Basis-Adaptive Selection Algorithm in dr-package

by Jae Keun Yoo

Abstract Sufficient dimension reduction (SDR) turns out to be a useful dimension reduction tool in high-dimensional regression analysis. Weisberg (2002) developed the **dr**-package to implement the four most popular SDR methods. However, the package does not provide any clear guidelines as to which method should be used given a data. Since the four methods may provide dramatically different dimension reduction results, the selection in the **dr**-package is problematic for statistical practitioners. In this paper, a basis-adaptive selection algorithm is developed in order to relieve this issue. The basic idea is to select an SDR method that provides the highest correlation between the basis estimates obtained by the four classical SDR methods. A real data example and numerical studies confirm the practical usefulness of the developed algorithm.

Introduction

Sufficient dimension reduction (SDR) in the regression of $y \in \mathbb{R}^1 | \mathbf{X} \in \mathbb{R}^p = (x_1, \ldots, x_p)^{\mathrm{T}}$ replaces the original p-dimensional predictors $\mathbf{X}$ with its lower-dimensional linearly transformed predictors $\mathbf{M}^{\mathrm{T}}\mathbf{X}$ without any loss of information on $y|\mathbf{X}$, which is equivalently expressed:

$$y \amalg \mathbf{X} | \mathbf{M}^{\mathrm{T}}\mathbf{X}, \tag{1}$$

where $\amalg$ stands for independence, $\mathbf{M}$ is a $p \times q$ matrix and $q \leq p$.

A space spanned by the columns of $\mathbf{M}$ to satisfy (1) is called a dimension reduction subspace. Hereafter, $\mathcal{S}(\mathbf{M})$ denotes the column subspace of a $p \times q$ matrix $\mathbf{M}$, and $\mathbf{M}^{\mathrm{T}}\mathbf{X}$ is called a sufficient predictor. The intersection of all possible dimension reduction subspaces is called the *central subspace* $\mathcal{S}_{y|\mathbf{X}}$, if it exists. The main goal of SDR is to infer $\mathcal{S}_{y|\mathbf{X}}$, which is done through the estimations of its true structural dimension d and orthonormal basis matrix.

According to Cook (1998a), for a non-singular transformation of $\mathbf{X}$ such that $\mathbf{Z} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$, the following relation holds: $\mathcal{S}_{y|\mathbf{X}} = \mathbf{A}\mathcal{S}_{y|\mathbf{Z}}$. Considering a standardized predictor $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\{\mathbf{X} - E(\mathbf{X})\}$, we have that $\mathcal{S}_{y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{y|\mathbf{Z}}$, where $\boldsymbol{\Sigma} = cov(\mathbf{X})$ and $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. Typically, SDR methods estimate $\mathcal{S}_{y|\mathbf{Z}}$ first, then back-transform it to $\mathcal{S}_{y|\mathbf{X}}$. Hereafter kernel matrices to restore $\mathcal{S}_{y|\mathbf{Z}}$ for each method will be denoted as $\mathbf{M}_\bullet$, and it will be assumed that $\mathbf{M}_\bullet$ is exhaustively informative to $\mathcal{S}_{y|\mathbf{Z}}$ so that $\mathcal{S}(\mathbf{M}_\bullet) = \mathcal{S}_{y|\mathbf{Z}}$. With $\mathbf{Z}$-scale predictors, the classical but most popularly used SDR methodologies include:

(i) **sliced inverse regression** (SIR; Li, 1991):
$\mathbf{M}_{\mathrm{SIR}} = cov\{E(\mathbf{Z}|y)\}$. If y is categorical, a sample version of $E(\mathbf{Z}|y)$ is straightforward. If y is many-valued or continuous, y is categorized by dividing its range into h slices.

(ii) **sliced average variance estimation** (SAVE; Cook and Weisberg, 1991):
$\mathbf{M}_{\mathrm{SAVE}} = E[\{\mathbf{I}_p - cov(\mathbf{Z}|y)\}\{\mathbf{I}_p - cov(\mathbf{Z}|y)\}^{\mathrm{T}}]$. The sample version of $cov(\mathbf{Z}|y)$ is constructed in the same way as SIR.

(iii) **principal Hessian directions** (pHd; Li, 1992; Cook, 1998b):
$\mathbf{M}_{\mathrm{pHd}} = \boldsymbol{\Sigma}_{rzz} = E[\{y - E(y) - \beta^{\mathrm{T}}\mathbf{Z}\}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}]$, where β is the ordinary least squares on the regression of $y|\mathbf{Z}$. The sample version of $\mathbf{M}_{\mathrm{pHd}}$ is constructed by replacing the population quantities with their usual moment estimators.

(iv) **covariance method** (cov$k$ Yin and Cook, 2002):
$\mathbf{M}_{\mathrm{cov}k} = \{E(\mathbf{Z}w), E(\mathbf{Z}w^2), \ldots, E(\mathbf{Z}w^k)\}$, where $w = \{y - E(y)\}/\sqrt{\mathrm{var}(y)}$. As can be easily seen, $E(\mathbf{Z}w^j)$ in $\mathbf{M}_{\mathrm{cov}k}$ is the covariance between w^j and $\mathbf{Z}$ as well as the ordinary least squares coefficient of $w^j|\mathbf{Z}$ for $j = 1, \ldots, k$.

These four SDR methods can be implemented in the **dr**-package. The package provides the basis estimates and dimension test results. A detailed review on the **dr**-package is given in Weisberg (2002). The following question is a critical issue in using **dr**: *which one among the four methods has to be used?* Even with the same data, one can obtain dramatically different dimension reduction results from the four methods in **dr**, yet there is no clear and up-to-date guideline on method selection. It is known that SIR and cov$k$ work better when a linear trend exists in regression, while SAVE and pHd are effective under a nonlinear trend, particularly in the case of a quadratic relationship. However, in practice, these may not be useful guidelines for choosing an SDR method, because it is not easy to check the

existence of non/linear trends in regression. Even in the case that a linear trend exists, it is still not clear which of SIR and cov*k* should be used, as they yield different dimension test results.

The purpose of the paper is to develop a basis-adaptive selection algorithm for selecting an SDR method. The basic idea is to select the one that gives the highest correlation between the basis estimates obtained by all possible pairs of the four SDR methods. To measure the correlation, a trace correlation r (Hooper, 1959) will be used. The algorithm is data-driven and can be enhanced to other SDR methods, although it is highlighted with the use of the **dr**-package in this paper.

The organization of this article is as follows. We develop the idea of a basis-adaptive selection algorithm, and discuss its selection criteria. Next, we present a real data example and simulation studies. Finally, we summarize our work.

Basis-adaptive selection

Development of algorithm

We begin this section with soil evaporation data in Section 6 of Yin and Cook (2002). The data contains 46 observations of daily soil evaporation, daily air and soil temperature curves, daily humidity curves and wind speed. For illustration purposes, we consider a regression analysis of daily soil evaporation given the integrated area of and the range of the daily air and soil temperatures. We will revisit the data in a later section.

```
## loading data
evaporat <- read.table("evaporat.txt", header=T); attach(evaporat)
Rat <- Maxat-Minat; Rvh <- Maxh-Minh; Rst <- Maxst-Minst
w <- c(scale(evaporat$Evap,center=TRUE,scale=TRUE)); detach(evaporat)
evaporat <- data.frame(evaporat, Rat, Rvh, Rst)
w2 <- cbind(w, w^2); w3 <- cbind(w2, w^3); w4 <- cbind(w3, w^4)

## dr-package fitting
library(dr)
sir5 <- dr(Evap~Avat+Rat+Avst+Rst, data=evaporat, method="sir", nslice=5)
save5 <- update(sir5, method="save"); phdres <- update(sir5, method="phdres")
cov2 <- dr(w2~Avat+Rat+Avst+Rst,data=evaporat, method="ols")
cov3 <- dr(w3~Avat+Rat+Avst+Rst,data=evaporat, method="ols")
cov4 <- dr(w4~Avat+Rat+Avst+Rst,data=evaporat, method="ols")

## dimension test
round(dr.test(sir5),3)
round(dr.test(save5),3)
round(dr.test(phdres, numdir=4),3)
set.seed(100);round(dr.permutation.test(cov2,npermute=1000)$summary,3)
set.seed(100);round(dr.permutation.test(cov3,npermute=1000)$summary,3)
set.seed(100);round(dr.permutation.test(cov4,npermute=1000)$summary,3)
```

Table 1: Dimension test results for soil evaporation data

	SIR with 5 slices	SAVE with 5 slices	pHd	cov2	cov3	cov4
$H_0 : d = 0$	0.000	0.048	0.698	0.000	0.000	0.002
$H_0 : d = 1$	0.022	0.261	0.723	0.007	0.001	0.001
$H_0 : d = 2$	0.202	0.546	0.751	N/A	0.011	0.001
$H_0 : d = 3$	0.814	0.755	0.452	N/A	N/A	0.004

The p-values for the dimension estimation by the four methods in **dr** are reported in Table 1. For SAVE and pHd, the p-values under normal distributions are reported. Since a permutation test should be used for cov*k*, `set.seed(100)` was used to have reproducible results. According to Table 1, with level 5%, the SIR and the SAVE determine that $\hat{d} = 2$ and $\hat{d} = 1$, respectively. The pHd estimates that $\hat{d} = 0$, while the cov*k* estimates that $\hat{d} \geq 4$. In the evaporation data, the dimension estimation results are completely different for each of the four methods. Then, what methodological result should we use for the dimension reduction of the data? Unfortunately, there is no clear guidance for this issue.

Before developing selection criteria among the four SDR methods, there are some aspects that must be considered first. A selection based on the criteria should be data-driven, unless it is purposely

pre-selected. In addition, the criteria should be generally extended to the other SDR methods, although the four methods in **dr** are highlighted in this paper.

In order to have some idea of how to select a method, consider a simulated example of $y|\mathbf{X} = (x_1, \ldots, x_{10})^{\mathrm{T}} = x_1 + \varepsilon$. The column of $\boldsymbol{\eta} = (1, 0, \ldots, 0)^{\mathrm{T}}$ spans $\mathcal{S}_{y|\mathbf{Z}}$, and the sufficient predictor of x_1 is well estimated not only by SIR but also by cov2. Therefore, it is expected that the estimates by SIR and cov2 will be more highly correlated than those of any other pairs of the four methods. Assuming that $\mathbf{X}$ and ε are independently normally-distributed with $\mu = 0$ and $\sigma^2 = 0.1^2$, the averages of the absolute correlations between the pairs of the estimates from the four methods as well as between x_1 and the estimate from each method are computed through 500 iterations.

```
sir.cov <- sir.save <- sir.phd <- save.cov <- save.phd <- phd.cov <- NULL
sir.eta <- save.eta <- phd.eta <- cov.eta <- NULL

set.seed(1)

## starting loop
for (i in 1:500){

## model construction
  X <- matrix(rnorm(100*10), c(100,10)); y <- X[,1] + rnorm(100)
  w <- c(scale(y,center=TRUE,scale=TRUE)); w2 <- cbind(w, w^2)

## obtaining basis estimates from the SDR methods
  dir.sir5<-dr.direction(dr(y~X, method="sir", nslice=5))[,1]
  dir.cov2<-dr.direction(dr(w2~X, method="ols"))[,1]
  dir.save5<-dr.direction(dr(y~X, method="save", nslice=5))[,1]
  dir.phd <-dr.direction(dr(y~X, method="phdres"))[,1]

## computing the absolute correlation between the pairs of the estimates from the four methods
  sir.cov[i]<-abs(cor(dir.sir5, dir.cov2)); sir.save[i]<-abs(cor(dir.sir5, dir.save5))
  sir.phd[i]<-abs(cor(dir.sir5, dir.phd)); save.cov[i]<-abs(cor(dir.save5, dir.cov2))
  save.phd[i]<-abs(cor(dir.save5, dir.phd)); phd.cov[i]<-abs(cor(dir.phd, dir.cov2))

## computing the absolute correlation between the estimate from each method and x1
  sir.eta[i]<-abs(cor(dir.sir5, X[,1])); cov.eta[i]<-abs(cor(dir.cov2, X[,1]))
  save.eta[i]<-abs(cor(dir.save5, X[,1])); phd.eta[i]<-abs(cor(dir.phd, X[,1]))
 }

## the averages of the absolute correlations by each pair of the methods
round(apply(cbind(sir.cov, sir.save, sir.phd, save.cov, save.phd, phd.cov), 2,mean),3)
 sir.cov sir.save  sir.phd save.cov save.phd phd.cov
   0.966    0.208    0.257    0.214    0.348   0.288

## the averages of the absolute correlations by each method and x1
round(apply(cbind(sir.eta, cov.eta, save.eta, phd.eta), 2, mean), 3)
 sir.eta  cov.eta save.eta  phd.eta
   0.941    0.939    0.215    0.267
```

The highest average of the six pairwise absolute correlations is highest between SIR and cov$k$; in addition, either of SIR or cov$k$ estimates x_1 well. If the regression model is changed to $y|\mathbf{X} = x_1^2 + \varepsilon$, the pair of SAVE and pHd yield the highest averages in the absolute correlations among the six pairs, and both SAVE and pHd are good SDR methods for the regression.

Let us think about this in a reverse manner. Suppose that the estimates of sufficient predictors by a pair of SAVE and pHd are more highly correlated than those of any other pairs of the SDR methods. Then, it would be not bad reasoning to believe that SAVE or pHd should be preferable to the data, although we do not know what the true model is. That is, the pair of the two methods that gives the highest correlation would be not a bad choice for estimating $\mathcal{S}_{y|\mathbf{X}}$, although it is not guaranteed to be the best. We will formalize this idea.

First, we list all six possible pairs of the four methods in the **dr**-package: (1) (SIR, SAVE); (2) (SIR, pHd); (3) (SIR, cov$k$); (4) (SAVE, pHd); (5) (SAVE, cov$k$); (6) (pHd, cov$k$). Let $\hat{\boldsymbol{\eta}}_a$ and $\hat{\boldsymbol{\eta}}_b$ be orthonormal basis estimates of $\mathcal{S}_{y|\mathbf{Z}}$ by any pair among the six under $d = m$. If $d = 1$, the correlation between $\hat{\boldsymbol{\eta}}_a^{\mathrm{T}}\mathbf{Z}$ and $\hat{\boldsymbol{\eta}}_b^{\mathrm{T}}\mathbf{Z}$ can be simply computed using the usual Pearson correlation coefficient. However, if $d \geq 2$, the correlation between $\hat{\boldsymbol{\eta}}_a^{\mathrm{T}}\mathbf{Z}$ and $\hat{\boldsymbol{\eta}}_b^{\mathrm{T}}\mathbf{Z}$ is not straightforward. Now it should be noted that the correlation between $\hat{\boldsymbol{\eta}}_a^{\mathrm{T}}\mathbf{Z}$ and $\hat{\boldsymbol{\eta}}_b^{\mathrm{T}}\mathbf{Z}$ depends on the similarity between $\hat{\boldsymbol{\eta}}_a$ and $\hat{\boldsymbol{\eta}}_b$ regardless of the value

of d. The similarity of two matrices with the same column rank is equivalent to the distance between their column subspaces. Finally, a correlation between of $\hat{\boldsymbol{\eta}}_a^{\mathrm{T}}\mathbf{Z}$ and $\hat{\boldsymbol{\eta}}_b^{\mathrm{T}}\mathbf{Z}$ can be alternatively measured by a trace correlation r_{tr} (Hooper, 1959) between $\mathcal{S}(\hat{\boldsymbol{\eta}}_a)$ and $\mathcal{S}(\hat{\boldsymbol{\eta}}_b)$:

$$r_{\mathrm{tr}} = \sqrt{\frac{1}{m} trace\{\hat{\boldsymbol{\eta}}_a\hat{\boldsymbol{\eta}}_a^{\mathrm{T}}\hat{\boldsymbol{\eta}}_b\hat{\boldsymbol{\eta}}_b^{\mathrm{T}}\}}.$$

The trace correlation r_{tr} varies between 0 and 1, and higher values indicate that $\mathcal{S}(\hat{\boldsymbol{\eta}}_a)$ and $\mathcal{S}(\hat{\boldsymbol{\eta}}_b)$ are closer. If the two subspaces coincide, then we have $r_{\mathrm{tr}} = 1$.

The trace correlation is expected to be maximized under the true dimension of d, which is usually unknown. For a smaller choice of d, the true basis are underestimated, so each method will not normally estimate the same parts of $\mathcal{S}_{y|\mathbf{X}}$. This implies that r_{tr} will be smaller compared to it under the true dimension. In contrast, in the case of a larger selection of d, the true basis are overestimated. Then, there is redundancy in the estimates, and the redundancy will be random for each method. This indicates that smaller correlations should be expected. This discussion is confirmed through the following simulation example.

```
sir.cov <- sir.save <- sir.phd <- save.cov <- save.phd <- phd.cov <- NULL
sir.cov2 <- sir.save2 <- sir.phd2 <- save.cov2 <- save.phd2 <- phd.cov2 <- NULL
sir.cov3 <- sir.save3 <- sir.phd3 <- save.cov3 <- save.phd3 <- phd.cov3 <- NULL
sir.eta2 <- save.eta2 <- phd.eta2 <- cov.eta2 <- NULL

set.seed(1)

## true basis matrix of the central subspace
eta <- cbind(c(1, rep(0,9)), c(0, 1, rep(0,8)) )

## starting loop
for (i in 1:500){

## model construction
  X <- matrix(rnorm(100*10), c(100,10)); y <- X[,1] + X[,1]*X[,2]+ rnorm(100)
  w <- c(scale(y,center=TRUE, scale=TRUE)); w2 <- cbind(w, w^2); w3<- cbind(w2, w^3)

## obtaining basis estimates from the four methods
  sir5b <- dr(y~X, method="sir", nslice=5)$raw.evectors
  cov2b <- dr(w2~X, method="ols")$raw.evectors
  cov3b <- dr(w3~X, method="ols")$raw.evectors
  save5b <- dr(y~X, method="save", nslice=5)$raw.evectors
  phdb <- dr(y~X, method="phdres")$raw.evectors

## computing trace correlations under d=1 for the six pairs
  sir.cov[i] <- tr.cor(sir5b[,1], cov2b[,1], 1)
  sir.save[i] <- tr.cor(sir5b[,1], save5b[,1], 1)
  sir.phd[i] <- tr.cor(sir5b[,1], phdb[,1], 1)
  save.cov[i] <- tr.cor(save5b[,1], cov2b[,1], 1)
  save.phd[i] <- tr.cor(save5b[,1], phdb[,1], 1)
  phd.cov[i] <- tr.cor(phdb[,1], cov2b[,1], 1)

## computing trace correlations under d=2 for the six pairs
  sir.cov2[i] <- tr.cor(sir5b[,1:2], cov2b[,1:2], 2)
  sir.save2[i] <- tr.cor(sir5b[,1:2], save5b[,1:2], 2)
  sir.phd2[i] <- tr.cor(sir5b[,1:2], phdb[,1:2], 2)
  save.cov2[i] <- tr.cor(save5b[,1:2], cov2b[,1:2], 2)
  save.phd2[i] <- tr.cor(save5b[,1:2], phdb[,1:2], 2)
  phd.cov2[i] <- tr.cor(phdb[,1:2], cov2b[,1:2], 2)
  sir.cov3[i] <- tr.cor(sir5b[,1:3], cov3b[,1:3], 3)

## computing trace correlations under d=3 for the six pairs
  sir.save3[i] <- tr.cor(sir5b[,1:3], save5b[,1:3], 3)
  sir.phd3[i] <- tr.cor(sir5b[,1:3], phdb[,1:3], 3)
  save.cov3[i] <- tr.cor(save5b[,1:3], cov3b[,1:3], 3)
  save.phd3[i] <- tr.cor(save5b[,1:3], phdb[,1:3], 3)
  phd.cov3[i] <- tr.cor(phdb[,1:3], cov3b[,1:3], 3)
 }
```

```
## averages of the trace correlations for each pair under d=1,2,3
round(apply(cbind(sir.cov, sir.save, sir.phd, save.cov, save.phd, phd.cov),
      2, mean), 3)
round(apply(cbind(sir.cov2, sir.save2, sir.phd2, save.cov2, save.phd2, phd.cov2),
      2, mean), 3)
round(apply(cbind(sir.cov3, sir.save3, sir.phd3, save.cov3, save.phd3, phd.cov3),
      2, mean), 3)

 sir.cov  sir.save   sir.phd  save.cov  save.phd   phd.cov
   0.580     0.224     0.476     0.317     0.451     0.616
 sir.cov2 sir.save2  sir.phd2 save.cov2 save.phd2  phd.cov2
   0.822     0.397     0.656     0.486     0.606     0.801
 sir.cov3 sir.save3  sir.phd3 save.cov3 save.phd3  phd.cov3
   0.773     0.517     0.665     0.577     0.679     0.753
```

In the model, the true structural dimension is equal to two, and $\mathcal{S}_{y|\mathbf{Z}}$ is spanned by the columns of $\boldsymbol{\eta} = \{(1,0,\ldots,0),(0,1,0,\ldots,0)\}^{\mathrm{T}}$. For $d = 1,2,3$, the averages of the trace correlations are computed for the six pairs. According to the results, the averages are maximized under the true structural dimension $d = 2$ for each pair, as expected. In addition, the maximum trace correlation is attained at the pair of SIR and cov2 under $d = 2$.

Based on this discussion, the selection algorithm can be developed as follows:

Algorithm

1. Fix the maximum value $d_{\max}$ of d. The value should be less than or equal to $\min\{k, (h_{\mathrm{SIR}} - 1), (p-1)\}$, where k stands for the maximum polynomial in cov$k$ and h_{SIR} is the number slices in SIR.
2. Compute r_{tr} between the basis estimates from each pair for $m = 1,\ldots,d_{\max}$. For $d = 1$ or 2, the cov2 is commonly used, and the cov$d$ is employed for $d \geq 3$.
3. Choose a pair of the methods to provide the maximum r_{tr} in Step 2. The pair chosen in this step will be called the *initial pair*.
4. Remove the initial pair and all of the other pairs not containing one of the methods of the initial pair for all values of d. After completing this step, the surviving pairs must contain one method of the initial pair.
5. Search a second pair that has the highest r_{tr} among all of the remaining pairs. This pair will be called the *final pair*.
6. The common method in the initial and final pairs is the representative SDR method to the data.

Steps 4–5 are required to select one of the methods of the initial pair in Step 2. This approach of selecting SDR methods through the proposed algorithm is called *basis-adaptive selection* (BAS). The BAS is data-driven and not necessarily limited in the four SDR methods of SIR, SAVE, pHd and cov$k$, as one can extend it to other SDR methods.

The bas1 function

The `bas1` function runs the BAS algorithm for SIR, SAVE, pHd and cov$k$. The function requires **dr** and has the following arguments:

```
bas1(formula, data, nsir=5, nsave=4, k=4, plot=TRUE)
```

The arguments of `nsir`, `nsave` and `k` determine the numbers of slices for SIR and SAVE as well as the moment for cov$k$, respectively. The default values are 5, 4 and 4, in order. If `plot=TRUE`, the function `bas1` returns a scatter plot of the trace correlations against the various choices of d, up to min(`nsir`-1, `k`) for the six pairs of the four methods in **dr**.

The values returned by the `bas1` function are `selection`, `sir`, `save`, `phd` and `covk`. Their descriptions are as follows:

- `selection`: a selected method among SIR, SAVE, pHd and cov$k$ by the BAS algorithm
- `sir`: the SIR application object with the number of slices equal to `nsir`
- `save`: the SAVE application object with the number of slices equal to `nsave`
- `pHd`: the pHd application object
- `covk`: a list type object by the cov$k$ application objects with the kth polynomial

Soil evaporation data: Revisited

We revisit the soil evaporation data. Here, the application of the BAS is more extensively studied under the cross-combinations of four or five numbers of slices for SIR and SAVE and the third or forth order of polynomials for cov$k$. We define "BASijk" for $i = 3, 4, 5$, $j = 4, 5$ and $k = 2, \ldots, i$. For example, in BAS453, the numbers of slice for SIR and SAVE are four and five, respectively, and the order of polynomial for cov$k$ is three.

```
## BAS application
BAS342 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=3, nsave=4, k=2, plot=FALSE)
BAS343 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=3, nsave=4, k=3, plot=FALSE)
BAS352 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=3, nsave=5, k=2, plot=FALSE)
BAS353 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=3, nsave=5, k=3, plot=FALSE)

BAS443 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=4, nsave=4, k=3, plot=FALSE)
BAS444 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=4, nsave=4, k=4, plot=FALSE)
BAS453 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=4, nsave=5, k=3, plot=FALSE)
BAS454 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=4, nsave=5, k=4, plot=FALSE)

BAS544 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=5, nsave=4, k=4, plot=FALSE)
BAS545 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=5, nsave=4, k=5, plot=FALSE)
BAS554 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=5, nsave=5, k=4, plot=FALSE)
BAS555 <- bas1(Evap~Avat+Rat+Avst+Rst, data=evaporat, nsir=5, nsave=5, k=5, plot=FALSE)

## Selection results
BAS342$selection; BAS343$selection; BAS352$selection; BAS353$selection
BAS443$selection; BAS444$selection; BAS453$selection; BAS454$selection
BAS544$selection; BAS545$selection; BAS554$selection; BAS555$selection

## dimension test
set.seed(100); round(dr.permutation.test(BAS342$covk$cov2, npermute=1000)$summary, 3)
set.seed(100); round(dr.permutation.test(BAS343$covk$cov3, npermute=1000)$summary, 3)
round(dr.test(BAS443$phd, numdir=4), 3)
round(dr.test(BAS544$sir), 3)
```

Both BAS342 and BAS352 recommended cov2, and cov3 was the selection of BAS343 and BAS353. For the cases of BAS4jk and BAS5jk, pHd and SIR were recommended regardless of the values of j and k. These selection results are summarized in Table 2. From Table 2, it can be seen that the selection results are different from the numbers of slices in SIR. For example, if the numbers of slices in SIR was 3, BAS recommended cov$k$. For $i = 4$ and 5, BAS chose pHd and SIR, respectively. Thus, additional work needs to be done in order to choose between cov$k$, pHd and SIR in the soil evaporation data.

For this purpose, we considered how reasonably the methods recommended in Table 2 estimated the true dimension of the central subspace. The dimension estimation results are already summarized in Table 1, and the nominal level 5% was used. First, the two methods of cov2 and cov3 were inspected. According to Table 1, cov2 and cov3 determine that $d \geq 2$ and $d \geq 3$, respectively. This indicates that the order of polynomial in cov$k$ should be bigger than or equal to 4 for further dimension determination. However, in the case $k = 4$, cov$k$ yields $\hat{d} \geq 4$, so the dimension reduction is meaningless because $p = 4$. Therefore, we conclude that cov$k$ would not be recommended one for this data. Next we consider pHd, which infers that $\hat{d} = 0$ according to Table 1. Therefore, the pHd should also be ruled out for the final choice. In Table 1, the SIR with 5 slices determines that $\hat{d} = 2$, which is the most reasonable among the three recommendations. Thus, one may continue the regression analysis with the two-dimensional sufficient predictors from the SIR results.

Table 2: Method selection results by BAS in soil evaporate data

	BAS3•2	BAS3•3	BAS4••	BAS5••
Recommendation	cov2	cov3	pHd	SIR

Numerical studies

Predictors $\mathbf{X} = (x_1, \ldots, x_{10})^{\mathrm{T}}$ and a random error ε were independently generated from $N(0,1)$. Under these variable configurations, the following four models were considered:

- Model 1: $y = x_1 + \varepsilon$
- Model 2: $y = x_1^2 + \varepsilon$
- Model 3: $y = x_1 + x_1 x_2 + \varepsilon$
- Model 4: $y = 1 + x_1 + \exp(x_2)\varepsilon$

In Models 1 and 2, the column of $\boldsymbol{\eta} = (1, 0, \ldots, 0)^{\mathrm{T}}$ spans $\mathcal{S}_{y|\mathbf{X}}$, so $d = 1$. The structural dimension of Models 3 and 4 is two, and $\mathcal{S}_{y|\mathbf{X}}$ is spanned by the columns of $\boldsymbol{\eta} = \{(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0)\}^{\mathrm{T}}$. The four models are commonly used in Cook and Weisberg (1991); Li (1991, 1992); Yin and Cook (2002). The desired SDR methods for each model are as follows: Model 1: SIR and cov*k*; Model 2: SAVE and pHd; Model 3: cov*k* and pHd; Model 4: SIR and cov*k*.

Each model was iterated 500 times with $n = 100$ and $n = 200$. Throughout all numerical studies, five slices were commonly used for SIR and SAVE, and four was the maximum polynomial in cov*k*.

As a summary, the selection percentages of the methods by BAS are reported in Table 3.

Table 3: Percentages of the selection of a method among SIR, SAVE, pHd, and cov*k* by BAS

	$n = 100$				$n = 200$			
	SIR	SAVE	pHd	cov*k*	SIR	SAVE	pHd	cov*k*
Model 1	41.2	0.6	0.2	58.0	49.4	0.0	0.0	50.6
Model 2	1.0	17.8	78.4	2.8	0.0	22.8	77.2	0.0
Model 3	6.4	2.0	25.6	66.0	3.6	1.2	25.0	70.2
Model 4	6.4	0.6	6.4	86.6	26.8	1.2	1.4	70.6

According to Table 3, the BAS selects SIR and cov*k* 99.8% of the time for Model 1 and SAVE and pHd 97.8% of the time for Model 2 with $n = 100$, respectively. These results are consistent with the discussion given in the previous section regarding the development of the algorithm. In Model 1, cov*k* is preferred to SIR with $n = 100$, but the two are almost equally selected with $n = 200$. For Model 2, the pHd is recommended more than SAVE with $n = 100$ and $n = 200$. In Model 3, the pHd and the cov*k* are two dominant methods according to BAS, although the cov*k* is selected more frequently with $n = 100$. For Model 4, the cov*k* is recommended most frequently, but the selection percentages of SIR rapidly grow with $n = 200$.

```
set.seed(5); sel1 <- sel2 <- sel3 <- sel4 <- sel12 <- sel22 <- sel32 <- sel42 <- NULL

## starting loop
for (i in 1:500){

## model construction for n=100
  X <- matrix(rnorm(100 * 10), c(100, 10))
  y1 <- X[,1] + rnorm(100)
  y2 <- X[,1]^2 + rnorm(100)
  y3 <- X[,1] + X[,1] * X[,2] + rnorm(100)
  y4 <- 1 + X[,1] + exp(X[,2]) * rnorm(100)

## model construction for n=200
  X2 <- matrix(rnorm(200 * 10), c(200,10))
  y12 <- X2[,1] + rnorm(200)
  y22 <- X2[,1]^2 + rnorm(200)
  y32 <- X2[,1] + X2[,1] * X2[,2] + rnorm(200)
  y42 <- 1 + X2[,1] + exp(X2[,2]) * rnorm(200)

## selection by BAS
  sel1[i] <- bas1(y1~X, plot=FALSE)$selection
  sel2[i]<-bas1(y2~X, plot=FALSE)$selection
  sel3[i] <- bas1(y3~X, plot=FALSE)$selection
  sel4[i]<-bas1(y4~X, plot=FALSE)$selection
```

```
  sel12[i] <- bas1(y12~X2, plot=FALSE)$selection
  sel22[i]<-bas1(y22~X2, plot=FALSE)$selection
  sel32[i] <- bas1(y32~X2, plot=FALSE)$selection
  sel42[i]<-bas1(y42~X2, plot=FALSE)$selection
}

## computing the selection percentages for model 1
# n=100
c(length(which(sel1=="sir")), length(which(sel1=="save")),
  length(which(sel1=="phd")), length(which(sel1=="covk"))) / 500
# n=200
c(length(which(sel12=="sir")), length(which(sel12=="save")),
  length(which(sel12=="phd")), length(which(sel12=="covk"))) / 500

## computing the selection percentages for model 2
# n=100
c(length(which(sel2=="sir")), length(which(sel2=="save")),
  length(which(sel2=="phd")), length(which(sel2=="covk"))) / 500
# n=200
c(length(which(sel22=="sir")), length(which(sel22=="save")),
  length(which(sel22=="phd")), length(which(sel22=="covk"))) / 500

## computing the selection percentages for model 3
# n=100
c(length(which(sel3=="sir")), length(which(sel3=="save")),
  length(which(sel3=="phd")), length(which(sel3=="covk"))) / 500
# n=200
c(length(which(sel32=="sir")), length(which(sel32=="save")),
  length(which(sel32=="phd")), length(which(sel32=="covk"))) / 500

## computing the selection percentages for model 4
# n=100
c(length(which(sel4=="sir")), length(which(sel4=="save")),
  length(which(sel4=="phd")), length(which(sel4=="covk"))) / 500
# n=200
c(length(which(sel42=="sir")), length(which(sel42=="save")),
  length(which(sel42=="phd")), length(which(sel42=="covk"))) / 500
```

Summary

Sufficient dimension reduction (SDR) is a useful dimension reduction method in regression. The popularly used SDR methods among the others include sliced inverse regression (Li, 1991), sliced average variance estimation (Cook and Weisberg, 1991), principal Hessian directions (Li, 1992) and covariance method (Yin and Cook, 2002). Currently, the **dr**-package is the only one to cover the four sufficient dimension reduction methods in R. However, users were left without practical guidelines as to which SDR method should be chosen. To remedy this, we developed the basis-adaptive selection (BAS) algorithm to recommend a SDR method in **dr** by maximizing a trace correlation (Hooper, 1959). A real data example and numerical studies confirm its potential usefulness in practice.

The BAS algorithm requires the two parts. The first is the basis estimates of the dimension reduction subspace, and the second is a quantity to measure the distances between the subspaces spanned by the columns of the estimates. For non-linear feature extractions through different kernel methods, there is no reason why the BAS algorithm cannot be applied if some quantity to measure the distance between the non-linear subspaces defined in different kernels is feasible.

If the sliced inverse regression applications with various numbers of slices replaces the other three methods in the BAS, the BAS algorithm can be utilized to find a good number of slices for this method.

The code for BAS is available through the personal webpage of the author: http://home.ewha.ac.kr/~yjkstat/bas.R.

Acknowledgments

For the corresponding author Jae Keun Yoo, this work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean Ministry of Education (NRF-2017R1A2B1004909).

Bibliography

R. D. Cook. *Regression Graphics: Idea for Studying Regressions through Graphics*. John Wiley & Sons, 1998a. [p124]

R. D. Cook. Principal hessian directions revisited. *Journal of the American Statistical Association*, 93(441): 84–94, 1998b. URL http://dx.doi.org/10.1080/01621459.1998.10474090. [p124]

R. D. Cook and S. Weisberg. Comment: Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):328–332, 1991. URL http://dx.doi.org/10.1080/01621459.1991.10475036. [p124, 130, 131]

J. Hooper. Simultaneous equations and canonical correlation theory. *Econometrika*, 27(2):245–256, 1959. URL http://dx.doi.org/10.2307/1909445. [p125, 127, 131]

K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):326–342, 1991. URL http://doi.org/10.1080/01621459.1991.10475035. [p124, 130, 131]

K. C. Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. URL http://dx.doi.org/10.1080/01621459.1992.10476258. [p124, 130, 131]

S. Weisberg. Dimension reduction regression in R. *Journal of Statistical Software*, 7(1):1–22, 2002. URL http://dx.doi.org/10.18637/jss.v007.i01. [p124]

X. Yin and R. D. Cook. Dimension reduction for the conditional kth moment in regression. *Journal of Royal Statistical Society Series B*, 64(2):159–175, 2002. URL http://dx.doi.org/10.1111/1467-9868.00330. [p124, 125, 130, 131]

Jae Keun Yoo
Department of Statistics, Ewha Womans University
Seoul 03760
Republic of Korea
peter.yoo@ewha.ac.kr