Lecture 14
PCR and Ridge Regression

04 November 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Notes

- problem set 5 posted; due next Wednesday

Notes

- problem set 5 posted; due next Wednesday
- problem set 6 will be due in two weeks, November 18th and is also posted (courtesy of DP)

Notes

- problem set 5 posted; due next Wednesday
- problem set 6 will be due in two weeks, November 18th and is also posted (courtesy of DP)
- the second midterm will be a take-home exam; available online on December 2nd, and due December 7th

Notes

- problem set 5 posted; due next Wednesday
- problem set 6 will be due in two weeks, November 18th and is also posted (courtesy of DP)
- the second midterm will be a take-home exam; available online on December 2nd, and due December 7th
- the final problem set, 7, is formally due the last day of classes (but we'll accept them through December 14th)

Goals for today

- ridge regression formulation and link to SVD
- principal component analysis
- applications to image data

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.

Ridge regression has an analytical solution. To see this 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. To see this 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

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

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:

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

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 $\hat{y}$.

Notice that because VV^t is equal to the identity matrix, we can write the inner term of this projection matrix in a nice form:

$$\begin{aligned} X^t X + \lambda I_p &= V\Sigma^2 V^t + \lambda VV^t \\ &= V(\Sigma^2 + \lambda)V^t \end{aligned}$$

Notice that because VV^t is equal to the identity matrix, we can write the inner term of this projection matrix in a nice form:

$$\begin{aligned} X^tX + \lambda I_p &= V\Sigma^2 V^t + \lambda VV^t \\ &= V(\Sigma^2 + \lambda)V^t \end{aligned}$$

And the inverse is given as:

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

Notice that because VV^t is equal to the identity matrix, we can write the inner term of this projection matrix in a nice form:

$$
\begin{aligned}
X^tX + \lambda I_p &= V\Sigma^2 V^t + \lambda VV^t \\
&= V(\Sigma^2 + \lambda)V^t
\end{aligned}
$$

And the inverse is given as:

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

Where D_λ is a diagonal matrix with entries:

$$
D_\lambda = \text{diag}\left(\frac{1}{\sigma_{max}^2 + \lambda}, \ldots, \frac{1}{\sigma_{min}^2 + \lambda}\right)
$$

Notice that because VV^t is equal to the identity matrix, we can write the inner term of this projection matrix in a nice form:

$$\begin{aligned} X^tX + \lambda I_p &= V\Sigma^2V^t + \lambda VV^t \\ &= V(\Sigma^2 + \lambda)V^t \end{aligned}$$

And the inverse is given as:

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

Where D_λ is a diagonal matrix with entries:

$$D_\lambda = \text{diag}\left(\frac{1}{\sigma_{max}^2 + \lambda}, \ldots, \frac{1}{\sigma_{min}^2 + \lambda}\right)$$

So what have we done to the

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:

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

What is the 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:

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

So we are shrinking the directions of the singular vectors, with more shrinkage on the smaller singular values.

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?

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