A Risk Comparison of Ordinary Least Squares vs Ridge Regression

Paramveer S. Dhillon  
Department of Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104, USA  
DHILLON@CIS.UPENN.EDU

Dean P. Foster  
Department of Statistics  
Wharton School, University of Pennsylvania  
Philadelphia, PA 19104, USA  
FOSTER@WHARTON.UPENN.EDU

Sham M. Kakade  
Microsoft Research  
One Memorial Drive  
Cambridge, MA 02142, USA  
SKAKADE@MICROSOFT.COM

Lyle H. Ungar  
Department of Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104, USA  
UNGAR@CIS.UPENN.EDU

Editor: Gabor Lugosi

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 (PCA-OLS) is within a constant factor (namely 4) of the risk of ridge regression (RR).

Keywords: risk inflation, ridge regression, pca

1. Introduction

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

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

where $\epsilon$ is independent noise in each coordinate, with the variance of $\epsilon_i$ being $\sigma^2$.

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

$$L(\beta) = \frac{1}{n} \mathbb{E}_Y[\|Y - X\beta\|^2],$$
Let \( \hat{\beta} \) be an estimator of \( \beta \) (constructed with a sample \( Y \)). Denoting
\[
\hat{\Sigma} := \frac{1}{n}X^T X,
\]
we have that the risk (i.e., expected excess loss) is:
\[
\text{Risk}(\hat{\beta}) := \mathbb{E}_{\hat{\beta}}[L(\hat{\beta}) - L(\beta)],
\]
where \( L(\beta) = \| \beta - \bar{\beta} \|_2^2 \), 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:
\[
\text{Risk}(\hat{\beta}) = \mathbb{E}_{\hat{\beta}}[\| \hat{\beta} - \bar{\beta} \|_2^2] + \| \bar{\beta} - \beta \|_2^2,
\]
where \( \bar{\beta} = \mathbb{E}[\hat{\beta}] \).

1.1 The Risk of Ridge Regression (RR)

Ridge regression or Tikhonov Regularization (Tikhonov, 1963) penalizes the \( \ell_2 \) norm of a parameter vector \( \beta \) and “shrinks” it towards zero, penalizing large values more. The estimator is:
\[
\hat{\beta}_\lambda = \arg\min_{\beta} \{ \| Y - X\beta \|_2^2 + \lambda \| \beta \|_2^2 \}.
\]

The closed form estimate is then:
\[
\hat{\beta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} \left( \frac{1}{n}X^TY \right).
\]

Note that
\[
\hat{\beta}_0 = \hat{\beta}_{\lambda=0} = \arg\min_{\beta} \{ \| Y - X\beta \|_2^2 \},
\]
is the ordinary least squares estimator.

Without loss of generality, rotate \( X \) such that:
\[
\hat{\Sigma} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p),
\]
where the \( \lambda_i \)'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), we have that:

**Lemma 1**
\[
\text{Risk}(\hat{\beta}_\lambda) = \frac{\sigma^2}{n} \sum_j \left( \frac{\lambda_j}{\lambda_j + \lambda} \right)^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1 + \frac{\lambda_j}{\lambda})^2}.
\]

The proof is straightforward and is provided in the appendix.
2. Ordinary Least Squares with PCA (PCA-OLS)

Now let us construct a simple estimator based on $\lambda$. Note that our rotated coordinate system where $\Sigma$ is equal to $\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ 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 $\lambda$) – no matter how the $\lambda$ is chosen e.g., using cross validation or any other strategy.

Our main theorem (Theorem 2) bounds the Risk Ratio/Risk Inflation of the PCA-OLS and the RR estimators.

**Theorem 2 (Bounded Risk Inflation)** For all $\lambda \geq 0$, we have that:

$$0 \leq \frac{\text{Risk}(\hat{\beta}_{PCA,\lambda})}{\text{Risk}(\hat{\beta}_\lambda)} \leq 4,$$

and the left hand inequality is tight.

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

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

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, let’s consider the case when $\lambda_j \geq \lambda$, then the ratio of $j$th terms is:

$$\frac{\frac{\sigma^2}{n} \left( \frac{\lambda_j}{\lambda_j + \lambda} \right)^2}{\frac{\sigma^2}{n} \left( \frac{\lambda_j}{\lambda_j + \lambda} \right)^2} \leq \frac{\text{Risk}(\hat{\beta}_\lambda)}{\text{Risk}(\hat{\beta}_{PCA,\lambda})} = \left( 1 + \frac{\lambda}{\lambda_j} \right)^2 \leq 4.$$

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

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

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

It is worth noting that the converse is not true and the ridge regression estimator (RR) can be arbitrarily worse than the PCA-OLS estimator. An example which shows that the left hand inequality is tight is given in the Appendix.

1. Risk Inflation has also been used as a criterion for evaluating feature selection procedures [Foster and George, 1994].
3. Experiments

First, we generated synthetic data with $p = 100$ and varying values of $n = \{20, 50, 80, 110\}$. The data was generated in a fixed design setting as $Y = X\beta + \epsilon$ where $\epsilon_i \sim N(0,1)$ $\forall i = 1, \ldots, n$. Furthermore, $X_{n \times p} \sim MVN(0, I)$ where $MVN(\mu, \Sigma)$ is the Multivariate Normal Distribution with mean vector $\mu$, variance-covariance matrix $\Sigma$ and $\beta_j \sim N(0,1)$ $\forall j = 1, \ldots, p$.

The results are shown in Figure 1. As can be seen, the risk ratio of PCA (PCA-OLS) and ridge regression (RR) is never worse than 4 and often its better than 1 as dictated by Theorem 2.

Next, we chose two real world datasets, namely USPS ($n=1500$, $p=241$) and BCI ($n=400$, $p=117$).

Since we do not know the true model for these datasets, we used all the $n$ observations to fit an OLS regression and used it as an estimate of the true parameter $\beta$. This is a reasonable approximation to the true parameter as we estimate the ridge regression (RR) and PCA-OLS models on a small subset of these observations. Next we choose a random subset of the observations, namely $0.2 \times p$, $0.5 \times p$ and $0.8 \times p$ to fit the ridge regression (RR) and PCA-OLS models.

The results are shown in Figure 2. As can be seen, the risk ratio of PCA-OLS to ridge regression (RR) is again within a factor of 4 and often PCA-OLS is better i.e., the ratio $< 1$.

4. 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. It turns out the converse is not true — this PCA estimator may be arbitrarily better than the ridge one.

Appendix A.

Proof of Lemma 1.

Proof. We analyze the bias-variance decomposition in Equation 1. For the variance,

$$\mathbb{E}_{Y} \| \hat{\beta}_\lambda - \bar{\beta}_\lambda \|_2^2 = \sum_j \lambda_j \mathbb{E}_{Y} ( (\hat{\beta}_\lambda)_j - (\bar{\beta}_\lambda)_j )^2$$

$$= \sum_j \frac{\lambda_j}{(\lambda_j + \lambda)^2} \frac{1}{n^2} \mathbb{E} \left[ \sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i])[X_i]_j \sum_{i'=1}^{n} (Y_i' - \mathbb{E}[Y_i'])[X_i']_j \right]$$

$$= \sum_j \frac{\lambda_j}{(\lambda_j + \lambda)^2} \frac{\sigma^2}{n} \sum_{i=1}^{n} Var(Y_i)[X_i]_j^2$$

$$= \sum_j \frac{\lambda_j}{(\lambda_j + \lambda)^2} \frac{\sigma^2}{n} \sum_{i=1}^{n} [X_i]_j^2$$

$$= \frac{\sigma^2}{n} \sum_j \frac{\lambda_j^2}{(\lambda_j + \lambda)^2}.$$
Figure 1: Plots showing the risk ratio as a function of $\lambda$, the regularization parameter and $n$, for the synthetic dataset. $p=100$ in all the cases. The error bars correspond to one standard deviation for 100 such random trials.

Figure 2: Plots showing the risk ratio as a function of $\lambda$, the regularization parameter and $n$, for two real world datasets (BCI and USPS—top to bottom).
Similarly, for the bias,
\[
\|\hat{\beta}_\lambda - \beta\|_2^2 = \sum_j \lambda_j (|\hat{\beta}_\lambda(j)| - |\beta_j|)^2
\]
\[
= \sum_j \beta_j^2 \lambda_j \left(\frac{\lambda_j}{\lambda_j + \lambda} - 1\right)^2
\]
\[
= \sum_j \beta_j^2 \frac{\lambda_j}{(1 + \frac{\lambda_j}{\lambda})^2},
\]
which completes the proof.

The risk for RR can be arbitrarily worse than the PCA-OLS estimator.

Consider the standard OLS setting described in Section 1 in which \(X\) is \(n \times p\) matrix and \(Y\) is a \(n \times 1\) vector.

Let \(X = diag(\sqrt{1 + \alpha}, 1, \ldots, 1)\), then \(\Sigma = X^T X = diag(1 + \alpha, 1, \ldots, 1)\) for some \((\alpha > 0)\) and also choose \(\beta = [2 + \alpha, 0, \ldots, 0]\). For convenience let’s also choose \(\sigma^2 = n\).

Then, using Lemma 1, we get the risk of RR estimator as

\[
\text{Risk}(\hat{\beta}_\lambda) = \left(\frac{1 + \alpha}{1 + \alpha + \lambda}\right)^2 + \left(\frac{p - 1}{(1 + \lambda)^2}\right) + (2 + \alpha)^2 \times \left(1 + \frac{\alpha}{(1 + \frac{1+\alpha}{\lambda})^2}\right).
\]

Let’s consider two cases

- **Case 1**: \(\lambda < (p - 1)^{1/3} - 1\), then \(II > (p - 1)^{1/3}\).
- **Case 2**: \(\lambda > 1\), then \(1 + \frac{1+\alpha}{\lambda} < 2 + \alpha\), hence \(III > (1 + \alpha)\).

Combining these two cases we get \(\forall \lambda, \text{Risk}(\hat{\beta}_\lambda) > \text{min}((p - 1)^{1/3}, (1 + \alpha))\). If we choose \(p\) such that \(p - 1 = (1 + \alpha)^3\), then \(\text{Risk}(\hat{\beta}_\lambda) > (1 + \alpha)\).

The PCA-OLS risk (From Theorem 2) is:

\[
\text{Risk}(\hat{\beta}_{PCA,\lambda}) = \sum_j \mathbb{1}_{\lambda_j \geq \lambda} + \sum_{j: \lambda_j < \lambda} \lambda_j \beta_j^2.
\]

Considering \(\lambda \in (1, 1 + \alpha)\), the first term will contribute 1 to the risk and rest everything will be 0. So the risk of PCA-OLS is 1 and the risk ratio is

\[
\frac{\text{Risk}(\hat{\beta}_{PCA,\lambda})}{\text{Risk}(\hat{\beta}_\lambda)} \leq \frac{1}{(1 + \alpha)}.
\]

Now, for large \(\alpha\), the risk ratio \(\approx 0\).
References
