Completion and Outlier detection

2021-09-13

Completion

Delaigle 세팅 - out_type = 3 (지난주에 보여드린 결과), 20번 반복

```
### Out X
  Method  PVE Reconstruction  Completion Eigenfunction Eigenfunction_cos
     Yao 1.00     0.11 (0.01) 0.19 (0.03)   0.06 (0.04)       0.56 (0.01)
   Kraus 0.91           <NA> 0.47 (0.28)   0.08 (0.04)       0.57 (0.02)
   Huber 0.89     0.17 (0.02) 0.40 (0.11)   0.04 (0.02)       0.56 (0.01)
  Boente 1.00     0.19 (0.03) 0.42 (0.13)   0.13 (0.03)       0.59 (0.01)
    Mest 0.92     0.16 (0.02) 0.34 (0.12)   0.11 (0.05)       0.58 (0.02)
 Mest-sm 0.98     0.13 (0.02) 0.26 (0.08)   0.09 (0.05)       0.58 (0.02)
      GK 0.93     0.16 (0.02) 0.36 (0.11)   0.12 (0.07)       0.59 (0.02)
   GK-sm 0.99     0.14 (0.02) 0.28 (0.08)   0.10 (0.07)       0.58 (0.02)
   PM-NA 0.96     0.14 (0.02) 0.30 (0.09)   0.09 (0.04)       0.57 (0.02)
PM-sm-NA 0.98     0.13 (0.02) 0.25 (0.07)   0.08 (0.04)       0.57 (0.02)
   PM-Im 0.96     0.14 (0.02) 0.31 (0.09)   0.09 (0.06)       0.58 (0.02)
PM-sm-Im 0.98     0.13 (0.02) 0.26 (0.08)   0.08 (0.06)       0.57 (0.02)


### Out O
  Method  PVE Reconstruction  Completion Eigenfunction Eigenfunction_cos
     Yao 0.98     0.42 (0.32) 0.82 (0.64)   0.84 (0.51)       0.78 (0.13)
   Kraus 0.95           <NA> 1.26 (0.70)   0.94 (0.43)       0.82 (0.11)
   Huber 0.89     0.19 (0.04) 0.43 (0.12)   0.06 (0.06)       0.56 (0.02)
  Boente 1.00     0.19 (0.04) 0.40 (0.17)   0.13 (0.04)       0.59 (0.02)
    Mest 0.89     0.18 (0.03) 0.40 (0.14)   0.14 (0.07)       0.60 (0.03)
 Mest-sm 0.95     0.16 (0.03) 0.33 (0.11)   0.13 (0.07)       0.59 (0.03)
      GK 0.91     0.17 (0.03) 0.39 (0.14)   0.17 (0.15)       0.60 (0.05)
   GK-sm 0.97     0.15 (0.03) 0.32 (0.12)   0.15 (0.15)       0.59 (0.05)
   PM-Im 0.93     0.16 (0.03) 0.36 (0.14)   0.17 (0.18)       0.60 (0.06)
PM-sm-Im 0.96     0.15 (0.03) 0.31 (0.13)   0.16 (0.18)       0.60 (0.06)


## Out O - 50번 반복
  Method  PVE Reconstruction  Completion Eigenfunction Eigenfunction_cos
    Mest 0.89     0.17 (0.03) 0.41 (0.15)   0.14 (0.08)       0.59 (0.03)
 Mest-sm 0.94     0.16 (0.03) 0.35 (0.12)   0.13 (0.08)       0.59 (0.03)
      GK 0.93     0.17 (0.04) 0.40 (0.15)   0.15 (0.11)       0.60 (0.04)
   GK-sm 0.97     0.16 (0.03) 0.36 (0.14)   0.14 (0.11)       0.59 (0.04)
      PM 0.92     0.16 (0.03) 0.39 (0.13)   0.15 (0.10)       0.60 (0.04)
   PM-sm 0.94     0.15 (0.03) 0.33 (0.12)   0.14 (0.10)       0.60 (0.04)
   PM-NA 0.94     0.15 (0.03) 0.35 (0.11)   0.13 (0.09)       0.59 (0.03)
PM-sm-NA 0.96     0.14 (0.03) 0.30 (0.10)   0.12 (0.09)       0.59 (0.03)
   PM-Im 0.94     0.15 (0.03) 0.32 (0.11)   0.12 (0.10)       0.59 (0.04)
```

```
PM-sm-Im 0.96    0.14 (0.02) 0.28 (0.10)   0.12 (0.10)       0.58 (0.04)
```

Kraus 세팅 - out_type = 2, sig = 0.01, 20번 반복

```
 Method  PVE Reconstruction      Completion Eigenfunction Eigenfunction_cos
    Yao 0.96    0.81 (2.16)     0.76 (1.50)   1.69 (0.30)       0.96 (0.05)
  Kraus 0.91           <NA>  139.71 (622.16)   1.77 (0.12)       0.99 (0.02)
 Boente 1.00    0.04 (0.01)     0.11 (0.04)   0.80 (0.06)       0.76 (0.02)
   Mest 0.83    0.04 (0.00)     0.06 (0.01)   0.66 (0.11)       0.72 (0.02)
Mest-sm 0.98    0.03 (0.00)     0.05 (0.01)   0.33 (0.16)       0.65 (0.04)
     GK 0.86    0.03 (0.00)     0.05 (0.01)   0.46 (0.13)       0.68 (0.03)
  GK-sm 0.98    0.02 (0.00)     0.04 (0.01)   0.27 (0.15)       0.63 (0.04)
     PM 0.88    0.03 (0.00)     0.04 (0.01)   0.52 (0.13)       0.69 (0.03)
  PM-sm 0.98    0.02 (0.00)     0.03 (0.00)   0.18 (0.12)       0.60 (0.03)
```

Outlier detection

- 이전 결과들이 모두 Boente 세팅으로만 요약되어 있어, Delaigle 세팅에도 적용하여 결과 확인
- `sensitivity`가 outlier를 얼마나 잘 detect했는지를 나타냄
- 여기서 `PM`, `PM-sm`은 이전 파일에서 `PM-NA`, `PM-sm-NA`를 나타냄
  - `PM`에 imputation을 적용한 방법은 completion에서의 결과가 좋긴 했지만, 따로 reference한 방법이 아닌 제 생각에 좋을 것 같아서 해본 방법이었으며, 이론적으로 적절한 방법인지에 대해서는 의문임
- 비교 방법론
  1. `PC1-adjbox` : 1st PC score에 adjusted boxplot
  2. `PC1-box` : 1st PC score에 boxplot
  3. `SD-adjbox` : Score distance에 adjusted boxplot
  4. `SD-box` : Score distance에 boxplot
  5. `robMah-comp` : Completion된 dense data에 `rainbow::foutliers(..., method = "robMah")` 함수로 outlier detection
  - 함수의 과정은 다음과 같음
    1. dense data를 robust PCA한 후의 PC score를 계산
    2. 계산한 PC score로 robust location, cov 계산 (여기서는 MCD estimator 사용)
    3. 2에서 계산한 estimate을 이용하여 Mahalanobis dist 계산하고 이에 대해 boxplot을 적용하여 outlier를 detect
  6. `robMah` : K PC score들로 계산한 robust Mahalanaobis distance를 boxplot으로 outlier detection
  - 이미 계산한 PC score(completion X)에 `robMah-comp`의 2~3번 과정을 적용
  7. `outmap` : Outlier map (Score distance vs Orthogonal distance) plot에서 1사분면에 속하는 데이터를 outlier로 detect

Boente 세팅 - model = 4

```
$Yao
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.004  0.179       0.862  0.875  0.257
specififity          1       1     0.968  1.000       1.000  1.000  0.998

$Mest
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.182   0.83       0.886  0.885  0.998
specififity          1       1     0.869   1.00       1.000  1.000  0.992

$GK
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
```

```
sensitivity          0       0     0.270  0.879       0.894  0.888  0.997
specififity          1       1     0.852  1.000       1.000  1.000  0.994

$PM
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.281  0.865       0.901  0.874  1.000
specififity          1       1     0.868  1.000       1.000  1.000  0.992
```

Delaigle 세팅 - out_type = 2

- 여기서의 outlier 세팅이 heavy-tailed distribution으로부터 noise를 생성하다보니, 이 값이 큰 경우도 있고 작은 경우도 존재
- 따라서 noise가 작은 경우는 PC score 등이 차이가 크지 않아서 outlier로 detect하지 못하는 것으로 보임

```
$Yao
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.142  0.140       0.062  0.288  0.105
specififity          1       1     0.976  0.983       0.981  0.990  0.998

$Mest
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.138  0.100       0.063  0.095  0.168
specififity          1       1     0.989  0.989       0.985  0.982  0.978

$GK
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.157  0.107       0.070  0.103  0.143
specififity          1       1     0.985  0.988       0.985  0.985  0.988

$PM
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.130  0.102       0.077  0.097  0.155
specififity          1       1     0.985  0.988       0.985  0.985  0.985
```

Delaigle 세팅 - out_type = 6(Boente 세팅의 model = 2)

- outlier generating을 Boente 세팅처럼 할 경우에는 Boente에서와 비슷한 형태를 보임
- 이 세팅은 앞의 Boente 세팅(model = 4)보다 spike 정도가 적은 세팅임

```
$Yao
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.323  0.917        0.72  0.960  0.812
specififity          1       1     0.892  1.000        1.00  0.998  1.000

$Mest
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.283  0.978       0.693  0.975  0.997
specififity          1       1     0.854  1.000       1.000  1.000  0.989

$GK
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.343  0.975        0.65  0.973  0.998
specififity          1       1     0.867  1.000        1.00  1.000  0.995

$PM
```

```
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.270  0.988       0.663  0.973  1.000
specififity          1       1     0.854  1.000       1.000  1.000  0.993
```

Delaigle 세팅 - out_type = 7(Boente 세팅의 model = 4)

```
$Yao
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.743      1       0.983      1   0.59
specififity          1       1     0.857      1       1.000      1   1.00

$Mest
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.402  0.988       0.973  0.988  0.998
specififity          1       1     0.850  1.000       1.000  1.000  0.988

$GK
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.427   0.99       0.975   0.99  0.997
specififity          1       1     0.865   1.00       1.000   1.00  0.995

$PM
            PC1-adjbox PC1-box SD-adjbox SD-box robMah-comp robMah outmap
sensitivity          0       0     0.445   0.99        0.98  0.988  0.998
specififity          1       1     0.855   1.00        1.00  1.000  0.993
```