データ解析特論 第 3 回　　[2017/10/30 201720690 小松 弘人]

1.

ファイル data-2D.txt にあるデータは，2 次元空間における平均ベクトル $(0,0)$，分散共分散行列 $\begin{bmatrix} 1^2 & 0 \\ 0 & 3^2 \end{bmatrix}$ の 2 変数正規分布に従って発生させた乱数の分布を，角度 θ だけ左に回転したものである．

(a) 主成分分析を用いて，回転角 θ を推定せよ．

(b) データを主軸上に射影し，1 次元のヒストグラムとして示せ．

(c) データの座標系 (x, y) における x 成分と y 成分の間の相関係数を求めよ

(d) データの第 1 主成分 p，第 2 主成分 q の座標系を用いて表したのち，再度両成分の間の相関係数を求めよ．

スクリプト

```
### 1. data-2d.のデータを主成分分析txt
# データ読み込み
dat <- read.csv("data-2d.txt", header=F)
dat.pca <- prcomp(dat, center = T, retx = T)

# グラフ描画
{
plot(c(), xlim=c(-10, 10), ylim=c(-10, 10), type="n")
points(dat[,1], dat[,2])
points(dat.pca$x[,1], dat.pca$x[,2], col="red")
points(-dat.pca$x[,2], dat.pca$x[,1], col="blue")
}

## (a) 回転角を推定
((atan2(dat.pca$rotation[2,1], dat.pca$rotation[1,1])/pi*180)-90)

## (b) データを主軸上に射影
hist(dat.pca$x[,1], breaks=40)

## (c) (x,y座標系における)x,成分の相関係数y
cor(dat[,1], dat[,2])

## (d) (p,q座標系における)p,成分の相関係数q
cor(dat.pca$x[,1], dat.pca$x[,2])
```

結果

data-2D.txt のデータの散布図を図 1，主成分分析を行い第 1, 第 2 主成分面へ射影したデータの散布図を図 2，分散共分散行列から推測される元データの散布図を図 3, これらを全て重ねたグラフを図 4 に示す．

また，各小問に対応するスクリプトを実行した結果を以下に示す．図 5 は (b) で表示したヒストグラムである．

```
> ## (a) 回転角を推定
> ((atan2(dat.pca$rotation[2,1], dat.pca$rotation[1,1])/pi*180)-90)
```

```
[1] 57.37136
> ## (b) データを主軸上に射影
> hist(dat.pca$x[,1], breaks=40)
> ## (c) (x,y) 座標系における x,y 成分の相関係数
> cor(dat[,1], dat[,2])
[1] -0.6301858
> ## (d) (p,q) 座標系における p,q 成分の相関係数
> cor(dat.pca$x[,1], dat.pca$x[,2])
[1] 7.195328e-17
```

考察

図 2 は，データの分布の長軸が x 軸上にあるが，乱数の確率密度関数の分散共分散行列から考えると，長軸が y 軸上にあるのが自然である．したがって，元の散布図は図 3 であったと考えられる．

回転角 θ は，`dat.pca$rotation` の回転行列から求められる．`dat.pca$rotation` の回転行列は，図 1 から図 2 への回転を表している．しかし，求める回転角は，図 3 から図 1 へ回転させた角度である．したがって，回転行列の 1 行 1 列成分および 2 行 1 列成分から角度を求め，さらに，図 2 から図 3 への回転角 90° を引く．よって，回転角 $\theta = 57.37136°$ と推測される．

(x, y) 座標系における x, y 成分の相関係数は -0.6301858，(p, q) 座標系における p, q 成分の相関係数は $7.195328e{-}17$ となった．(p, q) 座標系における p, q 成分の相関係数は，乱数の確率密度関数の相関係数とほぼ一致している．一方で，(x, y) 座標系における x, y 成分の相関係数が -0.6301858 となったのは回転

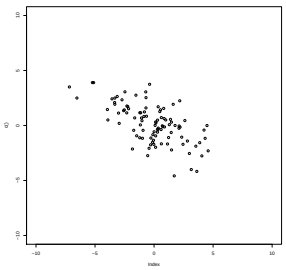

図 1: data-2D のデータ

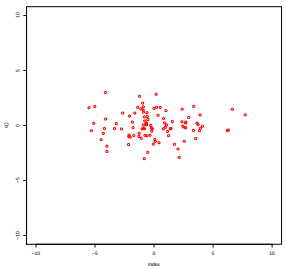

図 2: 第 1 第 2 主成分面への射影

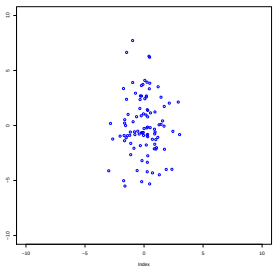

図 3: 回転前のデータ (推定)

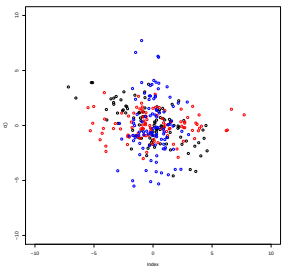

図 4: 3 つの散布図を重ねたグラフ

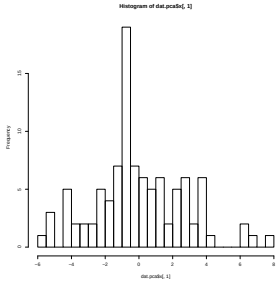

図 5: (b) の実行結果

の影響であると考えられる．

2.

5 次元の特徴空間に分布するデータ (ファイル data-5D.txt) を可視化したい．主成分分析を用いて第 1，第 2 主成分が張る部分空間に射影した後，散布図で示せ．

スクリプト

```
dat <- read.csv("data-5d.txt", header=F)
dat.pca <- prcomp(dat, center = T, retx = T)
plot(dat.pca$x[,1], dat.pca$x[,2])
```

結果

散布図を図 6 に示す．

考察

データは，3 つのグループに分けることができると考えられる．よって，乱数は 3 つの確率密度関数を合わせた確率密度関数から生成されたと考えられる．

3.

A さんが提案するアルゴリズムでは，既存手法に対して平均的な処理時間は確かに短縮されているように思われるのだが，既存手法とは大きくは変わらない．微小な差を有意に示すためにはどのようにすればいいか．ばらつきのある処理時間のデータを人工的に作成して，2 標本を t 検定 (平均値の差に関する検定) を使って示せ．また，検定の信頼性が低下するのはどのようなときか，考察せよ．

スクリプト

以下のスクリプトの `N`, `exist.mean`, `exist.sd`, `prop.mean`, `prop.sd` を変更しながら実行する．

```
N = 10000
exist.mean = 15; exist.sd = 1; prop.mean = 15; prop.sd = 1;
exist.time <- rnorm(N, mean=exist.mean, sd=exist.sd)
prop.time <- rnorm(N, mean=prop.mean, sd=prop.sd)
(res = t.test(exist.time, prop.time,
```

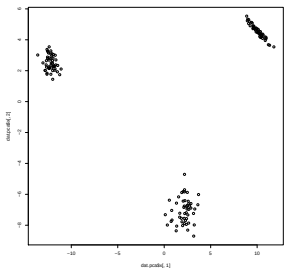

図 6: 散布図

```
        alternative="two.sided", var.equal=T, conf.level=0.95))
# 有意水準をとる5%
(res$p.value < 0.05) # 帰無仮説を棄却してもよいか？有意差がある（差がある）か？
```

結果

2 つの分散は等しいと仮定し，95% 信頼区間の両側検定を行う．帰無仮説は，“平均は等しい (差が 0)” である．パラメータを変更しながら，この帰無仮説が棄却できるか否かを調べた．結果を，以下に示す．

1. 既存手法 ($\mu = 15.5, \sigma^2 = 1^2$), 提案手法 ($\mu = 15, \sigma^2 = 1^2$) のとき，
   - N=10 のとき，FALSE
   - N=100 のとき，TRUE
   - N=1000 のとき，TRUE
   - N=10000 のとき，TRUE
   - N=100000 のとき，TRUE
2. 既存手法 ($\mu = 15.1, \sigma^2 = 1^2$), 提案手法 ($\mu = 15, \sigma^2 = 1^2$) のとき，
   - N=10 のとき，FALSE
   - N=100 のとき，FALSE
   - N=1000 のとき，FALSE
   - N=10000 のとき，TRUE
   - N=100000 のとき，TRUE
3. 既存手法 ($\mu = 15.1, \sigma^2 = 5^2$), 提案手法 ($\mu = 15, \sigma^2 = 5^2$) のとき，
   - N=10 のとき，FALSE
   - N=100 のとき，FALSE
   - N=1000 のとき，FALSE
   - N=10000 のとき，FALSE
   - N=100000 のとき，TRUE

考察

1. では，2 つの平均を比較的大きくした．その結果，サンプル数が 100 であっても平均値が違うことを推定できた．2. では，2 つの平均の差を小さくした．その結果，サンプル数が 10000 あれば平均値が違うことを推定できた．3. では，2 つの分散を大きくした．その結果，サンプル数が 100000 あれば平均値が違うことを推定できた．これらの結果から，サンプル数が少ないときや，データの分散が大きいときは検定の信頼性が低下すると考えられる．