Lecture 14
PCR and Ridge Regression

04 November 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Goals for today

- ridge regression

Ridge regression

The ridge regression estimator is the solution to the following modified least squares optimization problem for some value of $\lambda > 0$.

$$\widehat{\beta}_{ridge} = \arg\min_{b} \left\{ ||y - Xb||_2^2 + \lambda ||b||_2^2 \right\}$$

Ridge regression

The ridge regression estimator is the solution to the following modified least squares optimization problem for some value of $\lambda > 0$.

$$\widehat{\beta}_{ridge} = \arg\min_{b} \left\{ ||y - Xb||_2^2 + \lambda ||b||_2^2 \right\}$$

The equation shrinks the coefficients towards zero, adding some bias but reducing the variance of the estimator.

Ridge regression has an analytical solution. Write the criterion as a matrix equation:

$$(y - Xb)^t(y - Xb) + \lambda b^t b = y^t y + b^t X^t X b - 2y^t X b + \lambda b^t b$$

Ridge regression has an analytical solution. Write the criterion as a matrix equation:

$$(y - Xb)^t(y - Xb) + \lambda b^t b = y^t y + b^t X^t X b - 2y^t X b + \lambda b^t b$$

And take its derivative:

$$\frac{\partial}{\partial \beta}\left(y^t y + b^t X^t X b - 2y^t X b + \lambda b^t b\right) = 2X^t X b - 2X^t y + 2\lambda b$$

Setting this to zero yields

$$
\begin{aligned}
2X^tX\widehat{\beta} + 2\lambda\widehat{\beta} &= 2X^ty \\
(X^tX + I_p\lambda)\widehat{\beta} &= X^ty \\
\widehat{\beta} &= (X^tX + I_p\lambda)^{-1} \times X^ty
\end{aligned}
$$

Setting this to zero yields

$$
\begin{aligned}
2X^tX\widehat{\beta} + 2\lambda\widehat{\beta} &= 2X^ty \\
(X^tX + I_p\lambda)\widehat{\beta} &= X^ty \\
\widehat{\beta} &= (X^tX + I_p\lambda)^{-1} \times X^ty
\end{aligned}
$$

This is a useful analytical form, though as with least squares we would generally not invert the matrix directly but instead use a stable matrix decomposition.

Now consider the singular value decomposition $U\Sigma V^t$ of the matrix X. We can write the projection matrix P in terms of this as:

$$\begin{aligned} P &= X(X^tX)^{-1}X^t \\ &= U\Sigma V^t(V^t\Sigma^2 V)^{-1}V\Sigma U^t \\ &= U\Sigma V^t V\Sigma^{-2}V^t V\Sigma U^t \\ &= UU^t \end{aligned}$$

Now consider the singular value decomposition $U\Sigma V^t$ of the matrix X. We can write the projection matrix P in terms of this as:

$$\begin{aligned} P &= X(X^tX)^{-1}X^t \\ &= U\Sigma V^t(V^t\Sigma^2 V)^{-1}V\Sigma U^t \\ &= U\Sigma V^t V\Sigma^{-2}V^t V\Sigma U^t \\ &= UU^t \end{aligned}$$

Remember how important the projection matrix was? This is a very important result!

The analogue of the projection matrix for ridge regression is given by:

$$P_\lambda = X(X^tX + \lambda I_p)^{-1}X^t$$

Where P_0 is equal to the ordinary P.

The analogue of the projection matrix for ridge regression is given by:

$$P_\lambda = X(X^tX + \lambda I_p)^{-1}X^t$$

Where P_0 is equal to the ordinary P. As was the case last time, this matrix maps y into the predicted values $\hat{y}$.

What is the decomposition of P_λ in terms of the singular value decomposition?

$$\begin{aligned} P_\lambda &= X(X^tX + \lambda I_p)^{-1}X^t \\ &= U\Sigma V^t(V^t\Sigma^2 V)^{-1}V\Sigma U^t \\ &= U\Sigma(\Sigma^2 + \lambda I_p)^{-1}\Sigma U^t \\ &= UDU^t \end{aligned}$$

For the diagonal matrix D:

$$D = \text{diag}\left(\frac{\sigma_1^2}{\sigma_1^2 + \lambda}, \ldots, \frac{\sigma_p^2}{\sigma_p^2 + \lambda}\right) \tag{1}$$

What is the bias of the ridge regression?

$$
\begin{aligned}
Var\left(\widehat{\beta}|X\right) &= (X^tX + I_p\lambda)^{-1} \times X^tVar(y|X) \\
&= (X^tX + I_p\lambda)^{-1} \times X^tX\beta
\end{aligned}
$$

What is the bias of the ridge regression?

$$\begin{aligned}\mathbb{E}\widehat{\beta} &= (X^tX + I_p\lambda)^{-1} \times X^t\mathbb{E}y \\ &= (X^tX + I_p\lambda)^{-1} \times X^tX\beta\end{aligned}$$