Comparison between SDR and PCA on Hemoglobin data

Xuelong Wang

2020-03-26

Contents

1 Motivation

Since PCBs are usually high dimension and highly correlated, one approach is to apply the dimension reduction method to the original data, then apply an analysis method on the projected data set.

2 Goal

Here, we compare the performances of Sufficient Dimension Reduction with PCA, which are commonly used dimension reduction method.

1. For the simplicity, we just choose the linear regression as the analysis method
2. We use R^2 and RMSE as the evaluation methods
3. Data set: Hemogoblin data

3 Data

The data is a subset of the Hemoglobin data removing all the missing values.

- $n = 977$
- $p = 38$
- All the PCBs have been standardized and log-transformed

4 Dimension reduction

For PCA, we adopt a Cross-Validation method to choose the number of components. For SDR methods, we use the large sample tests for the results. The number of directions are the followings:

- PCA: 24
- SIR: 2
- SAVE: 14

5 Predictive result

- Full data: R^2: 0.1720779, Ajusted R^2: 0.1385374
- PCA: R^2: 0.1567446, Ajusted R^2: 0.1354861
- SIR: R^2: 0.1588594, Ajusted R^2: 0.1571322
- SAVE: R^2: 0.0455469, Ajusted R^2: 0.0316568

```
summary(lm.full)
```

```
##
## Call:
## lm(formula = LBXGH ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6354 -0.2232 -0.0302  0.1766  3.8927
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.3883316  0.0131052 411.158  < 2e-16 ***
## PCB199      -0.0327179  0.0569586  -0.574  0.56582
## PCB028       0.0008881  0.0263714   0.034  0.97314
## PCB044      -0.0762283  0.0553939  -1.376  0.16911
## PCB049       0.0464157  0.0560848   0.828  0.40811
## PCB052      -0.0075679  0.0467478  -0.162  0.87143
## PCB066      -0.0449042  0.0292486  -1.535  0.12506
## PCB074       0.0055954  0.0493000   0.113  0.90966
## PCB081      -0.0049366  0.0186373  -0.265  0.79116
## PCB087       0.0126421  0.0213907   0.591  0.55466
## PCB099       0.0417711  0.0551759   0.757  0.44921
## PCB101       0.0601248  0.0271954   2.211  0.02729 *
## PCB105      -0.0600041  0.0627788  -0.956  0.33942
## PCB110      -0.0266809  0.0448013  -0.596  0.55163
## PCB118      -0.0412604  0.0916873  -0.450  0.65281
## PCB126       0.1000357  0.0334596   2.990  0.00286 **
## PCB128       0.0137990  0.0168315   0.820  0.41252
## PCB138       0.2705724  0.1031461   2.623  0.00885 **
## PCB146      -0.0302316  0.0741704  -0.408  0.68366
## PCB149       0.0176199  0.0262914   0.670  0.50291
## PCB151       0.0046101  0.0227062   0.203  0.83916
## PCB153      -0.2429019  0.1336133  -1.818  0.06939 .
## PCB156       0.0046350  0.0447184   0.104  0.91747
## PCB157      -0.0402937  0.0410557  -0.981  0.32663
## PCB167       0.0105351  0.0340251   0.310  0.75691
```

```
## PCB169       0.0834229  0.0354688   2.352  0.01888 *
## PCB170       0.0641558  0.0726120   0.884  0.37717
## PCB172       0.0251044  0.0460867   0.545  0.58607
## PCB177       0.0712410  0.0412204   1.728  0.08426 .
## PCB178      -0.1009867  0.0532090  -1.898  0.05801 .
## PCB180       0.0277260  0.1045943   0.265  0.79100
## PCB183      -0.0355295  0.0428274  -0.830  0.40698
## PCB187      -0.0048102  0.0751001  -0.064  0.94894
## PCB189      -0.0314053  0.0170745  -1.839  0.06619 .
## PCB194       0.0160784  0.0422755   0.380  0.70379
## PCB195       0.0237603  0.0265822   0.894  0.37164
## PCB196      -0.0466508  0.0528280  -0.883  0.37743
## PCB206      -0.0070683  0.0673858  -0.105  0.91648
## PCB209       0.1128879  0.0453778   2.488  0.01303 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4096 on 938 degrees of freedom
## Multiple R-squared:  0.1721, Adjusted R-squared:  0.1385
## F-statistic:  5.13 on 38 and 938 DF,  p-value: < 2.2e-16
```

```
summary(lm.pca)
```

```
##
## Call:
## lm(formula = LBXGH ~ ., data = data.pca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6557 -0.2314 -0.0334  0.1832  3.8758
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.3883316  0.0131284 410.432  < 2e-16 ***
## PC1          0.0315476  0.0028433  11.096  < 2e-16 ***
## PC2          0.0014433  0.0050083   0.288  0.77326
## PC3         -0.0133887  0.0093009  -1.440  0.15034
## PC4         -0.0045707  0.0119921  -0.381  0.70318
## PC5          0.0040034  0.0139814   0.286  0.77468
## PC6          0.0005012  0.0154009   0.033  0.97404
## PC7          0.0276077  0.0174337   1.584  0.11362
## PC8         -0.0282944  0.0186022  -1.521  0.12859
## PC9         -0.0003983  0.0202382  -0.020  0.98430
## PC10         0.0423378  0.0211573   2.001  0.04566 *
## PC11        -0.0091071  0.0227551  -0.400  0.68908
## PC12        -0.0555274  0.0233440  -2.379  0.01757 *
## PC13         0.0679797  0.0238138   2.855  0.00440 **
## PC14         0.0417459  0.0250461   1.667  0.09589 .
## PC15         0.0168920  0.0263110   0.642  0.52102
## PC16         0.1013823  0.0289368   3.504  0.00048 ***
## PC17         0.0030678  0.0315975   0.097  0.92268
## PC18         0.0381365  0.0349858   1.090  0.27596
## PC19        -0.0440858  0.0363367  -1.213  0.22533
## PC20         0.0159710  0.0399134   0.400  0.68914
## PC21         0.0551405  0.0414951   1.329  0.18422
```

```
## PC22        -0.0749801  0.0445850  -1.682  0.09295 .
## PC23        -0.0818248  0.0474623  -1.724  0.08503 .
## PC24        -0.0810779  0.0483798  -1.676  0.09409 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4104 on 952 degrees of freedom
## Multiple R-squared:  0.1567, Adjusted R-squared:  0.1355
## F-statistic: 7.373 on 24 and 952 DF,  p-value: < 2.2e-16
```

```
summary(lm.sir)
```

```
##
## Call:
## lm(formula = LBXGH ~ ., data = data.sir)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6803 -0.2187 -0.0335  0.1623  3.9905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.38833    0.01296 415.669   <2e-16 ***
## Dir1         0.44329    0.03280  13.514   <2e-16 ***
## Dir2        -0.06240    0.05447  -1.146    0.252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4052 on 974 degrees of freedom
## Multiple R-squared:  0.1589, Adjusted R-squared:  0.1571
## F-statistic: 91.98 on 2 and 974 DF,  p-value: < 2.2e-16
```

```
summary(lm.save)
```

```
##
## Call:
## lm(formula = LBXGH ~ ., data = data.save)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4310 -0.2591 -0.0457  0.1819  4.0497
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.388332   0.013894 387.804  < 2e-16 ***
## Dir1         0.038527   0.068488   0.563 0.573879
## Dir2         0.067607   0.056335   1.200 0.230400
## Dir3        -0.003674   0.058575  -0.063 0.950005
## Dir4        -0.020984   0.062486  -0.336 0.737075
## Dir5        -0.069311   0.081176  -0.854 0.393403
## Dir6        -0.059575   0.085174  -0.699 0.484439
## Dir7         0.016205   0.064334   0.252 0.801176
## Dir8        -0.133869   0.042919  -3.119 0.001868 **
## Dir9         0.054616   0.059497   0.918 0.358868
## Dir10       -0.186364   0.058612  -3.180 0.001522 **
```

```
## Dir11       -0.040498   0.050900  -0.796 0.426437
## Dir12        0.069939   0.062025   1.128 0.259772
## Dir13       -0.177660   0.045598  -3.896 0.000104 ***
## Dir14        0.092289   0.041334   2.233 0.025793 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4343 on 962 degrees of freedom
## Multiple R-squared:  0.04555,    Adjusted R-squared:  0.03166
## F-statistic: 3.279 on 14 and 962 DF,  p-value: 3.895e-05
```