Partial Least Squares

Rather than pick directions based solely on the inputs (as PCA does), maybe we should pick directions based on how the variables relate to the response. *Partial Least Squares (PLS)* does this.

After centering the inputs and response, and usually rescaling the inputs to have standard deviation one, we find the first PLS direction from the sample covariance of the response with each of the inputs.

Let the covariance of $\mathbf{x}_j$ with y be $\hat{\varphi}_{1,j} = \mathbf{x}_j^T \mathbf{y}$. If we rescaled inputs to have std. dev. 1, $\hat{\varphi}_{1,j}$ will be the regression coefficient of y on x_j, but this isn't so in general.

We construct a derived input $\mathbf{z}_1 = \sum_{j=1}^{p} \hat{\varphi}_{1,j} \mathbf{x}_j$ pointing in the first PLS direction.

To get the second PLS direction, we modify the inputs by subtracting the projections of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ in the direction of $\mathbf{z}_1$, then repeat the above procedure to obtain $\mathbf{z}_2$, and so forth up to $\mathbf{z}_M$ for some $M < p$.

Finally, we do least-squares linear regression of $\mathbf{y}$ on $\mathbf{z}_1, \ldots, \mathbf{z}_M$.

Question: Does picking *directions* based on $\mathbf{y}$, and then using the *same* $\mathbf{y}$ to find *regression coefficients* produce overfitting? Would using different subsets of training cases for these two operations be better?

Comparison of Linear Regression Methods

The text tries out the methods we've discussed on the prostate cancer dataset, trying to predict `lpsa` from the eight other variables.

Here are the results in terms of average squared error on the test set:

Method	*Parameter*	*Test error*
Least squares with all variables	—	0.586
Least squares with 'best' subset	k = 2	0.574
Ridge regression	df = 4	0.540
Lasso	df = 5	0.491
Principal component regression	M = 7	0.527
Partial least squares	M = 2	0.636

The parameters of the methods were chosen by ten-fold cross validation.

Table 3.3 and Figures 3.5, 3.6, 3.7, and 3.9 in the text have more information.