Lecture 15
Ridge Regression and PCR

04 November 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

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

- a note on numerical and statistical noise
- ridge regression formulation and link to SVD
- principal component analysis
- applications to image data

Statistical noise as numerical noise

In problem set 5 you are going to construct the pseudo-inverse A^+ of an arbitrary matrix A in terms of A's singular value decomposition.

Statistical noise as numerical noise

In problem set 5 you are going to construct the pseudo-inverse A^+ of an arbitrary matrix A in terms of A's singular value decomposition.

Consider the standard description of a statistical linear model:

$$y = X\beta + \epsilon$$

If we have some sort of inverse of X, we can try to write this as:

$$y' = X(\beta + X^+\epsilon)$$

And now the error is in β rather than in y.

Statistical noise as numerical noise, cont.

It turns out that the y values in the second equation will not be exactly the same as those generated by the original model for the same error terms. However, the least squares estimate of β will be the same.

Statistical noise as numerical noise, cont.

It turns out that the y values in the second equation will not be exactly the same as those generated by the original model for the same error terms. However, the least squares estimate of β will be the same.

So, when considering statistical linear models we know that there is an equivalent problem involving the same X matrix and β vector for which the noise is only due to numerical or measurement error in β.

Statistical noise as numerical noise, cont.

It turns out that the y values in the second equation will not be exactly the same as those generated by the original model for the same error terms. However, the least squares estimate of β will be the same.

So, when considering statistical linear models we know that there is an equivalent problem involving the same X matrix and β vector for which the noise is only due to numerical or measurement error in β.

This is purely to justify why we care about condition numbers and linking concepts in numerical analysis with those in statistics. We would never actually convert the problem to this alternative format, partially because we cannot without knowledge of the error terms.

Ridge regression

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

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

Why the ridge penalty?

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

Why the ridge penalty?

1. The equation shrinks the coefficients towards zero, adding some bias but reducing the variance of the estimator.
2. Using the ℓ_2-norm keeps the equation rotationally invariant.

Why the ridge penalty?

1. The equation shrinks the coefficients towards zero, adding some bias but reducing the variance of the estimator.
2. Using the ℓ_2-norm keeps the equation rotationally invariant.
3. Ridge regression has an analytical solution.

To see this write the criterion as a matrix equation:

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

To see this write the criterion as a matrix equation:

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

And take its derivative:

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

Setting this to zero yields

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

Setting this to zero yields

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

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.

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:

$$\begin{aligned} P &= X(X^tX)^{-1}X^t \\ &= U\Sigma V^t(V\Sigma^2V^t)^{-1}V\Sigma U^t \\ &= U\Sigma V^tV\Sigma^{-2}V^tV\Sigma U^t \\ &= U\begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix}U^t \end{aligned}$$

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:

$$\begin{aligned} P &= X(X^tX)^{-1}X^t \\ &= U\Sigma V^t(V\Sigma^2V^t)^{-1}V\Sigma U^t \\ &= U\Sigma V^tV\Sigma^{-2}V^tV\Sigma U^t \\ &= U\begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix}U^t \end{aligned}$$

This can be written as UU^t if we remember to use the *thin SVD*.

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 $\hat{y}$.

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:

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

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:

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

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:

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

Where E_λ is a diagonal matrix with entries:

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

Remember that the condition number is the ratio of the largest and smallest singular value of a matrix.

Remember that the condition number is the ratio of the largest and smallest singular value of a matrix.What is the condition number of $X^t X$?

$$\text{cond}(X^t X) = \frac{\sigma_{max}^2}{\sigma_{min}^2} \tag{1}$$

Remember that the condition number is the ratio of the largest and smallest singular value of a matrix.What is the condition number of X^tX?

$$\mathrm{cond}(X^tX) = \frac{\sigma^2_{max}}{\sigma^2_{min}} \tag{1}$$

How about the condition number of $X^tX + \lambda I_p$?

$$\mathrm{cond}(X^tX + \lambda I_p) = \frac{\sigma^2_{max} + \lambda}{\sigma^2_{min} + \lambda} \tag{2}$$

Remember that the condition number is the ratio of the largest and smallest singular value of a matrix.What is the condition number of X^tX?

$$\text{cond}(X^tX) = \frac{\sigma^2_{max}}{\sigma^2_{min}} \tag{1}$$

How about the condition number of $X^tX + \lambda I_p$?

$$\text{cond}(X^tX + \lambda I_p) = \frac{\sigma^2_{max} + \lambda}{\sigma^2_{min} + \lambda} \tag{2}$$

How does the incorporation of λ change our ability to invert the matrix?

Back to the projection matrix, what is the decomposition of P_λ in terms of the singular value decomposition?

Back to the projection matrix, what is the decomposition of P_λ in terms of the singular value decomposition?

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

Back to the projection matrix, what is the decomposition of P_λ in terms of the singular value decomposition?

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

For the diagonal matrix D:

$$D = \mathrm{diag}\left(\frac{\sigma_1^2}{\sigma_1^2 + \lambda}, \dots, \frac{\sigma_p^2}{\sigma_p^2 + \lambda}\right) \tag{3}$$

Back to the projection matrix, what is the decomposition of P_λ in terms of the singular value decomposition?

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

For the diagonal matrix D:

$$
D = \operatorname{diag}\left(\frac{\sigma_1^2}{\sigma_1^2 + \lambda}, \dots, \frac{\sigma_p^2}{\sigma_p^2 + \lambda}\right) \tag{3}
$$

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

Finally, and similarly, we can write the solution $\widehat{\beta}_\lambda$ as:

$$\widehat{\beta}_\lambda = V \cdot \operatorname{diag}\left(\frac{\sigma_1}{\sigma_1^2 + \lambda}, \ldots, \frac{\sigma_p}{\sigma_p^2 + \lambda}\right) \cdot U^t y$$

Application of ridge to a single photo

Principal component analysis

The principal components of the matrix X is a linear reparameterization $T = XW$ of the matrix X such that:

Principal component analysis

The principal components of the matrix X is a linear reparameterization $T = XW$ of the matrix X such that:

1. Each new coordinate is uncorrelated with the others; specifically, W is an orthogonal matrix called the *loadings*

Principal component analysis

The principal components of the matrix X is a linear reparameterization $T = XW$ of the matrix X such that:

1. Each new coordinate is uncorrelated with the others; specifically, W is an orthogonal matrix called the *loadings*
2. The first component has the largest variance of all linear combinations of the columns of X, the second has the highest variance conditioned on being uncorrelated with the first, and so forth.

Considering the first column of the matrix W, we can write the condition as follows:

$$\underset{w:\,||w||_2=1}{\arg\max}\ \{||Xw||_2\}$$

Considering the first column of the matrix W, we can write the condition as follows:

$$\underset{w:\, ||w||_2=1}{\arg\max} \; \{||Xw||_2\}$$

However, we already know that this is maximized when w is a multiple of the first right singular vector. That is, the first column of V in the singular value decomposition $U\Sigma V^t$ of X.

Likewise, we can argue that the second column of W is the second column of V, and so forth for all of the principal components.

Likewise, we can argue that the second column of W is the second column of V, and so forth for all of the principal components.

Therefore, the principal components are given by $T = XV$. This gives:

$$\begin{aligned} T &= XV \\ &= U\Sigma V^t V \\ &= U\Sigma \end{aligned}$$

So the components are the weighted columns of the left singular values.

Principal component regression (PCR) uses the first k columns of T as the design matrix, which we will denote $T_k = U_k\Sigma_k$.

Principal component regression (PCR) uses the first k columns of T as the design matrix, which we will denote $T_k = U_k \Sigma_k$. The regression vector is then defined as:

$$\widehat{\beta}_k = V_k (T_k^t T_k)^{-1} T_k^t y$$

Principal component regression (PCR) uses the first k columns of T as the design matrix, which we will denote $T_k = U_k \Sigma_k$. The regression vector is then defined as:

$$\widehat{\beta}_k = V_k (T_k^t T_k)^{-1} T_k^t y$$

Notice that this can be simplified as:

$$\begin{aligned} \widehat{\beta}_k &= V_k (\Sigma_k U^t U \Sigma_k)^{-1} \Sigma_k U_k y \\ &= V_k \Sigma_k^{-1} U_k^t y \end{aligned}$$

On problem set 5, you will show that when k is equal to p, the last line is equal to the ordinary least squares solution.

The variance matrix of the regression vector can be calculated as:

$$\begin{aligned}\mathrm{Var}(V_k\Sigma_k^{-1}U_k^t y) &= \sigma^2 \cdot V_k\Sigma_k^{-1}U_k^t U_k\Sigma_k^{-1}V_k^t \\ &= \sigma^2 \cdot V_k\Sigma_k^{-2}V_k^t\end{aligned}$$

The variance matrix of the regression vector can be calculated as:

$$\begin{aligned}\mathrm{Var}(V_k\Sigma_k^{-1}U_k^t y) &= \sigma^2 \cdot V_k\Sigma_k^{-1}U_k^t U_k\Sigma_k^{-1}V_k^t \\ &= \sigma^2 \cdot V_k\Sigma_k^{-2}V_k^t\end{aligned}$$

And the trace of this is given by:

$$\begin{aligned}\mathrm{tr}\left(\mathrm{Var}\widehat{\beta}_k\right) &= \sigma^2 \cdot \mathrm{tr}\left(V_k\Sigma_k^{-2}V_k^t\right) \\ &= \sigma^2 \cdot \mathrm{tr}\left(\Sigma_k^{-2}\right) \\ &= \sum_{i=1}^{k} \frac{\sigma^2}{\sigma_i^2}\end{aligned}$$

Therefore, we have:

$$\operatorname{tr}\left(Var(\widehat{\beta}_1)\right) \leq \operatorname{tr}\left(Var(\widehat{\beta}_2)\right) \leq \ldots \leq \operatorname{tr}\left(Var(\widehat{\beta}_p)\right) = \operatorname{tr}\left(Var(\widehat{\beta}_{ols})\right)$$

So PCR is another form of variance reduction.

Application of PCR to a single photo

Ridge vs. PCR: similarities

1. Both methods try to reduce variance by using the largest singular values of the design matrix X.

Ridge vs. PCR: similarities

1. Both methods try to reduce variance by using the largest singular values of the design matrix X.
2. Both have easy to compute, analytic solutions.

Ridge vs. PCR: similarities

1. Both methods try to reduce variance by using the largest singular values of the design matrix X.
2. Both have easy to compute, analytic solutions.
3. Efficient method for calculating the solution for multiple values of the tuning parameter. PCR is just one regression for all k and ridge uses the same SVD decomposition, so each λ is just a single matrix multiplication.

Ridge vs. PCR: similarities

1. Both methods try to reduce variance by using the largest singular values of the design matrix X.
2. Both have easy to compute, analytic solutions.
3. Efficient method for calculating the solution for multiple values of the tuning parameter. PCR is just one regression for all k and ridge uses the same SVD decomposition, so each λ is just a single matrix multiplication.
4. Both are invariant to rotations of the data matrix X

Ridge vs. PCR: similarities

1. Both methods try to reduce variance by using the largest singular values of the design matrix X.
2. Both have easy to compute, analytic solutions.
3. Efficient method for calculating the solution for multiple values of the tuning parameter. PCR is just one regression for all k and ridge uses the same SVD decomposition, so each λ is just a single matrix multiplication.
4. Both are invariant to rotations of the data matrix X
5. Both are sensitive to the scale and means of the columns X; typically a good idea to standardize these unless naturally on the same scale to begin with (color pixels is one example)

Ridge vs. PCR: differences

1. Ridge smoothly shrinks the singular vectors whereas PCR just throws out the worst

Ridge vs. PCR: differences

1. Ridge smoothly shrinks the singular vectors whereas PCR just throws out the worst
2. Ridge can be fit for any positive lambda, but there are only p possible values for the tuning parameter in PCR

Ridge vs. PCR: differences

1. Ridge smoothly shrinks the singular vectors whereas PCR just throws out the worst
2. Ridge can be fit for any positive lambda, but there are only p possible values for the tuning parameter in PCR
3. Ridge is, therefore, preferable when being used with a very small λ to simply stabilize the solution rather than perform drastic shrinkage
4. Because we do not know whether β lives in the first k principal components, it is difficult to get any universal results on the bias of PCR.

Ridge vs. PCR: differences

1. Ridge smoothly shrinks the singular vectors whereas PCR just throws out the worst
2. Ridge can be fit for any positive lambda, but there are only p possible values for the tuning parameter in PCR
3. Ridge is, therefore, preferable when being used with a very small λ to simply stabilize the solution rather than perform drastic shrinkage
4. Because we do not know whether β lives in the first k principal components, it is difficult to get any universal results on the bias of PCR.
5. The principal components provide dimension reduction in addition to shrinkage. The PCR can be preferable when you have a large number of variables and but want to preserve some sort of interpretability.

Ridge vs. PCR: differences

1. Ridge smoothly shrinks the singular vectors whereas PCR just throws out the worst
2. Ridge can be fit for any positive lambda, but there are only p possible values for the tuning parameter in PCR
3. Ridge is, therefore, preferable when being used with a very small λ to simply stabilize the solution rather than perform drastic shrinkage
4. Because we do not know whether β lives in the first k principal components, it is difficult to get any universal results on the bias of PCR.
5. The principal components provide dimension reduction in addition to shrinkage. The PCR can be preferable when you have a large number of variables and but want to preserve some sort of interpretability.
6. The principal components are also great for visualizations and as inputs in other machine learning algorithms.

What's next

Amazingly, we only have 4 more lectures before Thanksgiving break.

1. 11-09: Logistic regression revisited
2. 11-11, 11-16, 11-18: Lasso regression