いて式を整理すれば帰無仮説 $H:\frac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\cdots+\lambda_p}\le1-\delta$ と同値である．つまり $m=1,2,...$ に対するこの仮説検定は，特定の量 $1-\delta$ を設定し，第 m 主成分の累積寄与率とその量を比較することにより，第 m 主成分以降の主成分を無視する（つまり $m-1$ 番目までの主成分が母集団に対する上手い推定を与える）根拠を得たいという仮説検定であることがわかる．テキストで考えている帰無仮説では，$m=1,2,...$ に対して第 $m+1$ 主成分が対応するが，同値として示した帰無仮説を考える場合は $m=1,2,...$ に対して第 m 主成分が対応する．前者は捨てたい $p-m$ 個の主成分に対する帰無仮説だが，後者は使用した m 個の主成分に対する帰無仮説であると考えられる．なぜ番号 m と主成分の対応がわかり難く，解析もしにくそうな前者が本文に載っているのか考えると，テキストの流れとして，ある程度元のデータを次元縮小した後に捨てるものについて検定を行いたいというモチベーションがあるためこのような帰無仮説を挙げているのだと考えられる．また，前者は仮説検定を行っている m に対して第 $m+1$ 主成分以降の固有値の和が固有値全体の和に対して δ 未満であれば棄却するという仮説検定であり，これは設けた有意水準 α に対して固有値が小さいためある主成分以降は使わない，と解釈がしやすい．しかし，後者は仮説検定を行っている m に対して第 m 主成分の累積寄与率が $1-\delta$ より大きければ棄却するという仮説検定であり，有意水準 α に対して累積寄与率が大きいためある主成分以降は使わない，と解釈が直感的ではない（また，$m=1,2,...$ であるから，$m=1$ の時万が一棄却されれば第 0 主成分までを使用すればいいというおかしな解釈になるため，m を 2 以上としなければならない.）．

例（$m=5$ で棄却される場合）

1. 帰無仮説 $H:\frac{\lambda_{5+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}=\frac{\lambda_6+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\ge\delta$ が棄却される → 第 6 主成分以降の固有値の和は全体の固有値の和に対して占める割合が設定した量より小さいため，第 6 主成分以降は使用しないという根拠になる．
2. 帰無仮説 $H:\frac{\lambda_1+\cdots+\lambda_5}{\lambda_1+\cdots+\lambda_p}\le1-\delta$ が棄却される → 第 5 主成分の累積寄与率が設定した量より大きいため，第 5 主成分

[*12] （左）片側検定を考えるのであれば，不等号を逆にして，主成分の番号を後ろから進めればよい．

[*13] ここで，仮説検定を考える場合，帰無仮説の λ_i^0 を正の値のみで設定することに注意したい．この注意は，これまでのグラフのように信頼区間の下限と上限が常に正になっている，つまり

(5) $\frac{l_i}{1+\sqrt{2/nz}(\alpha)}\le\lambda_i\le\frac{l_i}{1-\sqrt{2/nz}(\alpha)}$,

より標本固有値 l_i が常に正になっていることや，始めに述べた仮定では標本固有値と母固有値について正となるように仮定がされていることからも当然の注意であり，コードでも常に正で求まるようになっている．また，特に上限（右辺）の分母についても注意が必要で，分母が 0 とならないかつ常に正となるために $(0\le)\sqrt{2/nz}(\alpha)<1$，$(n=N-1)$ を満たすように有意水準 α に対して標本数 N を取らなければならない．仮に，標本固有値や母固有値が正を取るとは限らず，かつそれぞれを大小付けないとすれば，それらが対応するように片側・両側検定や帰無仮説を設定しなければならない．

以降は使用しないという根拠になる．

また，テキストで考えている仮説検定は小さい固有値たちを考えるため，高次元のような一つ一つの固有値は小さいが和は無視できないような場合にも有効である．

話が逸れたが，spam データに対する両側検定に戻る．

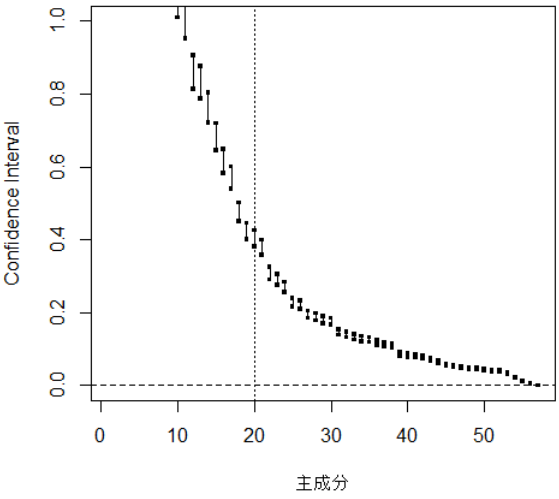

図 15　グラフの上限 (y.upper) が 1 で仮説検定の有意水準 A/100 が 1/100=0.01

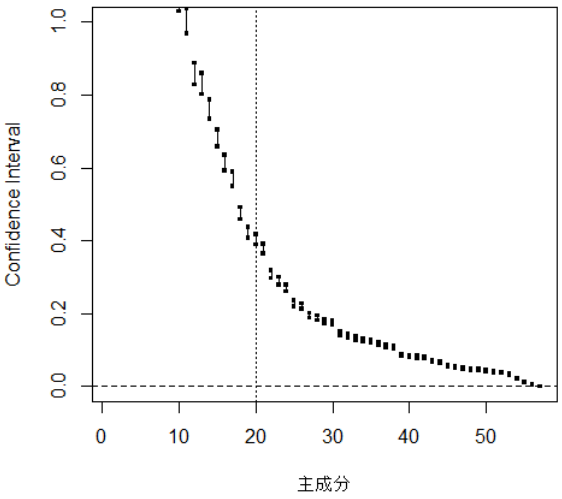

図 16　仮説検定の有意水準 A/100 が 10/100=0.10

図 15，16 は有意水準が 0.01，0.10 と異なるグラフであり，信頼区間との対応を見れば 99% 信頼区間，90% 信頼区間に対応するため，有意水準が小さく受容域が広い図 15 は 16 に比べて信頼区間が広くなる．わかり難いが，両方の図で第 20 主成分に線を引いている．それぞれの信頼区間の幅を比較すれば確認出来る．

図 17 は AndersonContInt(1, df2, 0, 0, 5) の結果である．上記との違いは，標本として df2（spam データのラベル"spam" のみのデータセット）を使用している点である．

図 18 は AndersonContInt(1, df3, 0, 0, 5) の結果である．上記との違いは，標本として df3（spam データのラベル"nonspam"のみのデータセット）を使用している点である．

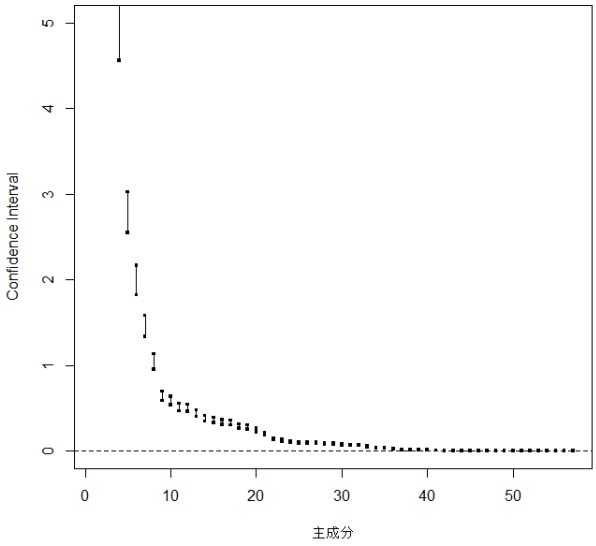

図 17　spam データのラベル"spam"のみのデータセット

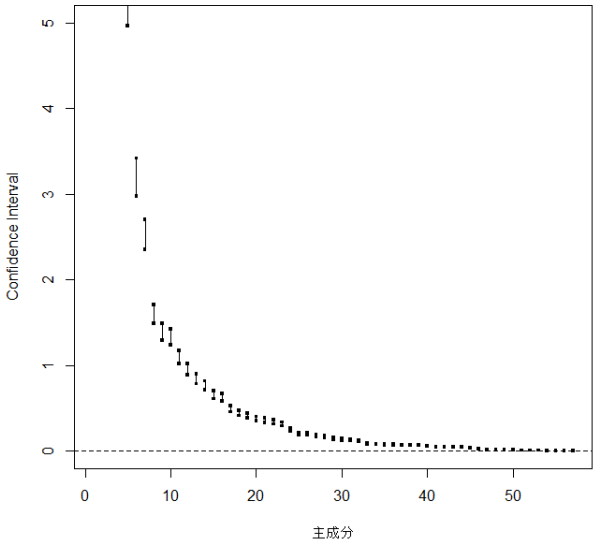

図 18　spam データのラベル"nonspam"のみのデータセット

図 17 と 18 を比べると，図 17 は広い信頼区間を持つ主成分が第 8 主成分まで続き，図 18 は第 6 主成分まで続いていることがわかる．

図 19 は AndersonContInt(1, df4, 0, 0, 35000) の結果である．上記との違いは，標本として df4（decathlon データのデータセット）を使用している点である．

図 20 は AndersonContInt(1, df5, 0, 0, 5) の結果である．上記との違いは，標本として df5（iris データのデータセット）を使用している点である．

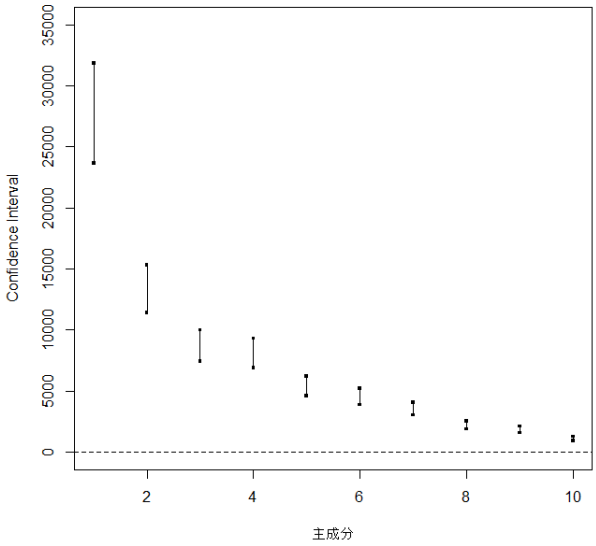

図 19　decathlon データのデータセット

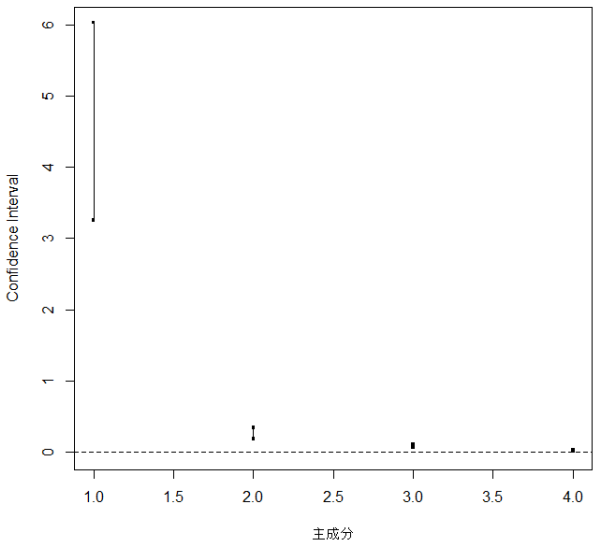

図 20　iris データのデータセット

図 19 より，decathlon データは 10 次元であるからか各主成分に対応する信頼区間が帰無仮説 $H:\lambda_i=0$ から遠い（グラフの上限に注意）．やはり先ほど述べたとおり，次元縮小したいのであれば帰無仮説は片側検定で設定した方が良いように思う．また，図 10 を思い出せば，decathlon データは信頼区間を構成するのに必要な標本数はデータセットを母集団とすると少なかった．従って，データセット全てを使わなくてもデータセットから少ない標本を抽出し，それをデータセットとしてコードに入れることで計算時間は短縮できそうである．

図 20 より，iris データは第 1 主成分に対応する信頼区間以外が帰無仮説 $H:\lambda_i=0$ に近く，4 次元なので効果は小さいが次元縮小しやすそうである．

図 21 は AndersonContInt(1, df6, 0, 0, 1000000) の結果である．上記との違いは，標本として spam データの期待値ベクトルと共分散行列に従う多変量正規分布のデータセット（データ数 2000）df6 を使用している点である．

図 22 は AndersonContInt(1, df6, 0, 0, 0.01) の結果である．上記との違いは，プロットする上限を落としている．

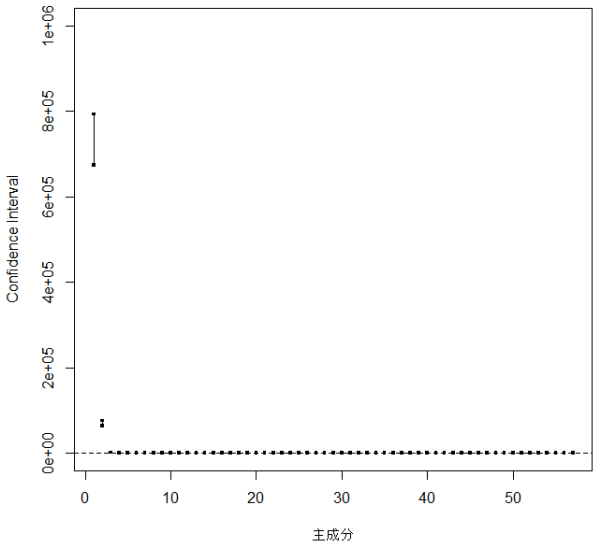

図 21　spam データの期待値ベクトルと共分散行列に従う多変量正規分布から生成したデータセット

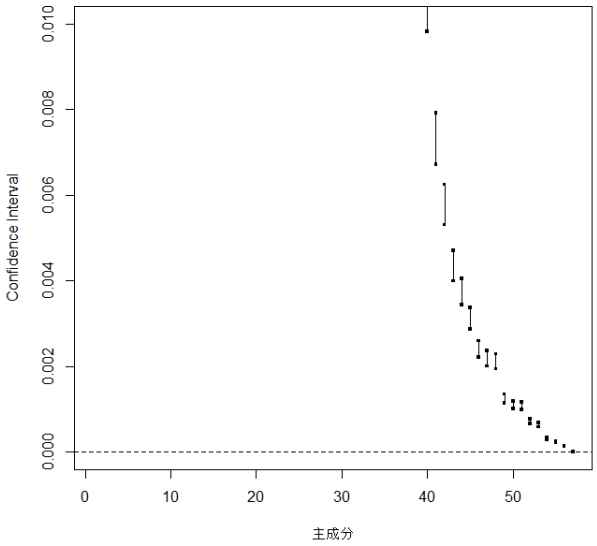

図 22　グラフの上限 (y.upper) が 0.01

これまでの両側検定に関する図全てを通してもわかるが，図 21，22 を見ると，対応する主成分毎に大きく異なるような値を上限と下限として持つ信頼区間を図示すると，見たい信頼区間によってグラフの縦の尺度を操作する必要があり，分析をしにくいことがわかる．主成分毎に固有値が大きく異なることを考えれば，固有値の値そのものに対して何かを仮定するような帰無仮説であれば，信頼区間の上限と下限が一つ一つの固有値そのものに依ってしまい，このような分析のしにくさを引き起こすように思う．

従って，主成分分析で仮説検定を行う場合，主成分毎に大きく値が異なる固有値で構成しても信頼区間の上限と下限が大きく異なりすぎないような仮説検定を行いたいと感じる．固有値を用いる上でそのような仮説を立てられる値は，寄与率や累積寄与率が想起しやすい．それが Section 11.7.2.(pp.479-480) で考えられている帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\geq\delta$ を用いた仮説検定であり，最後に適用を載せている．

次は単にこれまでの両側検定を片側検定にしたコードとその結果である．

1.6　T.W.Anderson(2003) で与えられる漸近的な受容域の適用を行うコード 2

以下では，データセットを標本として考え，母集団の未知な母固有値 λ_i について有意水準 α の仮説検定（片側検定）を行う．

ただし，帰無仮説 H：$\lambda_i \leq \lambda_i^0$，対立仮説 K：$\lambda_i > \lambda_i^0$ である．

データセットから標本固有値を導出し，T.W.Anderson で与えられる信頼係数 $1-\alpha$ を持つ λ_i の以下の信頼区間[*14]

(5)′　$\frac{l_i}{1+\sqrt{2/n}z(\alpha)} \leq \lambda_i$,

の下限（左辺）を求める．これを各主成分で求めて信頼区間を図示し，区間が帰無仮説 H：$\lambda_i \leq \lambda_i^0$ と重なるかどうか確認する．

リスト 4　構成した漸近的な受容域 **(4)′** $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0} \leq z(\alpha)$ の適用を行う関数のプログラムコードとその実行

```
#lambda_iの信頼区間(5)'の下限を返す関数
Anderson <- function(A, N, L){ #仮説検定の有意水準A/100のA%,
                               #標本数N，標本固有値Lを受け取る．

  z <- qnorm(1 - A / 100, 0, 1) # 100(A/100)%点z（両側検定であれば100(A/2/100)%を作る）

  ci <- L / (1 + sqrt(2 / (N - 1)) * z) #lambda_iの信頼区間の下限，(5)の左辺

 return(ci) #lambda_iの信頼区間の下限を返す

}

#データセット全てを標本とし，それより第一主成分から順に対応する標本固有値を導出する．
#それぞれの標本固有値でAnderson()関数を用いて信頼区間の下限を導出する．
#構成した信頼区間が帰無仮説H：lambda_i<=lambda_i^0と重なるかグラフで確認するため，
#各主成分に対応する標本固有値における信頼区間と帰無仮説の位置関係をプロットする関数.
AndersonConInt <- function(A, df, Lam0, y.lower, y.upper){
                          #仮説検定の有意水準A/100のA%，標本として受け取るデータセットdf,
                          #帰無仮説H：lambda_i<=Lam0，グラフ自体のy軸の上限(y.upper)と下限(y.lower)を受け取る．
  M <- ncol(df) #データセットの次元数
  plot(NULL, xlab="主成分", ylab="Confidence Interval", xlim=c(1, M),
       ylim=c(y.lower, y.upper)) #各主成分に対応するように信頼区間を表示する．

  for(i in 1:M){
   ci.lower <- Anderson(A, nrow(df), (prcomp(df)$sdev[i]) ^ 2)
                                    #第i主成分に対応する標本固有値でAnderson()関数を呼び出し，
                                    #信頼区間の下限をci.lowerに代入する．
   if( ci.lower > Lam0){
    points(x=i, y=ci.lower, pch=15, cex=0.5, col=1)       #信頼区間に帰無仮説H：lambda_i<=Lam0が重ならなければ，
    segments(x0=i, y0=ci.lower, x1=i, y1=2*y.upper, col=1)#その信頼区間を黒色で描画する．
   }
   else{
    points(x=i, y=ci.lower, pch=15, cex=0.5, col=2)       #信頼区間に帰無仮説H：lambda_i<=Lam0が重なれば，
    segments(x0=i, y0=ci.lower, x1=i, y1=2*y.upper, col=2)#その信頼区間を赤色で描画する．
   }
  }
  abline(h=Lam0, lty=2)   #帰無仮説H：lambda_i<=Lam0の位置に点線を引く．

}

library(kernlab)
```

[*14] 漸近分布により区間が求められるから，正確には近似的な信頼区間である．

```
data(spam)
library(scar)
data(decathlon)

df1 <- spam[,1:57]              #spamデータ全てのデータセット
df2 <- spam[1:1813, 1:57]       #spamデータのラベル"spam"のみのデータセット
df3 <- spam[1814:4601, 1:57] #spamデータのラベル"nonspam"のみのデータセット
df4 <- decathlon                #decathlonデータのデータセット
df5 <- iris[, 1:4]              #irisデータのデータセット

Mu <- prcomp(df2)$center        #df2のデータセットの期待値ベクトル
Sigma <- cov(df2)               #df2のデータセットの共分散行列
a <- 2000                       #標本のデータセットとして作りたい標本の数
library(MASS)
df6 <- mvrnorm(a, Mu, Sigma) #標本として考えたいデータセットを生成する.

AndersonConInt(1, df1, 1.0, 0, 5)  #仮説検定の有意水準A/100を1/100=0.01として,
                                   #標本として受け取るデータセットdfをdf1,
                                   #ここではspamデータ全てのデータセットとして,
                                   #帰無仮説H:lambda_i<=Lam0をLam0=1.0,
                                   #グラフ自体のy軸の上限を5, 下限を0としてAndersonConInt()関数を呼び出す.
```

1.7　リスト 3 の実行結果と考察

図 23 は，AndersonConInt(1, df1, 1.0, 0, 5) の実行結果である．つまり，仮説検定の有意水準 A/100 を 1/100=0.01 として，標本として受け取るデータセット df を df1，ここでは spam データ全てのデータセットとして，帰無仮説 $H : \lambda_i \leq \lambda_i^0$ を $\lambda_i^0 = 1.0$，グラフ自体の y 軸の上限を 5，下限を 0 として AndersonConInt() 関数を呼び出している．

帰無仮説が「棄却されている」のが「黒色」で，『棄却されていない』のが『赤色』の信頼区間として描画されている．横軸に平行な点線より下限が上にあれば棄却され，下にあれば棄却されない．

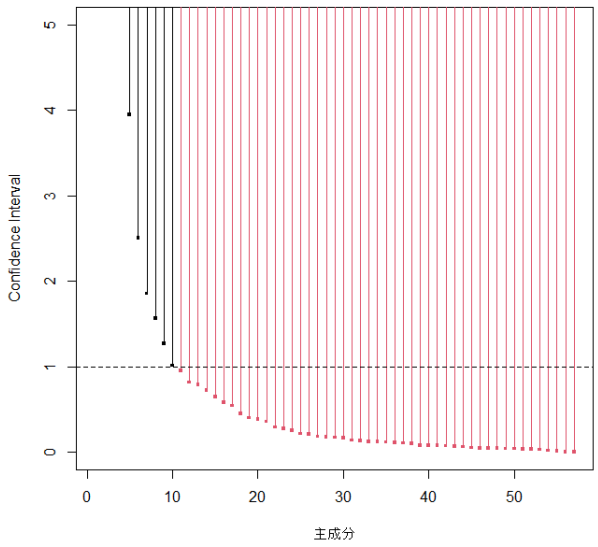

図 23　リスト 4 によるプロット（AndersonConInt(1, df1, 1.0, 0, 5)）

図 23 を見ると，第 11 主成分以降で帰無仮説 $H : \lambda_i \leq 1.0$ が棄却されなくなることがわかる．分析前に「固有値が 1.0

以下になるまでの主成分を用いる」というような設定がされているのであれば，この仮説検定を根拠として第 10 主成分までの主成分を用いたと主張出来る（ただし，先述の「棄却されない」ことに対する注意や，今 λ_i^0 として 1.0 を設定したのは結果ありきで設定していることへの注意は必要である）．

図 24 は，AndersonConInt(1, df4, 0, 0, 30000) の実行結果である．つまり，仮説検定の有意水準 A/100 を 1/100=0.01 として，標本として受け取るデータセット df を df4，ここでは decathlon データ全てのデータセットとして，帰無仮説 H：$\lambda_i \leq \lambda_i^0$ を $\lambda_i^0 = 0$，グラフ自体の y 軸の上限を 30000，下限を 0 として AndersonConInt() 関数を呼び出している．

図 25 は，AndersonConInt(1, df4, 5000, 0, 30000) の実行結果である．つまり，仮説検定の有意水準 A/100 を 1/100=0.01 として，標本として受け取るデータセット df を df4，ここでは decathlon データ全てのデータセットとして，帰無仮説 H：$\lambda_i \leq \lambda_i^0$ を $\lambda_i^0 = 5000$，グラフ自体の y 軸の上限を 30000，下限を 0 として AndersonConInt() 関数を呼び出している．

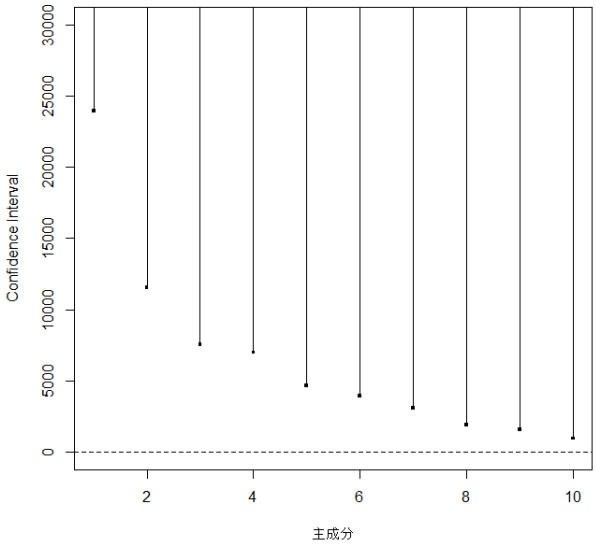

図 24　AndersonConInt(1, df4, 0, 0, 30000)

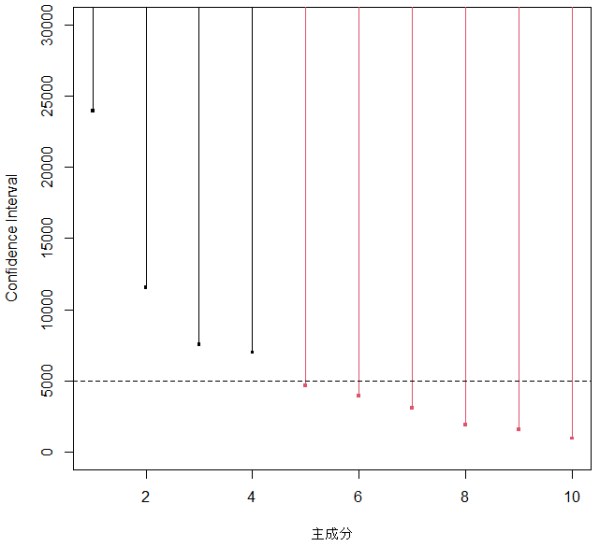

図 25　AndersonConInt(1, df4, 5000, 0, 30000)

図 24，25 を見ると，先ほど述べた「固有値が λ_i^0 以下になるまでの主成分を用いる」というような設定をする場合，データセット毎に指標となる λ_i^0 を変えなければならないということがわかる．つまり，どのようなデータセットに対しても同じ λ_i^0 で帰無仮説を立てられるとは限らないということである．decathlon データのように，最も小さい固有値に対応する主成分でも，その固有値は固有値全体で見て小さい（寄与率が小さい）だけで固有値自体は大きいということがあり，λ_i^0 はただ小さければいいというわけでもない．

また，図 25 では，帰無仮説 H：$\lambda_i \leq \lambda_i^0$ を $\lambda_i^0 = 5000$ と設定した上で，第 5 主成分以降が棄却されないために第 4 主成分までを用いることに根拠を持たせられるが，棄却されない理由として「固有値の値が 5000 以下であるから」というのはその値の小ささに（感覚的だが）説得力を感じない．

固有値一つに対して帰無仮説を設定するような仮説検定においては，グラフの描画における難点だけでなく，帰無仮説として設定する λ_i^0 もデータセットによって変えなければならないことや，帰無仮説として設ける基準が大きくなってしまうことがあるといった難点も見つかった．

以下は，これらを解決するような仮説検定として累積寄与率を用いた仮説検定であり，Section 11.7.2.(pp.479-480) で考

えられている帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\geq\delta$ を用いた仮説検定である[*15].

1.8 T.W.Anderson(2003) で与えられる漸近的な受容域の適用を行うコード 3

以下では，データセットを標本として考え，$m=1,2,...$ に対して，母集団の未知なパラメータ $\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}$（第 $m+1$ 主成分の母固有値 λ_{m+1} への帰無仮説と対応）について有意水準 α の仮説検定（片側検定）を行う．

ただし，特定の量 δ に対して，帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\geq\delta$，対立仮説 $K:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}<\delta$ である．

データセットから標本固有値を導出し，T.W.Anderson(2003) の Section 11.7.2.(pp.479-480) で与えられる信頼係数 $1-\alpha$ を持つ $\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}$ の以下の信頼区間[*16]

$$\textbf{(8)}\quad \frac{\sum_{i=m+1}^{p}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\leq\frac{\sum_{i=m+1}^{p}l_i}{\sum_{i=1}^{p}l_i}+z\frac{\left[2\left(\sum_{i=m+1}^{p}l_i\right)^2\sum_{i=1}^{m}l_i^2+2\left(\sum_{i=1}^{m}l_i\right)^2\sum_{i=m+1}^{p}l_i^2\right]^{\frac{1}{2}}}{\sqrt{n}\left(\sum_{i=1}^{p}l_i\right)^2},$$

の上限（右辺）を求める[*17]．ただし，$n=N-1$，$z=z(2\alpha)$[*18]．これを各主成分で求めて信頼区間を図示し，区間が帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\geq\delta$ と重なるか確認する．

これまでの仮説検定と対応するように説明すると，この仮説検定は，特定の量 δ を設定し，$m=1,2,...$ に対して，第 $m+1$ 主成分から第 p 主成分の固有値の和が固有値全体に占める割合と δ を比較することにより，第 $m+1$ 主成分以降の主成分を無視する（つまり m 番目までの主成分が母集団に対する上手い推定を与える）根拠を得たいという仮説検定であることがわかる．$m=1,2,...$ と第 $m+1$ 主成分に対して仮説検定を進めていく．第 $m+1$ 主成分において帰無仮説が棄却されれば，それは第 $m+1$ から第 p 主成分に対応する固有値の和が固有値全体の和に占める割合が設定していた δ 未満ということになる．つまり，固有値全体の和に占める割合が δ 未満になるような固有値たちを主成分として用いる意味が無いという基準の下で，第 $m+1$ 主成分以降を捨てることができる．

これまでの仮説検定と違い，棄却されるまでの主成分を考えるため，「棄却されること」を根拠として用いることができる．

リスト 5　信頼区間 **(8)** に対応するような，構成した漸近的な受容域の適用を行う関数のプログラムコードとその実行

```
#信頼区間(8)の上限を返す関数
Anderson <- function(A, df, M){ #仮説検定の有意水準A/100のA%,
                                #標本として受け取るデータセットdf,
                                #第M主成分（呼び出し時はM=m+1）の番号を受け取る.

  SL.mp <- sum( (prcomp(df)$sdev[M:ncol(df)])^2 ) #第m+1主成分を考えていることがわかりやすいように,
  SL.sp <- sum( (prcomp(df)$sdev[1:ncol(df)])^2 ) #また引数がM=m+1となるように，関数を定義している.
  SL.sm <- sum( (prcomp(df)$sdev[1:(M-1)])^2 )

  SL.mp2 <- sum( (prcomp(df)$sdev[M:ncol(df)])^4 )
  SL.sm2 <- sum( (prcomp(df)$sdev[1:(M-1)])^4 )

  z <- qnorm(1 - A / 100 , 0, 1) # 100(A/100)%点z（両側検定であれば100(A/2/100)%を作る）

  s.error.n <- z * sqrt( 2*((SL.mp)^2)*(SL.sm2) + 2*((SL.sm)^2)*(SL.mp2) ) #第二項目の分子
  s.error.d <- sqrt(nrow(df)-1) * ((SL.sp)^2)                               #第二項目の分母
  s.error <- s.error.n / s.error.d                                          #第二項目
```

[*15] 累積寄与率以外にも，共分散行列の対角成分の和（分散の和及び固有値の和）に対する対角成分の割合，相関行列の対角成分の和（次元 p であれば p）に対する対角成分の割合等も考えられるが，累積寄与率と発想は同じである．

[*16] 漸近分布により区間が求められるから，正確には近似的な信頼区間である．

[*17] テキストに信頼区間の導出が無いため，ここでは成り立つものとして適用する．標本の主成分の累積寄与率を用いた，母集団の主成分の累積寄与率の近似だと考える．

[*18] ただし，$z(2\alpha)$ は $N(0,1)$ の上側 $100(\alpha)$% 点である．すなわち $P_Z^H\{Z>z(2\alpha)\}=\alpha$ を満たす点である（ただし $Z\sim N(0,1)$）．一般に，確率変数 X の確率密度関数 f_X を持つ分布について，$0<\alpha<1$ に対して $\int_z^\infty f_X(x)dx=\alpha$ となる z を上側 100α% 点という．

```
18
19    ci <- SL.mp/SL.sp + s.error #信頼区間(8)の上限
20   return(ci)                   #信頼区間(8)の上限を返す
21
22  }
23
24  #データセット全てを標本とし，それぞれの標本固有値でAnderson()関数を用いて
25  #信頼区間の上限を導出する.
26  #構成した信頼区間が帰無仮説Hと重なるかグラフで確認するため，
27  #各主成分に対応する標本固有値における信頼区間と帰無仮説の位置関係をプロットする関数.
28  sdev.Onesided <- function(A, df, Delta){
29                              #仮説検定の有意水準A/100のA%，標本として受け取るデータセットdf，
30                              #帰無仮説Hのdeltaを受け取る.
31
32    plot(NULL, xlab="m", ylab="Confidence Interval", xlim=c(1, ncol(df)-1),
33         ylim=c(0,1.1))              #各主成分に対応するように信頼区間を表示する.
34
35    for(m in 1:(ncol(df)-1)){ #Anderson()関数にあるsumがm+1より始まるから，
36                              #p+1を考えないようにするため，-1をしている.
37
38     ci.upper <- Anderson(A, df, m+1) #第m+1主成分に対応する標本固有値でAnderson()関数を呼び出し，
39                                      #信頼区間の上限をci.upperに代入する.
40
41     if(ci.upper >= Delta){
42      points(x=c(m, m), y=c(0, ci.upper), pch=15, cex=0.5, col=2)       #信頼区間に帰無仮説Hが重なれば，
43      segments(x0=m, y0=0, x1=m, y1=ci.upper, col=2)                    #その信頼区間を赤色で描画する.
44     }
45     else{
46      points(x=c(m, m), y=c(0, ci.upper), pch=15, cex=0.5, col=1)  #信頼区間に帰無仮説Hが重ならなければ，
47      segments(x0=m, y0=0, x1=m, y1=ci.upper, col=1)               #その信頼区間を黒色で描画する.
48
49     }
50    }
51
52    abline(h=Delta, lty=2)
53  }
54  library(kernlab)
55  data(spam)
56  library(scar)
57  data(decathlon)
58
59  df1 <- spam[,1:57]            #spamデータ全てのデータセット
60  df2 <- decathlon              #decathlonデータのデータセット
61
62  sdev.Onesided(5, df1, 0.6)  #仮説検定の有意水準A/100を5/100=0.05として，
63                              #標本として受け取るデータセットdfをdf1，
64                              #ここではspamデータ全てのデータセットとして，
65                              #帰無仮説Hのdeltaをdelta=0.6としてsdev.Onesided()関数を呼び出す.
```

1.9 リスト 5 の実行結果と考察

図 26 は，sdev.Onesided(5, df1, 0.6) の実行結果である．つまり，仮説検定の有意水準 A/100 を 5/100=0.05 として，標本として受け取るデータセット df を df1，ここでは spam データ全てのデータセットとして，帰無仮説 H：$\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p} \geq \delta$ を $\delta = 0.6$ として sdev.Onesided() 関数を呼び出している．

図 27 は，sdev.Onesided(5, df2, 0.6) の実行結果である．つまり，仮説検定の有意水準 A/100 を 5/100=0.05 として，標本として受け取るデータセット df を df2，ここでは decathlon データ全てのデータセットとして，帰無仮説 H：$\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p} \geq \delta$ を $\delta = 0.6$ として sdev.Onesided() 関数を呼び出している．

帰無仮説が「棄却されている」のが「黒色」で，『棄却されていない』のが『赤色』の信頼区間として描画されている．横軸に平行な点線より上限が上にあれば棄却されず，下にあれば棄却される．

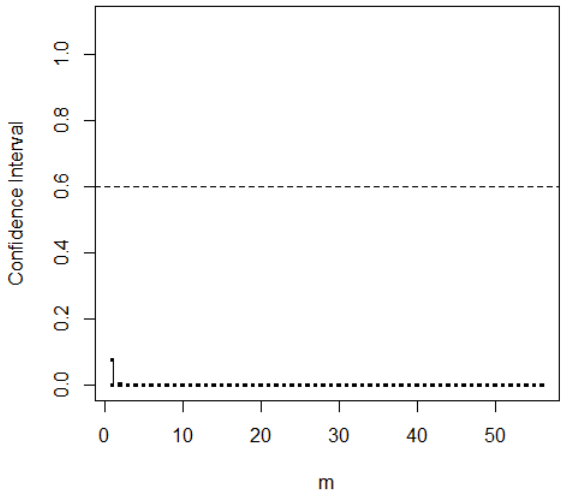

図 26　リスト 5 によるプロット（sdev.Onesided(5, df1, 0.6)）．m には第 $m+1$ 主成分への帰無仮説が対応する．

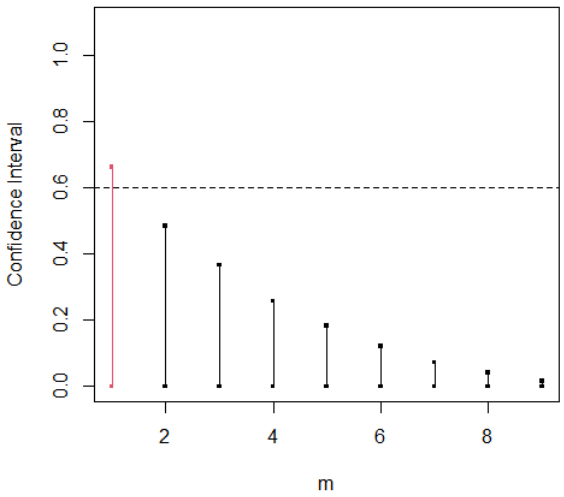

図 27　sdev.Onesided(5, df2, 0.6)．m には第 $m+1$ 主成分への帰無仮説が対応する．

図 26，27 を見ると，固有値一つに対して帰無仮説を設定するような仮説検定において見つかっていた，グラフの上限と下限をデータセットによって設けなければならなかったこと，帰無仮説として設定する値もデータセットによって変えなければならないこと，帰無仮説として設ける基準が大きくなってしまうことといった難点は解決されているように思う．これは先述したとおり，累積寄与率というどのようなデータセットの標本固有値でも $[0,1]$ に収まるような値に対応する仮説検定であるから解決している．

ただし，データセット毎にグラフの上限を設けなくてもいいように，一部の固有値が固有値全体に占める割合を用いたこの仮説検定は，正規性が保たれない spam データについてこれまで以上に適用できていないことがわかる．

図 27 を仮説検定として見ていく．$m=2$ で棄却されていることがわかる．つまり，第 $m+1=2+1=3$ 主成分以降に対応する固有値の和が固有値全体の和に占める割合は $\delta=0.6$ を下回っているということであり，固有値全体の和に占める割合が 0.6 未満になるような固有値たちを主成分として用いる意味が無いという基準の下で第 3 主成分以降の主成分を捨てることができる．[*19]

累積寄与率と帰無仮説 $H:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\geq\delta$，対立仮説 $K:\frac{\lambda_{m+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}<\delta$ の対応を考えると，ここで $\delta=0.6$ と設定しているのは大きすぎることがわかる．$\delta=0.6$ はつまり，「元のデータに対して主成分が保存する情報（分散）は 40% で十分である（活用する主成分以外の主成分たちは元のデータの情報を 60% 未満なら持っていてもいい）」ということである．具体的に図 27 で言えば，「第 2 主成分までを使えば元のデータの 40% の情報は持つから 60% 未満の情報を持つ他の主成分たちは使わなくてもいい」という判断を仮説検定から行うことになる．感覚としては，主成分は元のデータの大体 70% 以上の情報は保って欲しいため，δ としては 0.3 以下で設定するのが適切である．

[*19] ここでは $m=1,2,\ldots$ と考えているが，$m=0$ があった場合，$m=0$ での仮説検定は，帰無仮説 $H:\frac{\lambda_{0+1}+\cdots+\lambda_p}{\lambda_1+\cdots+\lambda_p}\geq 0.6$（第 $m+1=0+1=1$ 主成分への帰無仮説）に対応する．帰無仮説の左辺は 1 に等しいから（第 p 主成分の累積寄与率と同値），棄却されないはずである．実際，上限の値は母集団に対する正規性の仮定や，信頼区間が漸近的なものであることから 1.0 には一致していないとしても，spam データと decathlon データどちらの場合でも 1.0 に近い値が求められる．よって，どのような δ に対しても $m=0$ では棄却されないことがわかる．

図 28 と 29 は，spam データに対して $\delta = 0.05$，0.001 とした実行結果である．
図 30 から 32，は，decathlon データに対して $\delta = 0.2$，0.1，0.05 とした実行結果である．

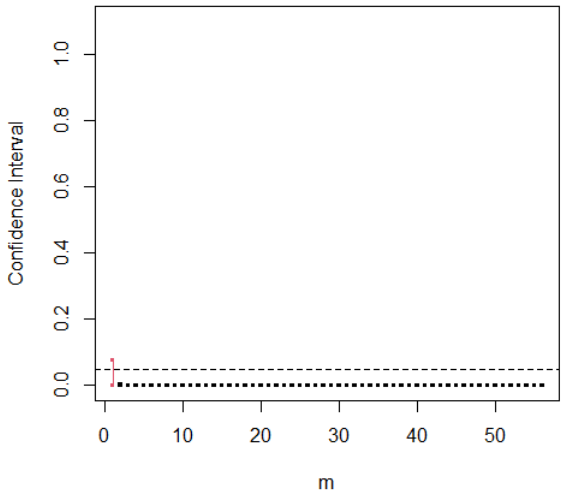

図 28　spam データに対して，$\delta = 0.05$

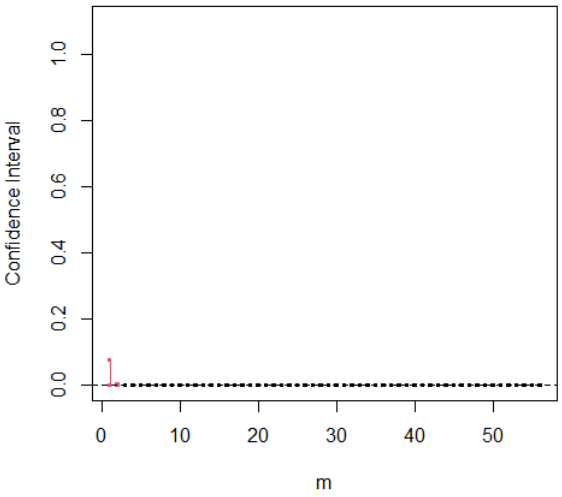

図 29　spam データに対して，$\delta = 0.001$

図 28 より，$\delta = 0.05$ のとき，つまり「spam データの母集団に対して主成分が保存する情報（分散）は 95% で十分である（活用する主成分以外の主成分たちは元のデータの情報を 5% 未満なら持っていてもいい）」というとき，「($m = 1$ で帰無仮説は棄却されるから）第 2 主成分までを使えば元のデータの 95% の情報は持つから 5% 未満の情報を持つ他の主成分たちは使わなくてもいい」という判断を仮説検定から行える．

図 29 より，$\delta = 0.001$ のとき，つまり「spam データの母集団に対して主成分が保存する情報（分散）は 99.9% で十分である（活用する主成分以外の主成分たちは元のデータの情報を 0.1% 未満なら持っていてもいい）」というとき，「($m = 2$ で帰無仮説は棄却されるから）第 3 主成分までを使えば元のデータの 99.9% の情報は持つから 0.1% 未満の情報を持つ他の主成分たちは使わなくてもいい」という判断を仮説検定から行える．

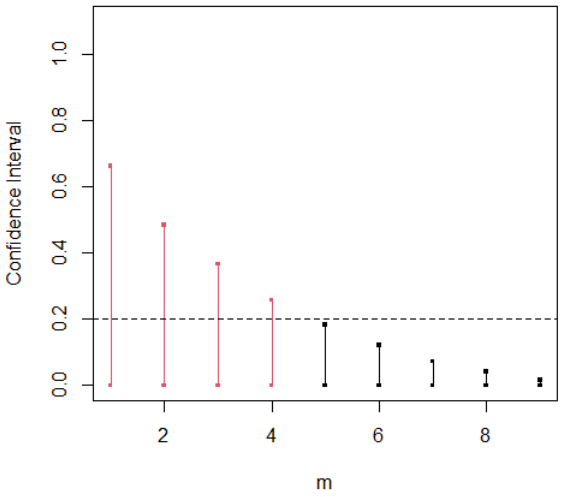

図 30　decathlon データに対して，$\delta = 0.2$

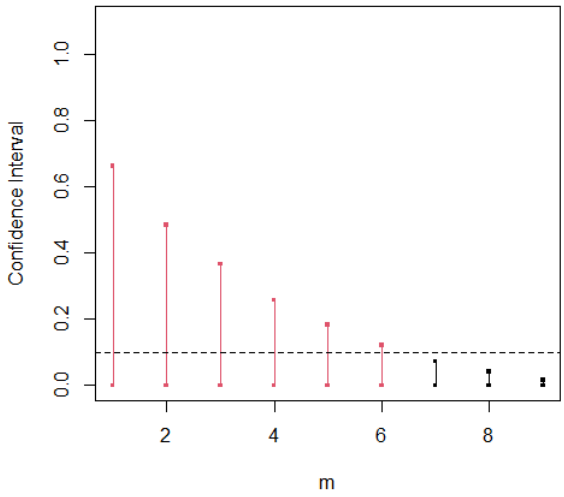

図 31　decathlon データに対して，$\delta = 0.1$

図 30 より，$\delta = 0.2$ のとき，つまり「decathlon データの母集団に対して主成分が保存する情報（分散）は 80% で十分である（活用する主成分以外の主成分たちは元のデータの情報を 20% 未満なら持っていてもいい）」というとき，「($m = 4$ で帰無仮説は棄却されるから）第 5 主成分までを使えば元のデータの 80% の情報は持つから 20% 未満の情報を持つ他の主成分たちは使わなくてもいい」という判断を仮説検定から行える．

図 31 より，$\delta = 0.1$ のとき，つまり「decathlon データの母集団に対して主成分が保存する情報（分散）は 90% で十分である（活用する主成分以外の主成分たちは元のデータの情報を 10% 未満なら持っていてもいい）」というとき，「($m = 5$ で帰無仮説は棄却されるから）第 6 主成分までを使えば元のデータの 90% の情報は持つから 10% 未満の情報を持つ他の主成分たちは使わなくてもいい」という判断を仮説検定から行える．

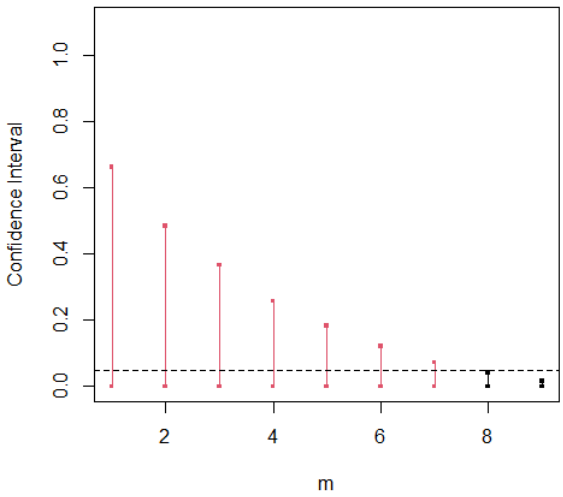

図 32　decathlon データに対して，$\delta = 0.05$

図 32 より，$\delta = 0.05$ のとき，つまり「decathlon データの母集団に対して主成分が保存する情報（分散）は 95% で十分である（活用する主成分以外の主成分たちは元のデータの情報を 95% 未満なら持っていてもいい）」というとき，「（$m = 7$ で帰無仮説は棄却されるから）第 8 主成分までを使えば元のデータの 95% の情報は持つから 5% 未満の情報を持つ他の主成分たちは使わなくてもいい」という判断を仮説検定から行える．