The `dimension` Package

Wenlan Zang
wenlan.zang@yale.edu,

Michael.J.Kane
michael.kane@yale.edu

December 9, 2019

1 Introduction

The `dimension` package provides an efficient way to determine the dimension f a signal rich subspace in a large matrix. It also provides a cleaned estimator of the original matrix and correlation matrix. Source code is maintained at https://github.com/WenlanzZ/dimension.

The `dimension` package estimates the intrinsic dimension of a signal-rich subspace in large matrix "real- and complex value dense R matrices and real-valued saprse matrices from the `Matrix` package") by decomposing matrix into a signal-plus-noise space and approximate the signal-rich subspace with a rank K approximation $\hat{X} = \sum_{k=1}^{K} d_k u_k {v_k}^T$. To estimate rank K, it follows a simple procedure assuming that matrix X is composed of a low-rank signal matrix S and an average general noise random matrix $\bar{N}$. It has been shown that the average eigenvalues of random matrices N follows a universal Marcěnko-Pastur (MP) distribution. We hypothesize that the deviation of eigenvalues of X from the MP distribution indicates the intrinsic dimension of signal-rich subspace.

The package included the following main functions:

- subspace() - Greate a subspace class with scaled eigenvalue and eigenvectors and simulated noise eigenvalues for specified ranks.
- print.subspace()- Get a brief summary of subspace class.
- plot.subspace() - Get the scree plot of subspace class.
- dimension() - Get the dimension of a signal-subspace in a large high-dimensional matrix.
- clipped() - Get a cleaned estimator of the original matrix, its covairance matrix and correlation matrix.

- modified_legacyplot() - Produces modified summary plots of bcp() output.

A demostration of the main functions and with a brief sample is as follow.

2 Subspace

Let $X \in \mathbf{R}^{n \times p}$ be a simulated multivariate normal matrix with *ncc* correlated columns.

```
> library('dimension')
> X <- Xsim(n = 150, p = 100, ncc = 30, var = c(rep(10,5),rep(3,25)))
> t1 <- proc.time()
> Subspace <- subspace(X, rank = 1:50, times = 10, basis = "eigen")
> print(proc.time() - t1)
   user  system elapsed
  0.946   0.293   0.767
> gc()
  used  (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
Ncells 9590962 512.3  17937651 958.0       NA 16475288 879.9
Vcells 16208511 123.7 27151251 207.2    16384 22559284 172.2
> plot(Subspace, annotation = 30)
```

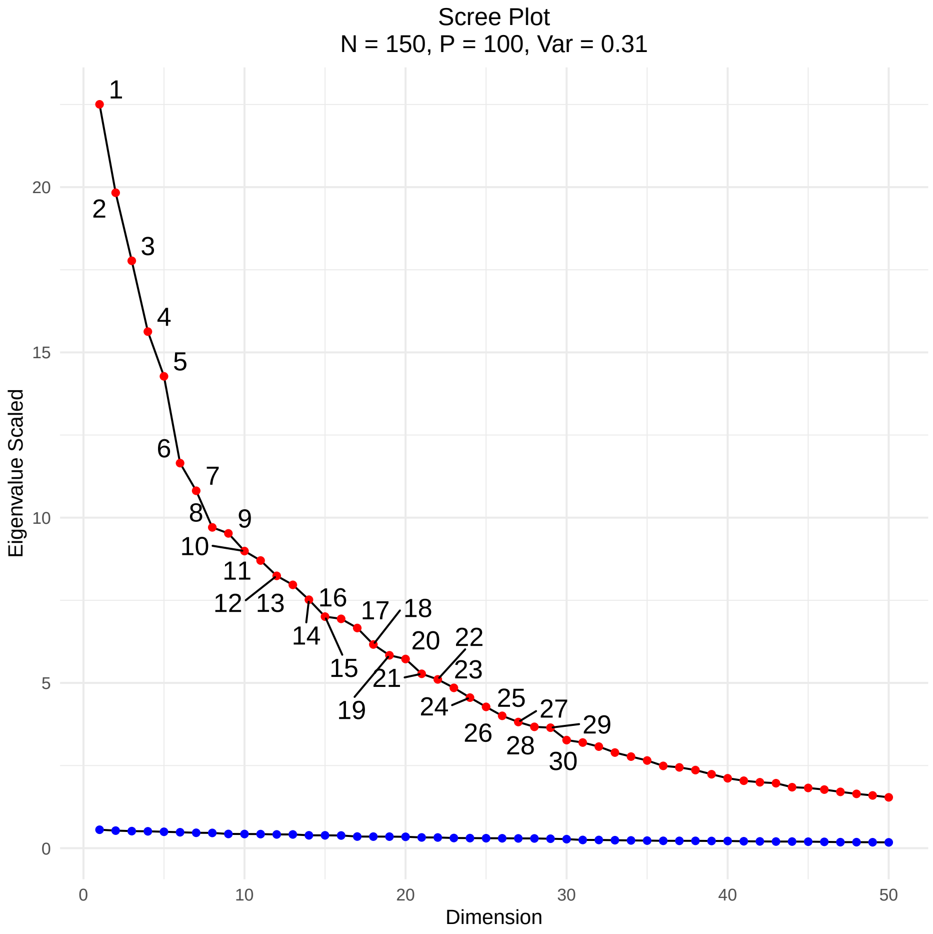
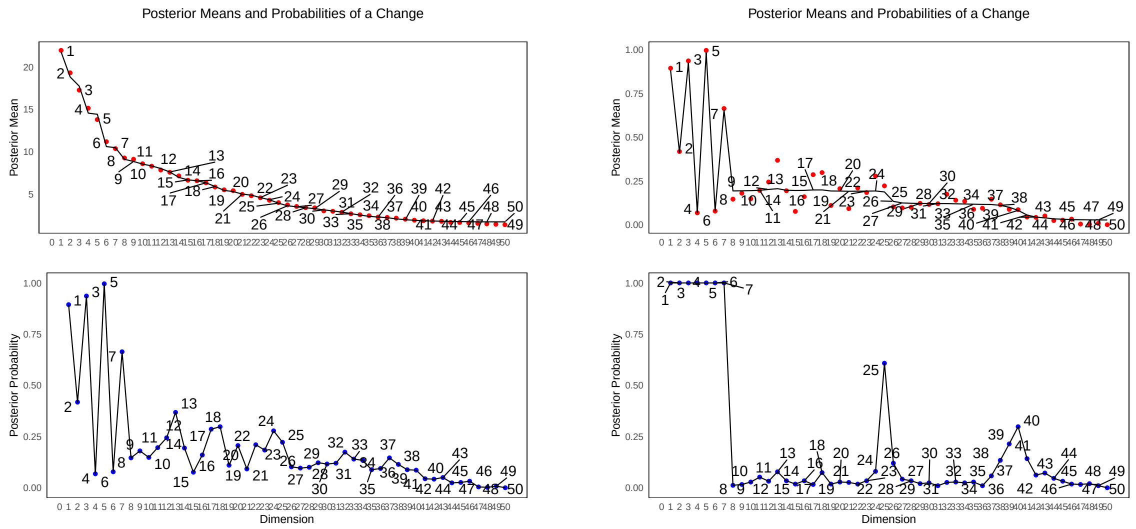


Figure 1

```
> t1 <- proc.time()
> results <- dimension(subspace_ = Subspace)
# equivelantly, if subsapce has not been calcualted
> results <- dimension(X, rank = 1:50, times = 10, basis="eigen")
> print(proc.time() - t1)
   user  system elapsed
  0.125   0.014  10.654
> gc()
 used  (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
Ncells 9403721 502.3 17937651 958.0       NA 17937651 958.0
Vcells 15834973 120.9 27151251 207.2    16384 22559284 172.2
> str(results)
> plot(results$Subspace,
       Changepoint = results$Changepoint$dimension,
       annotation = 30)
```

```
> modified_legacyplot(results$Changepoint$bcp_irl, annotation = 50)
> modified_legacyplot(results$Changepoint$bcp_post, annotation = 50)
```

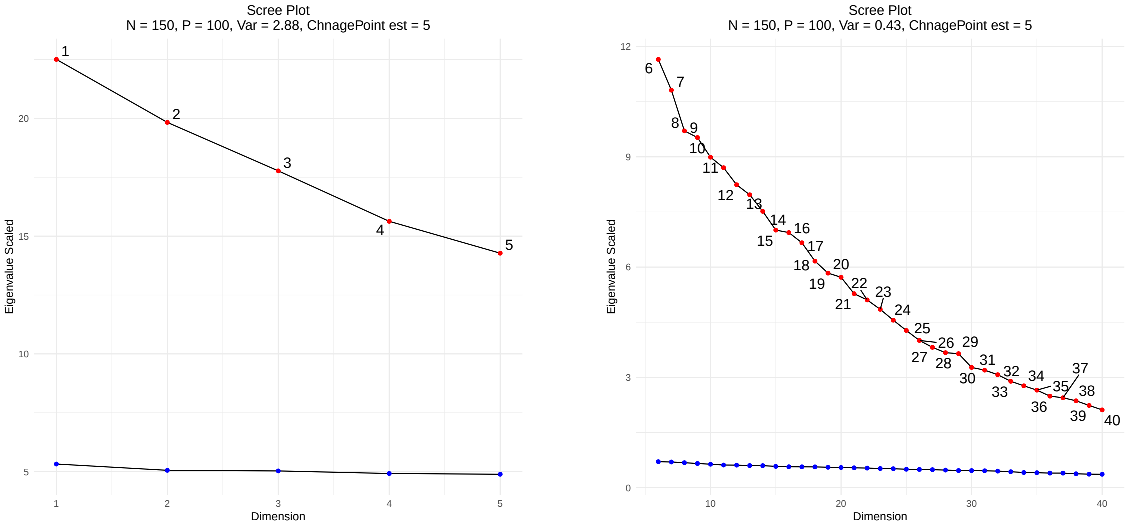


(a) label 1

(b) label 2

Figure 2: bcp

```
> t1 <- proc.time()
> TopSubspace <- subspace(X, rank = 1:5, times = 10, basis = "eigen")
> TopSubspace
  An object of class subspace within X matrix with 150 samples and 100 features.
  Estimated rank range from 1 to 5
> MidSubspace <- subspace(X, rank = 6:40, times = 10, basis = "eigen")
  An object of class subspace within X matrix with 150 samples and 100 features.
  Estimated rank range from 6 to 40
> print(proc.time() - t1)
   user  system elapsed
  0.974   0.292   0.788
> gc()
  used  (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
Ncells  9604856 513.0  17937651 958.0        NA 17937651 958.0
Vcells 16231491 123.9 27151251 207.2    16384 22559284 172.2
> plot(TopSubspace, Changepoint = results$Changepoint$dimension, annotation = 5)
> plot(MidSubspace, Changepoint = results$Changepoint$dimension, annotation = 40)
```



(a) label 1

(b) label 2

Figure 3: bcp