Comparing in-sample and out-of-sample error for ridge regression

Charles Zheng*

August 25, 2015

1 Introduction

1.1 Ordinary least squares

Consider a linear model where $y = X\beta + \epsilon$, with ϵ having independent, zero-mean entries, all with the same variance σ^2. We observe y and X and estimate β by $\hat{\beta} = (X^T X)^{-1} X^T y$. Now consider the problem of predicting an independent set of observations

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

where ϵ^* is an independent copy of ϵ and where the design matrix X is unchanged from before. We can predict the values of y^* by $\hat{y} = X\hat{\beta}$. Define the *in-sample* prediction risk by

$$\mathrm{r}_{in} = \frac{1}{n}\mathbf{E}||\hat{y} - y^*||^2$$

where the term "in-sample" refers to the design matrix X is the same for the predicted observations as for the training data. The in-sample error r_{in} can equally well be defined by the error of prediction for a single observation y^* conditional on observing its covariate x^*, where x^* is drawn uniformly at random among the n rows of X. It is well-known that under the previous assumptions,

$$\mathrm{r}_{in} = \sigma^2\left(1 + \frac{p}{n}\right)$$

*with thanks to Lucas Janson and Zhou Fan

To prove this fact, note that $\hat{y} = Hy$, where H is the projection onto the column space of X, and that $\text{tr}(H) = p$. Then

$$\begin{aligned}
\text{r}_{in} &= \frac{1}{n}\mathbf{E}||y^* - \hat{y}||^2 \\
&= \frac{1}{n}||y^* - Hy||^2 \\
&= \frac{1}{n}||(X\beta + \epsilon^*) - H(X\beta + \epsilon)||^2 \\
&= \frac{1}{n}||(I - H)X\beta + \epsilon^* - H\epsilon||^2 \\
&= \frac{1}{n}||\epsilon^* - H\epsilon||^2 \text{ (since } HX = X) \\
&= \frac{1}{n}||\epsilon||^2 + ||H\epsilon||^2 \\
&= \frac{1}{n}\text{tr}(\sigma^2 I) + \text{tr}(\sigma^2 H) \\
&= \frac{1}{n}\sigma^2(n + p)
\end{aligned}$$

which yields the desired formula.

The concept of in-sample error arises naturally in problems where the design matrix X is fixed, e.g. controlled experiments. In observational data it is more natural to suppose that observations (x_i, y_i) are drawn from some joint distribution F. Supposing we observe i.i.d. realizations (x_i, y_i) are drawn iid from F, then forming the design matrix X by stacking the x_i, we again obtain a least-squares estimate $\hat{\beta}$ for the coefficients of the best linear approximation of y conditional on x. Now suppose we obtain a new independent realization (x^*, y^*) from F, but only observe x^*. As before, we predict $\hat{y} = (x^*)^T\hat{\beta}$, and now we define the average *out-of-sample* prediction risk by

$$\text{r}_{out} = \mathbf{E}(\hat{y} - y^*)^2$$

With suitable assumptions we can derive a similar formula for the out-of-sample risk. The following result is due to Lucas Janson. Suppose that F is a multivariate gaussian with mean 0 and covariance

$$\Sigma_{xy} = \begin{pmatrix} \Sigma & \Sigma\beta \\ \beta^T\Sigma & \beta^T\Sigma\beta + \sigma^2 \end{pmatrix}$$

i.e. $x \sim N(0, \Sigma)$ and $y|x \sim N(x^T\beta, \sigma^2)$. Then using the fact that $(\hat{\beta} -$

$\beta)|X \sim N(0, \sigma^2(X^TX)^{-1})$ we have

$$\begin{aligned}
\mathrm{r}_{out} &= \mathbf{E}(y^* - \hat{y})^2 \\
&= \mathbf{E}(\beta^T x^* + \epsilon^* - \hat{\beta}^T x^*)^2 \\
&= \mathbf{E}((\beta - \hat{\beta})^T x^* + \epsilon^*)^2 \\
&= \mathbf{E}((\beta - \hat{\beta})^T x^*)^2 + \mathbf{E}(\epsilon^*)^2 \\
&= \sigma^2 + \mathbf{E}((\beta - \hat{\beta})^T x^*)^2 \\
&= \sigma^2 + \mathrm{tr}\mathbf{E}(x^*(x^*)^T(\beta - \hat{\beta})(\beta - \hat{\beta})^T)
\end{aligned}$$

using independence of (X, y) and x^*,

$$\begin{aligned}
&= \sigma^2 + \mathrm{tr}[\mathbf{E}[x^*(x^*)^T]\mathbf{E}[(\beta - \hat{\beta})(\beta - \hat{\beta})^T]] \\
&= \sigma^2 + \mathrm{tr}[\Sigma\mathbf{E}[(\beta - \hat{\beta})(\beta - \hat{\beta})^T]] \\
&= \sigma^2 + \mathrm{tr}\mathbf{E}[\Sigma\mathbf{E}[(\beta - \hat{\beta})(\beta - \hat{\beta})^T|X]] \\
&= \sigma^2 + \mathbf{E}[\mathrm{tr}[\Sigma(\sigma^2(X^TX)^{-1})]]] \\
&= \sigma^2 + \sigma^2\mathbf{E}[\mathrm{tr}[\Sigma^{1/2}((X^TX)^{-1})\Sigma^{1/2}]]]
\end{aligned}$$

Note that $\Sigma^{1/2}(X^TX)^{-1}\Sigma^{1/2}$ has an inverse-Wishart distribution with identity scale matrix and n degrees of freedom. Hence

$$\mathbf{E}[\mathrm{tr}[\Sigma^{1/2}((X^TX)^{-1})\Sigma^{1/2}]]] = \frac{p}{n-p-1}$$

and thus

$$\mathrm{r}_{out} = \sigma^2\left(1 + \frac{p}{n-p-1}\right)$$

Comparing with r_{in}, we see that r_{out} is strictly larger, since p/n has been replaced by $p/(n-p-1)$.

1.2 Ridge regression

We see in the OLS case that out-of-sample risk is greater than in-sample risk, with the difference becoming more and more pronounced as p increases relative to n. Hence it is especially interesting to consider the relationship between out-of-sample risk and in-sample risk in an extremely high-dimensional setting. Of course, since OLS cannot be applied when $p > n$, we could only derive the formulas for a method such as ridge regression.

Ridge regression can be used to estimate a linear model when $p > n$ by using the estimator

$$\hat{\beta}_\lambda = (X^T X + n\lambda)^{-1} X^T y$$

where $\lambda > 0$ is a regularization parameter.

Dobriban and Wager (2015) obtain asymptotic expressions for r_{out} of ridge regression; using similar methods, we obtain expressions for the in-sample error r_{in}.

Dobriban and Wager (2015) consider a sequence of multivariate normal models for (x, y), but in which β is also a random variate, and in an asymptotic regime where both p and n grow to infinity, approaching a ratio $\gamma = p/n$. Since p is changing, the covariance matrix Σ_p must be different for each model in the sequence, but one assumes that the distribution of eigenvalues of Σ_p converges in distribution to a limiting eigenvalue distribution $H(\lambda)$ on the real line. Meanwhile, it is assumed that $\beta \sim N(0, \frac{\alpha^2\sigma^2}{p} I)$ so that $\frac{||\beta||^2}{\sigma^2}$ approaches a constant α^2. It is shown that under such a setup, the asymptotically the optimal value of λ is given by

$$\lambda^* = \frac{\gamma}{\alpha^2}$$

and using this value of λ, one obtains

$$\mathrm{r}_{out} = \mathbf{E}(y^* - \hat{\beta}_{\lambda^*}^T x^*)^2 = \sigma^2 \left(\frac{1}{\lambda^* v_{H,\gamma}(-\lambda^*)} \right)$$

where $v_{H,\gamma}$ will be defined below.

Using similar methods we derive an expression for r_{in}. The key fact from random matrix theory we use is that if $\hat{\Sigma}_p$ is the empirical covariance matrix for a sequence of distributions $N(0, \Sigma_p)$ where Σ_p have limiting spectrum $H(\lambda)$, then

$$\lim \frac{1}{p} \mathrm{tr}((\hat{\Sigma}_p - zI_{p\times p})) = m_{H,\gamma}(z)$$

for all $z \in \mathbb{C} \setminus \mathbb{R}^+$, where $m_{H,\gamma}(z)$ is a well-known function from random matrix theory, which can be computed for distribution H from the fixed-point formula

$$m_H(z) = \int_{t=0}^{\infty} \frac{dH(t)}{t(1 - \gamma - \gamma z m z(z)) - z}$$

which is known as the Marchenko-Pasture formula, or Silverstein formula. Meanwhile, the function $v_{H,\gamma}$ appearing in the out-of-sample risk formula is related to $m_{H,\gamma}$ by

$$\gamma(m(z) + 1/z) = v(z) + 1/z$$

The limit

$$\lim \frac{1}{p}\mathrm{tr}((\hat{\Sigma}_p - zI_{p\times p})) = m_{H,\gamma}(z)$$

can also be expressed as

$$\lim \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z} \to m_{H,\gamma}(z)$$

where λ_i are the sample eigenvalues. Hence we also have

$$\begin{aligned}
\lim \frac{1}{p}\mathrm{tr}(\hat{\Sigma}(\hat{\Sigma}_p - zI_{p\times p})) &= \lim \frac{1}{p}\sum_{i=1}^{p}\frac{\lambda_i}{\lambda_i - z} \\
&= \lim \frac{1}{p}\sum_{i=1}^{p}\left(1 + \frac{z}{\lambda_i - z}\right) \\
&= \lim 1 + z\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z} \\
&= 1 + zm_{H,\gamma}(z)
\end{aligned}$$

Our result is as follows. Note that

$$\begin{aligned}
\hat{\beta}_\lambda - \beta &= (X^TX + n\lambda I)^{-1}X^Ty \\
&= ((\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma} - I)\beta + \frac{1}{n}(\hat{\Sigma} + \lambda I)^{-1}X^T\epsilon \\
&= (\hat{\Sigma} + \lambda I)^{-1}\left(-\lambda\beta + \frac{1}{n}X^T\epsilon\right)
\end{aligned}$$

For $y^* = X\beta + \epsilon^*$ where ϵ^* is an independent copy of ϵ, we have (as

$n, p \to \infty$)

$$
\begin{aligned}
\mathrm{r}_{in} \stackrel{def}{=} & \frac{1}{n}\mathbf{E}||y^* - X\hat{\beta}_{\lambda^*}||^2 \\
= & \frac{1}{n}\mathbf{E}||X\beta + \epsilon^* - X\hat{\beta}_{\lambda^*}||^2 \\
= & \sigma^2 + \frac{1}{n}\mathbf{E}||X(\beta - \hat{\beta}_{\lambda^*})||^2 \\
= & \sigma^2 + \frac{1}{n}\mathbf{E}(\beta - \hat{\beta}_{\lambda^*})^T X^T X(\beta - \hat{\beta}_{\lambda^*}) \\
= & \sigma^2 + \mathbf{E}(\beta - \hat{\beta}_{\lambda^*})^T \hat{\Sigma}(\beta - \hat{\beta}_{\lambda^*}) \\
= & \sigma^2 + \mathbf{E}(X^T\epsilon/n - \lambda\beta)^T(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}(X^T\epsilon/n - \lambda\beta)) \\
= & \sigma^2 + (1/n^2)\mathbf{E}[\epsilon^T X(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}X^T\epsilon] \\
& + \lambda^{*2}\mathbf{E}[\beta^T(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\beta] \\
= & \sigma^2 + (\sigma^2/n)\mathrm{tr}\mathbf{E}[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}] \\
& + \lambda^{*2}\frac{\alpha^2\sigma^2}{p}\mathrm{tr}\mathbf{E}[(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}] \\
& = \frac{\sigma^2}{n}\left[\mathrm{tr}\mathbf{E}[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}] + \lambda^*\mathrm{tr}\mathbf{E}[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2}]\right] \\
& = \sigma^2 + \frac{\sigma^2}{n}\mathrm{tr}\mathbf{E}[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}]
\end{aligned}
$$

where in the last line we used

$$
\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1} = \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1} - \lambda\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2}
$$

Thus

$$
\lim_{n\to\infty} \mathrm{r}_{in} = \sigma^2\left(1 + \lim_{n\to\infty}\frac{\gamma}{p}\mathrm{tr}\mathbf{E}[\hat{\Sigma}(\hat{\Sigma} + \lambda^* I)^{-1}]\right) = \sigma^2(1 + \gamma(1 - \lambda^* m(-\lambda^*)))
$$

TODO:

- These formulae have been confirmed numerically. Todo: include the plots and tables
- Interpret the formulae for special cases, e.g. identity covariance and AR-1 covariance
- From these formulae, derive a simpler formula for the relationship of out-of-sample to in-sample risk