A Risk Comparison of Ordinary Least Squares vs Ridge Regression

Paramveer Dhillon[1], Dean P. Foster[2], Sham M. Kakade[2], and Lyle Ungar[1]

[1]Department of Computer Science, University of Pennsylvania
[2]Department of Statistics, Wharton School, University of Pennsylvania

Abstract

We compare the risk of ridge regression to a simple variant of ordinary least squares, in which one simply projects the data onto a finite dimensional subspace (as specified by a Principal Component Analysis) and then performs an ordinary (un-regularized) least squares regression in this subspace. This note shows that the risk of this ordinary least squares method is within a constant factor (namely 4) of the risk of ridge regression.

1 Introduction

Consider the fixed design setting where we have a set of n vectors $\mathcal{X} = \{X_i\}$, and let $\mathbf{X}$ denote the matrix where the i^{th} row of $\mathbf{X}$ is X_i. The observed label vector is $Y \in \mathbb{R}^n$. Suppose that:

$$Y = X\beta + \epsilon$$

where ϵ is independent noise in each coordinate, with the variance of ϵ_i being σ^2.

The objective is to learn $\mathbb{E}[Y] = X\beta$. The expected loss of a vector w is estimator is:

$$L(w) = \frac{1}{n}\mathbb{E}_{\mathbb{Y}}[\|Y - Xw\|^2]$$

Let $\hat{\beta}$ be an estimator of β (constructed with a sample Y). Denoting

$$\mathbf{\Sigma} := \frac{1}{n}\mathbf{X}^T\mathbf{X}$$

we have that the risk (i.e. expected excess loss) is:

$$\text{Risk}(\hat{\beta}) := \mathbb{E}_{\hat{\beta}}[L(\hat{\beta}) - L(\beta)] = \mathbb{E}_{\hat{\beta}}\|\hat{\beta} - \beta\|^2_{\mathbf{\Sigma}}$$

where $\|x\|_{\mathbf{\Sigma}} = x^\top \mathbf{\Sigma} x$ and where the expectation is with respect to the randomness in Y.

We show that a simple variant of ordinary (un-regularized) least squares always compares favorably to ridge regression (as measured by the risk). This observation is based on the following bias variance decomposition:

$$\text{Risk}(\hat{\beta}) = \underbrace{\mathbb{E}\|\hat{\beta} - \bar{\beta}\|^2_{\mathbf{\Sigma}}}_{\text{Variance}} + \underbrace{\|\bar{\beta} - \beta\|^2_{\mathbf{\Sigma}}}_{\text{Prediction Bias}} \tag{1.1}$$

where $\bar{\beta} = \mathbb{E}[\hat{\beta}]$.

1.1 The Risk of Ridge Regression

Ridge regression or Tikhonov Regularization [Tikhonov, 1963] penalizes the ℓ_2 norm of a parameter vector w and "shrinks" β towards zero, penalizing large values more. The estimator is:

$$\hat{\beta}_\lambda = \operatorname*{argmin}_w \{\|Y - \mathbf{X}w\|^2 + \lambda\|w\|^2\}$$

The closed form estimate is then:

$$\hat{\beta}_\lambda = (\mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \left(\frac{1}{n}\mathbf{X}^T Y\right)$$

Note that

$$\hat{\beta}_0 = \hat{\beta}_{\lambda=0} = \operatorname*{argmin}_w \{\|Y - \mathbf{X}w\|^2\}$$

is the ordinary least squares estimator.

Without loss of generality, rotate $\mathbf{X}$ such that:

$$\mathbf{\Sigma} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$$

where λ_i's are ordered in decreasing order.

To see the nature of this shrinkage observe that:

$$[\hat{\beta}_\lambda]_j := \frac{\lambda_j}{\lambda_j + \lambda}[\hat{\beta}_0]_j$$

where $\hat{\beta}_0$ is the ordinary least squares estimator.

Using the bias-variance decomposition, (Equation 1.1), we have that:

Lemma 1. *We have:*

$$\mathrm{Risk}(\hat{\beta}_\lambda) = \frac{\sigma^2}{n}\sum_j \left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1 + \frac{\lambda_j}{\lambda})^2}$$

The proof is straightforward and provided in the appendix.

2 Ordinary Least Squares with PCA

Now let us construct a simple estimator based on λ. Note that our rotated coordinate system where $\mathbf{\Sigma}$ is equal to $diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ corresponds the PCA coordinate system.

Consider the following ordinary least squares estimator on the "top" PCA subspace — it uses the least squares estimate on coordinate j if $\lambda_j \geq \lambda$ and 0 otherwise.

$$[\hat{\beta}_{PCA,\lambda}]_j = \begin{cases} [\hat{\beta}_0]_j & \text{if } \lambda_j \geq \lambda \\ 0 & \text{otherwise} \end{cases}$$

The following claim shows this estimator compares favorably to the ridge estimator (for every λ).

Theorem 2.1. *(Bounded Risk Inflation) For all $\lambda \geq 0$, we have that:*

$$\text{Risk}(\hat{\beta}_{PCA,\lambda}) \leq 4 \ \text{Risk}(\hat{\beta}_\lambda)$$

Proof. Using the bias variance decomposition of the risk we can write the risk as:

$$\text{Risk}(\hat{\beta}_{PCA,\lambda}) = \frac{\sigma^2}{n} \sum_j \mathbb{1}_{\lambda_j \geq \lambda} + \sum_{j:\lambda_j < \lambda} \lambda_j \beta_j^2$$

The first term represents the variance and the second the bias.

The ridge regression risk is given by Lemma 1. We now show that the j^{th} term in the expression for the PCA risk is within a factor 4 of the j^{th} term of the ridge regression risk. First, lets consider the case when $\lambda_j \geq \lambda$, then the ratio of j^{th} terms is:

$$\frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}\left(\frac{\lambda_j}{\lambda_j+\lambda}\right)^2 + \beta_j^2 \frac{\lambda_j}{(1+\frac{\lambda_j}{\lambda})^2}} \leq \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}\left(\frac{\lambda_j}{\lambda_j+\lambda}\right)^2} = \left(1 + \frac{\lambda}{\lambda_j}\right)^2 \leq 4$$

Similarly, if $\lambda_j < \lambda$, the ratio of the j^{th} terms is:

$$\frac{\lambda_j \beta_j^2}{\frac{\sigma^2}{n}\left(\frac{\lambda_j}{\lambda_j+\lambda}\right)^2 + \beta_j^2 \frac{\lambda_j}{(1+\frac{\lambda_j}{\lambda})^2}} \leq \frac{\lambda_j \beta_j^2}{\frac{\lambda_j \beta_j^2}{(1+\frac{\lambda_j}{\lambda})^2}} = \left(1 + \frac{\lambda_j}{\lambda}\right)^2 \leq 4$$

Since, each term is within a factor of 4 the proof is complete. $\square$

3 Conclusion

We showed that the risk inflation of a particular ordinary least squares estimator (on the "top" PCA subspace) is within a factor 4 of the ridge estimator.

References

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl 4*, pages 501–504, 1963.

Appendix

Proof. We analyze the bias-variance decomposition in Equation 1.1. For the variance,

$$\begin{aligned}
\mathbb{E}_Y\|\hat{\beta}_\lambda - \bar{\beta}_\lambda\|_{\mathbf{\Sigma}}^2 &= \sum_j \lambda_j \mathbb{E}_Y([\hat{\beta}_\lambda]_j - [\bar{\beta}_\lambda]_j)^2 \\
&= \sum_j \frac{\lambda_j}{(\lambda_j+\lambda)^2}\frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i])[X_i]_j \sum_{i'=1}^n (Y_i' - \mathbb{E}[Y_i'])[X_i']_j\right] \\
&= \sum_j \frac{\lambda_j}{(\lambda_j+\lambda)^2}\frac{\sigma^2}{n}\sum_{i=1}^n Var(Y_i)[X_i]_j^2 \\
&= \sum_j \frac{\lambda_j}{(\lambda_j+\lambda)^2}\frac{\sigma^2}{n}\sum_{i=1}^n [X_i]_j^2 \\
&= \frac{\sigma^2}{n}\sum_j \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}
\end{aligned}$$

Similarly, for the bias,

$$\begin{aligned}
\|\bar{\beta}_\lambda - \beta\|_{\mathbf{\Sigma}}^2 &= \sum_j \lambda_j([\bar{\beta}_\lambda]_j - [\beta]_j)^2 \\
&= \sum_j \beta_j^2 \lambda_j \left(\frac{\lambda_j}{\lambda_j+\lambda} - 1\right)^2 \\
&= \sum_j \beta_j^2 \frac{\lambda_j}{(1+\frac{\lambda_j}{\lambda})^2}
\end{aligned}$$

which completes the proof. $\square$