Non-OLS Linear SEL Prediction: Principal Components Regression

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Linear predictors based on derived inputs

Another possible approach to the problem of finding an appropriate level of complexity in a fitted linear SEL prediction rule is to consider regression on some number $M < p$ of predictors derived from the original inputs x_j. Two such methods are those of Principal Components Regression and Partial Least Squares Regression. Here we continue to assume that **the columns of X have been standardized and Y has been centered**.

Principal components regression

Here the p columns of predictors in $\mathbf{X}$ are replaced with the first M principal components of $\mathbf{X}$

$$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = d_j\mathbf{u}_j$$

The vector of fitted predictions is thus

$$\widehat{\mathbf{Y}}^{\text{PCR}} = \sum_{j=1}^{M} \frac{\langle \mathbf{Y}, \mathbf{z}_j \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle} \mathbf{z}_j = \sum_{j=1}^{M} \langle \mathbf{Y}, \mathbf{u}_j \rangle \mathbf{u}_j$$

Comparing this to ridge regression, we see that ridge regression shrinks the coefficients of the principal components $\mathbf{u}_j$ according to their importance in making up $\mathbf{X}$, while principal components regression "zeros out" those least important in making up $\mathbf{X}$. Since the $\mathbf{u}_j$ constitute an orthonormal basis for $C(\mathbf{X})$, for $rank(\mathbf{X}) = r$,

$$\left\| \widehat{\mathbf{Y}}^{\text{PCR}} \right\|^2 = \sum_{j=1}^{M} \langle \mathbf{Y}, \mathbf{u}_j \rangle^2 \leq \sum_{j=1}^{r} \langle \mathbf{Y}, \mathbf{u}_j \rangle^2 = \left\| \widehat{\mathbf{Y}}^{\text{OLS}} \right\|^2$$

PCR and shrinkage

Further, $\widehat{\mathbf{Y}}^{\text{PCR}}$ can be written in terms of the original inputs as

$$\widehat{\mathbf{Y}}^{\text{PCR}} = \sum_{j=1}^{M} \langle \mathbf{Y}, \mathbf{u}_j \rangle \frac{1}{d_j} \mathbf{X}\mathbf{v}_j = \mathbf{X} \left(\sum_{j=1}^{M} \frac{1}{d_j^2} \langle \mathbf{Y}, \mathbf{X}\mathbf{v}_j \rangle \mathbf{v}_j \right)$$

so that

$$\widehat{\beta}^{\text{PCR}} = \sum_{j=1}^{M} \frac{1}{d_j^2} \langle \mathbf{Y}, \mathbf{X}\mathbf{v}_j \rangle \mathbf{v}_j$$

$\widehat{\beta}^{\text{OLS}}$ is the $M = rank\,(\mathbf{X})$ version of $\widehat{\beta}^{\text{PCR}}$. As the $\mathbf{v}_j$ are orthonormal, it is clear (as on the previous slide for $\widehat{\mathbf{Y}}$) that

$$\left\| \widehat{\beta}^{\text{PCR}} \right\| \leq \left\| \widehat{\beta}^{\text{OLS}} \right\|$$

So principal components regression shrinks both $\widehat{\mathbf{Y}}^{\text{OLS}}$ toward $\mathbf{0}$ in $\Re^N$ and $\widehat{\beta}^{\text{OLS}}$ toward $\mathbf{0}$ in $\Re^p$.

Choice of M and a caveat

The number of components used in PCR is a complexity parameter. The larger is M, the more complex is the predictor. So rational choice of M is in practice based on cross-validation.

There is, of course, nothing guaranteeing that directions in $\Re^p$ of large variation in training set inputs are directions in which y has large rates of change with respect to inputs. That said, PCR is a popular form of "dimension-reduction" for inputs that is sometimes highly effective in sorting out a low-dimensional "signal" in a "large p"-dimensional input.