Short Introduction to the Usage of Pakage MAVE

Weiqiang Hang Hongfan Zhang Yingcun Xia
National University of Singapore, Singapore

May 8, 2019

1 Introduction

Package MAVE provides several methods, including MAVE and OPG methods proposed by [4, 5, 6], to find the central space (CS) and the central mean space (CMS). It also implements sliced inverse regression of a kernel version; see [1, 4]. Formal definition of the central space and the central mean space can be found in [2, 3]. For comparison, a package dr in CRAN also contains other sufficient dimension reduction methods[7].

The main part of package MAVE is written in C++ based on RcppArmadillo package. If there is any problem during installation, please update your Rcpp package and install RcppArmadillo package and try again.

2 Usage

The primary function in this package is MAVE. The input arguments include an $n \times p$ covariate matrix X, an $n \times 1$ respond matrix Y, and the method argument for dimension reduction. The options for the method argument are 'csopg', 'csmave','meanopg','meanmave' and 'ksir', and the default is 'csopg'. 'csopg' and 'csmave' are methods of finding CS by OPG and MAVE respectively, 'meanopg' and 'meanmave' are methods of finding CMS by OPG and MAVE, 'ksir' is the sliced inverse regression of kernel version. Since OPG method is time-saving compared with MAVE and the result of OPG is as good as that of MAVE, we recommend using OPG. Argument `max.dim sets the maximum dimension for mave to compute. The default value is 10 meaning that it will only calculate dimension reduction spaces of dimension up to 10. It will help user save computation time when the data is high dimensional. We will use examples to illustrate the usage of the package.`

```
> library(MAVE)
> data(Concrete)
> set.seed(1234)
> train <- sample(1:1030)[1:500]
> x.train <- as.matrix(Concrete[train, 1:8])
```

```
> y.train <- as.matrix(Concrete[train, 9])
> x.test  <- as.matrix(Concrete[-train, 1:8])
> y.test  <- as.matrix(Concrete[-train, 9])
> dr.mave <- mave(y.train~x.train,method = 'MEANOPG', max.dim = 8)
> dr.mave

Call:
mave(formula = y.train ~ x.train, method = "MEANOPG", max.dim = 8)

central mean space  of dimensions  1 2 3 4 5 6 7 8  are computed
```

The object returned by mave or mave.compute contains information of call, data and basis matrices of dimension reduction spaces with different dimensions. The basis matrix of a given dimension, 2, for example, can be obtained by

```
> dir2 <- coef(dr.mave, dim = 2)
```

Then the reduced data is obtained from original data multiplied by the basis matrix of dimension reduction space. The reduced data can be calculated by mave.data. The following is an example to apply mars in pacakge mda to the reduced data for prediction.

```
> library(mda)
> x.train.mave <- mave.data(dr.mave, x = x.train , dim = 2)
> x.test.mave <- mave.data(dr.mave,x = x.test, dim = 2)
> model.mars <- mars(x.train.mave, y.train, degree=2)
> y.pred.mars <- predict(model.mars, x.test.mave)
> mean((y.pred.mars - y.test) ^ 2)

[1] 113.0373
```

For convenience, the package provides predict to implement the above procedure with some modifications. The argument degree will be passed to mars which specifies the maximum interaction degree. More arguments like thresh or penalty can be passed to mars by placing them after dim in the predict method.

```
> y.pred <- predict(dr.mave, newx = x.test, dim = 2, degree = 2)
> mean((y.pred - y.test) ^ 2)

[1] 88.48931
```

In MAVE package of version 1.3.8, mave function allows mutiple reponse, which means that argument y can a $n \times q$ matrix. mave.dim implements the selection of dimension of the CS or CMS discussed in section ??. It returns an object with additional information of cross-validation values of different dimensions. Below is a simple example to illustrate its usage.

```
> set.seed(12345)
> n=800
> x <- matrix(rnorm(n*5), n, 5)
> beta1 <- matrix(c(0.717,0.717,0,0,0))
> beta2 <- matrix(c(0,0,0.717,0.717,0))
> beta3 <- matrix(c(0,0,0,0,1))
> err1 <- matrix(rnorm(n))
> err2 <- matrix(rnorm(n))
> y1 <- as.matrix((x %*% beta1) / (1 + 2 * (x %*% beta2) ^ 2) + (x %*% beta3) * err1)
> y2 <- as.matrix((x%*%beta3)^2) + err2
> y = cbind(y1,y2)
> dr.mave <- mave(y~x, method = 'CSOPG')
> dr.mave.dim <- mave.dim(dr.mave)
> dr.mave.dim

Call:
mave.dim(dr = dr.mave)

The cross-validation is run on dimensions of 1 2 3 4 5
Dimension        1         2         3         4         5
CV-value        0.26         0.24         0.25         0.26         0.27

The selected dimension of  central space  is  2
```

The code below can be used to find the selected dimension with minimum cross-validation value.

```
> which.min(dr.mave.dim$cv)

[1] 2
```

From the result, the estimated dimension reduction space with dimension 3 has the minimum cross-validation value. The estimated basis vectors of CS of dimension 3 can be accessed by coef method. From the result, the estimated basis vector falls in the linear space generated by $(\beta_1, \beta_2, \beta_3)$ with small deviation. Although in this example, the estimated basis vectors are close to $(\beta_1, \beta_2, \beta_3)$, we should note that the original basis vectors like $(\beta_1, \beta_2, \beta_3)$ are unidentifiable, only the space generated by the the original basis vectors is identifiable.

```
> coef(dr.mave,dim=3)

           dir1        dir2        dir3
x1 -0.002290743  0.71487066 -0.12210258
x2  0.005650579  0.69502519 -0.01138889
x3 -0.006105577 -0.07085736 -0.63225614
x4  0.034851445 -0.02713024 -0.76417882
x5  0.999355253  0.01196297  0.03527250
```

In MAVE package of version 1.3.8, we use screening method to select import variables in high dimensional data. The default number of variables retained after screening is $n/\log(n)$. The following is an example about it.

```
> set.seed(12345)
> n <- 200
> p <- 500
> x <- matrix(rnorm(n*p), n, p)
> y <- x[,1]+x[,2]+rnorm(n)
> dr.mave <- mave(y~x, method = 'MEANOPG')
```

screening method is using to select import variables.

```
> dr.mave.dim <- mave.dim(dr.mave)
> dr.mave.dim

Call:
mave.dim(dr = dr.mave)

The cross-validation is run on dimensions of 1 2 3 4 5 6 7 8 9 10
Dimension        1         2         3         4         5         6         7         8
CV-value        0.73         0.62         0.63         0.74         0.87         0.92

The selected dimension of  central mean space  is  2
```

References

[1] Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414), 316-327.

[2] Cook, R.D.(1998),*Regression Graphics*. New York: Wiley

[3] Cook, R. D., and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 455-474.

[4] Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 363-410.

[5] Xia, Y. (2007) A constructive approach to the estimation of dimension reduction directions. *Annals of Statistics*, 35(3), 2654-2690

[6] Wang, H., and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482), 811-821.

[7] Weisberg, S. (2002). Dimension reduction regression in R. Journal of Statistical Software, 7(1), 1-22.