STAT452/652 Solution to HW05 - Lecture 7

Due on Oct 16, 2020

1 Applications - Insurance

1.1 Question 1

(a) After constructing indicator variables, our data frame has 1797 rows and 21 columns.

1.2 Question 2

(a) Using the rule of only keeping PCs which explain an above average proportion of variance, we would keep 15 of the 20 components.

(b) See Figures 1 and 2 for the scree and cumulative variance plots respectively. Note that we add the point (0,0) to the cumulative variance plot to illustrate the effect of the first component (if 0 components are included, the proportion of variability explained is 0).

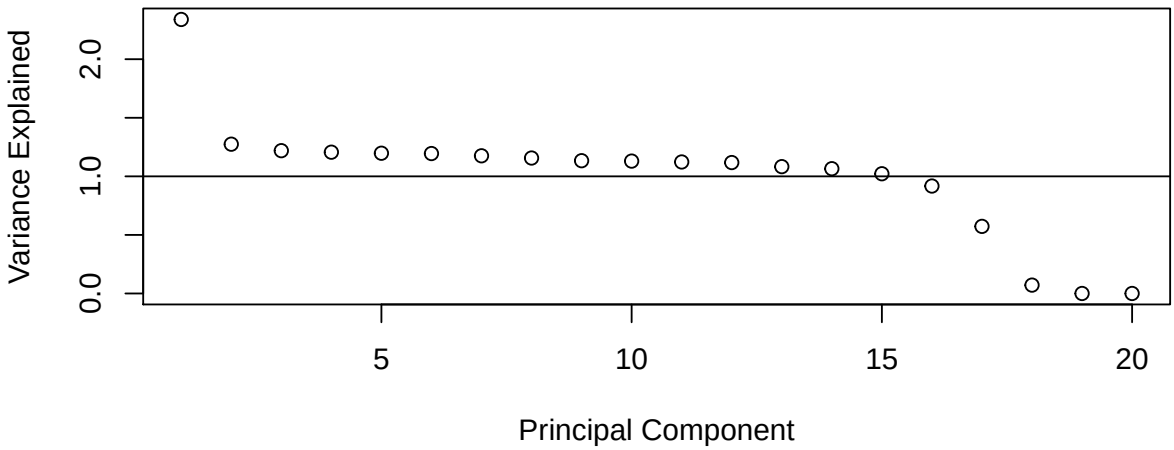

Figure 1: Scree plot for insurance data.

(c) Based on Figures 1 and 2, there is no real obvious cutoff. The nearly constant scree plot between 2 and 16 components creates a nearly linear cunulative variance plot, and makes it hard to see how to reduce

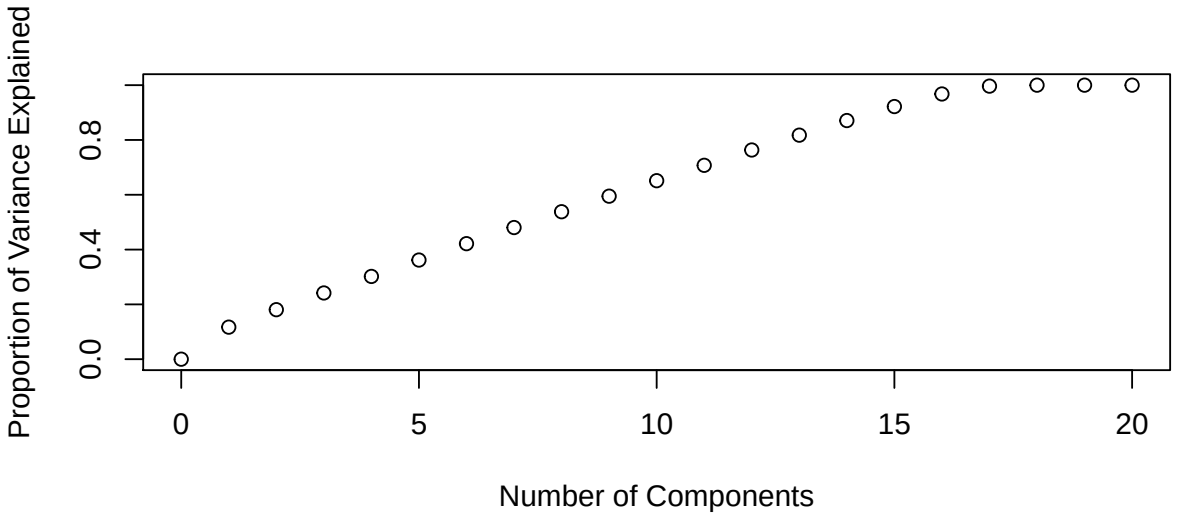

Figure 2: Cumulative variance plot for insurance data.

dimension substantially. One could argue that a single component is best for dimension reduction, byt this explains less than 20% of the variance.

One might also prefer to choose 15 or 16 components, because most components up to this point are about equally valuable (other than the first component), and after this point there is little variance left to be explained. Otherwise, it would be a completely arbitrary decision to choose something in between. Various rules might blindly pick a cutpoint, but there's no good reason to choose one number over the next between 2 and 15.

2 Applications - Air Quality

(a) The chosen number of components and MSPE on each fold is as follows.

```
##              1   2   3   4   5   6   7   8    9  10
## Components   5   5   3   3   3   3   3   3    5   3
## MSPE       260 323 514 138 146 417 266 656 1056 699
```

(b-c) The full-data MSPE for PLS is 447.

Boxplots of MSPEs and RMSPEs are given in Figures 3 and 4 respectively.

(d) Based on the MSPE boxplots in Figure 3, PLS appears to perdorm pretty similarly to the other models. Looking very closely, it is competitive with stepwise selection as a second-tier model (i.e. better than LASSO-1SE, but worse than the rest). But this is really making a lot of pretty small differences.

(e) Based on the RMSPE boxplots in Figure 4, PLS appears to be almost competitive with LS and Ridge, which have the lowest median RMSPEs. While PLS wins the competition at least 3 times (the bottom of its box, i.e. the first quartile, is at 1), this model's median RMSPE is worse than either LS or Ridge. It seems pretty comparable to stepwise.

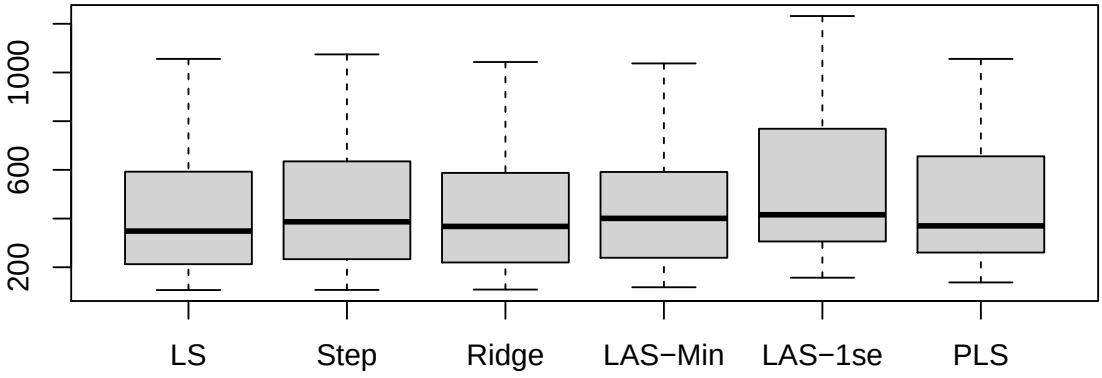

Figure 3: MSPE Boxplots

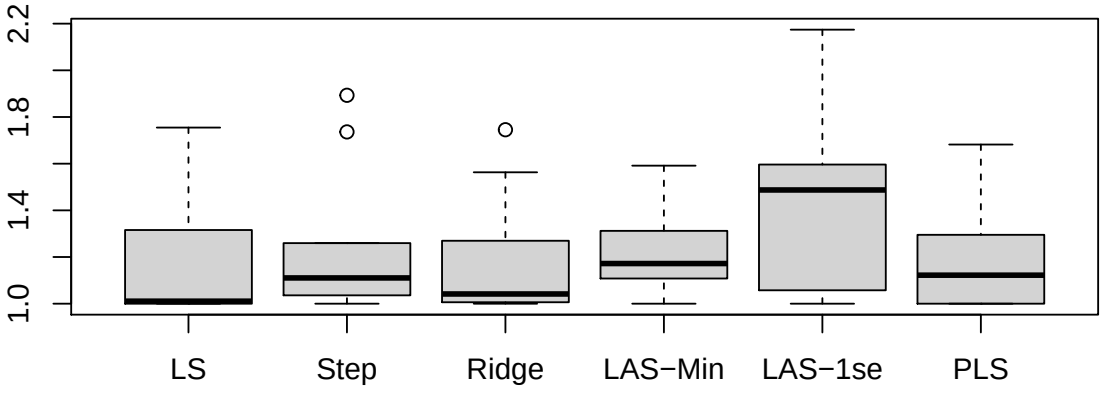

Figure 4: RMSPE Boxplots