Stat689: Statistical Aspects of Machine Learning and Data Mining, Fall 2009;

Homework 2 (10/15/09)

Instructor: Marc G.Genton
Student: Ganggang Xu

5.4 (a). To show $c^T\hat\beta$ is linear in Y and is unbiased for $c^T\beta$.
Let $A = Z^TZ$ and let A^- denote the generalized inverse of A, i.e.$AA^-A = A$. Then it is widely known that all solutions of the consistent linear system $A\beta = Z^TY$ can be expressed as

$$\hat\beta_{ols} = A^-Z^TY + (I - A^-A)b \tag{1}$$

where b is an arbitary $n \times 1$ vector. Since $c^T\beta$ is estimable, $c^T\beta = E(a^TY) = a^TZ\beta$ for all β, which implies $c^T = a^TZ$. Plug this back to equation (1), one has

$$c^T\hat\beta_{ols} = a^TZA^-Z^TY + a^TZ(I - A^-A)b$$

Let $T = Z(I - A^-A)$, then

$$T^TT = (I - A^-A)^TZ^TZ(I - A^-A) = (I - A^-A)^T(A - AA^-A) = 0$$

thus $T = 0$ or $Z = ZA^-A$ and

$$c^T\hat\beta_{ols} = a^TZA^-Z^TY = c^TA^-Z^TY$$

and using the previous conclusion $Z = ZA^-A$, one has

$$E\left(c^T\hat\beta_{ols}\right) = a^TZA^-Z^TZ\beta = a^T(ZA^-A)\beta = a^TZ\beta = c^T\beta$$

therefore, $c^T\hat\beta_{ols}$ is linear in Y and unbiased for $c^T\beta$.

(b). Show that $c^T\hat\beta_{ols}$ has minimum variance.
Let b^TY be any unbiased estimator of $c^T\beta$, then again, we must have $c^T = b^TZ$, use the same argument in the first part, one has

$$c^T\hat\beta_{ols} = b^TZA^-Z^TY$$

thus

$$\begin{aligned} Var(b^TY) = Var(b^TY - c^T\hat\beta_{ols} + c^T\hat\beta_{ols}) &= Var(b^TY - c^T\hat\beta_{ols}) + Var(c^T\hat\beta_{ols}) \\ &+ 2Cov(b^TY - c^T\hat\beta_{ols}, c^T\hat\beta_{ols}) \end{aligned}$$

focus on the third part,

$$\begin{aligned} Cov(b^TY - c^T\hat\beta_{ols}, c^T\hat\beta_{ols}) &= Cov(b^TY - b^TZA^-Z^TY, b^TZA^-Z^TY) \\ &= b^T(I - ZA^-Z^T)Var(Y)ZA^-Z^Tb \\ &= \sigma^2 b^T(I - ZA^-Z^T)ZA^-Z^Tb \\ &= \sigma^2 b^T(ZA^-Z^T - ZA^-AA^-Z^T)b \\ &= 0 \end{aligned}$$

therefore, we have

$$Var(b^T Y) = Var(b^T Y - c^T \hat{\beta}_{ols}) + Var(c^T \hat{\beta}_{ols}) \geq Var(c^T \hat{\beta}_{ols})$$

which implies that $c^T \hat{\beta}_{ols}$ has minimum variance among all linear unbiased estimators of $c^T \beta$.

5.6 It is a special case of Question 5.4, where equation (1) with $b = 0$. So the proof of this question is omitted.

5.7 To minimize

$$ESS(\beta) = (Y - X\beta)^T (Y - X\beta) \qquad \text{subject to} \qquad \beta^T \Omega \beta \leq c$$

Use the *Lagrange Mulitiplier*, it is equialent to minimize

$$(Y - X\beta)^T (Y - X\beta) + k(\beta^T \Omega \beta - c)$$

Taking derivative with respect to β and set it to 0, one has

$$-2X^T (Y - X\beta) + 2k\Omega\beta = 0$$

solving the equation leads us to

$$\hat{\beta} = (X^T X + k\Omega)^{-1} X^T Y$$

which is exactly the ridge regression estimator.

5.9 For the standardized data, we have following results:

(a). The PCR coefficient estimates:

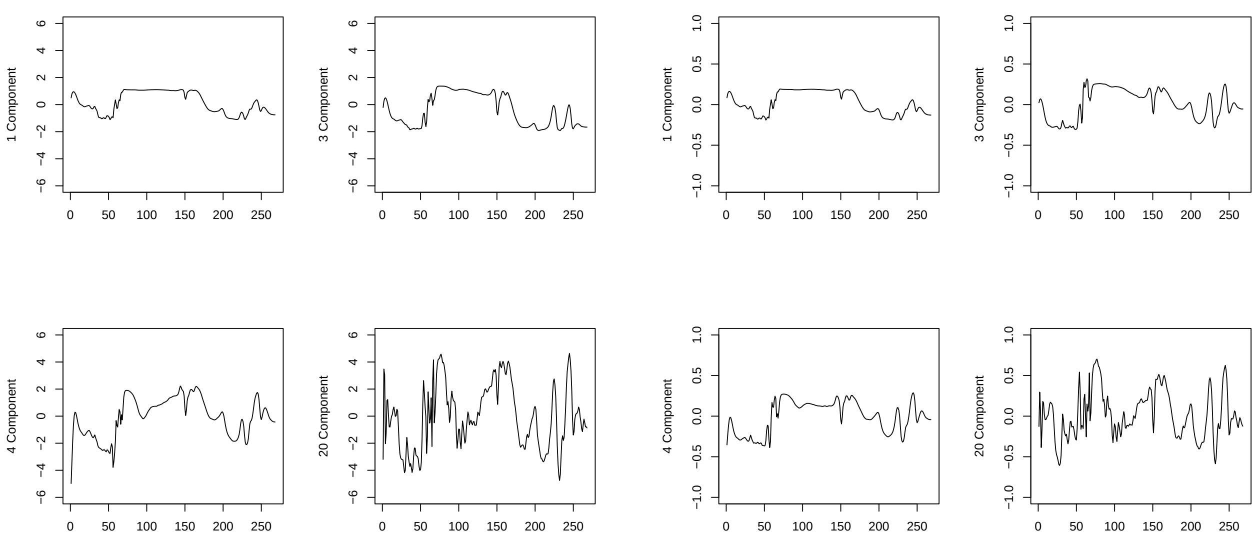

Figure 1: PCR estimates for the PET yarn data. Centered data (left) and Standardized data (right)

we can see from Figure 1 that the trends of the estimates doesn't change much but the magnitudes of the estimates changes a lot. This is because the magnitudes of 268 covariates are very similar, and the standardization only affects the magnitudes of the estimates. Centering or not centering the data has no effect on the trend of the estimates.

(b). The PLSR coefficient estimates

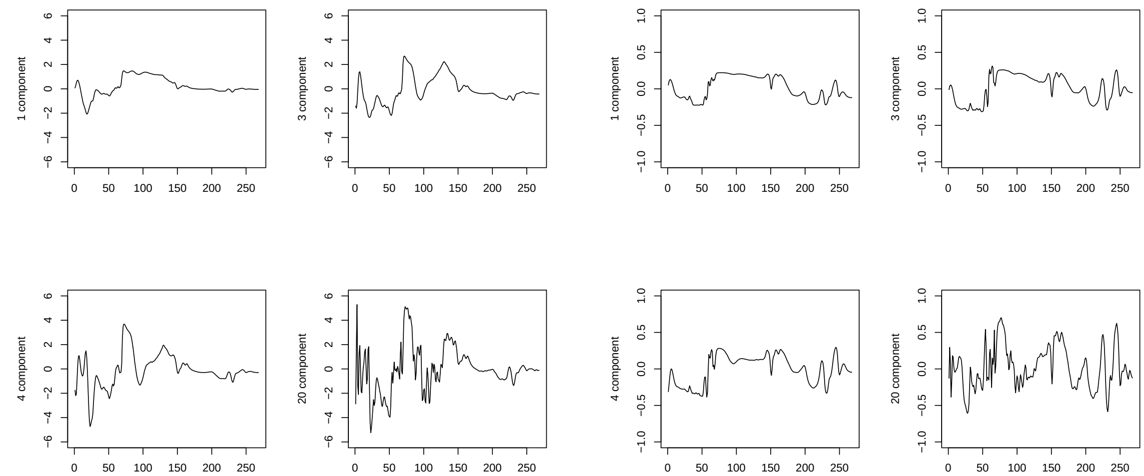

Figure 2: PLSR estimates for the PET yarn data. Centered data (left) and Standardized data (right)

we can see from Figure 2 that both of trends and magnitudes of the estimates were changed by the standardization of the datat. In fact, the PLSR estimates of the standardized data looks exactly the same as the PCR estimates for some reason. Centering or not centering the data has no effect on the trend of the estimates.

(c). Ridge Regression estimates:

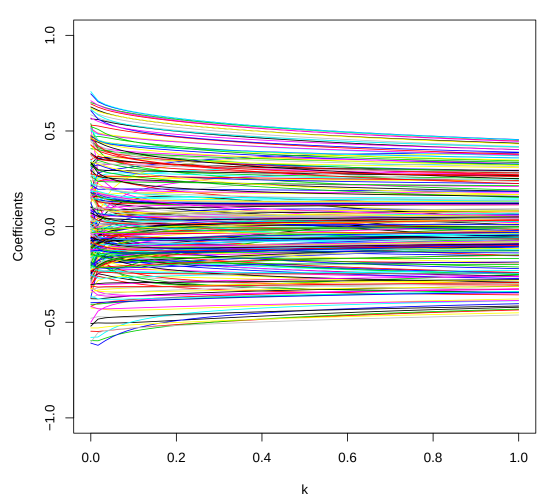

Figure 3: *Ridge Regression estimates for the PET yarn data. Standardized data*

For ridge regression, again, because the magnitudes of 268 covariates are very similar, hance the trends of estimates are about the same but with different magnitudes. Centering or not centering the data has no effect on the trend of the estimates.

5.18 The results are summarized in Table 1. The resuls of LASSO, LARS, Stepwise, and Subset selection methods are very similar, all of them dropped predictors $INDUS$ and AGE while OLS and Forward Stagewise selecting full model. But the coefficients of $INDUS$ and AGE are very small in OLS and Forward Stagewise Models. As for the estimates, the LASSO and LARS estimates are tend to be smaller in magnitude than other methods because of the shrinkage method. Using the cp criterion, the subsets selection and stepwise selection end up with exactly the same model, which is not surprising because the number of potential predictors is moderate.

Table 1: Model estimates ($\times 10^{-2}$.) of Boston Housing data using different methods. All models are selected as the one with lowest CP scores.

Selection Method	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
OLS	-1.03	0.12	0.25	10.09	-77.84	9.08	0.02	-4.91	1.43	-0.06	-3.83	0.04	-2.90
LASSO	-0.97	0.08	N/A	10.28	-65.35	9.47	N/A	-4.55	1.05	-0.04	-3.64	0.04	-2.86
LARS	-0.97	0.08	N/A	10.28	-65.35	9.47	N/A	-4.55	1.05	-0.04	-3.64	0.04	-2.86
Forward Stagewise	-1.01	0.11	0.18	10.15	-73.91	9.49	0.01	-4.80	1.31	-0.06	-3.76	0.04	-2.87
Stepwise	-1.03	0.11	N/A	10.51	-72.17	9.07	N/A	-5.17	1.34	-0.06	-3.74	0.04	-2.86
Subsets	-1.03	0.11	N/A	10.51	-72.17	9.07	N/A	-5.17	1.34	-0.06	-3.74	0.04	-2.86