**Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods** **Lecture: 6**

# Dimensionality Reduction and Learning: Ridge Regression vs. PCA

*Instructor: Sham Kakade*

## 1 Intro

The theme of these two lectures is that for $L_2$ methods we need not work in infinite dimensional spaces. In particular, we can unadaptively find and work in a low dimensional space and achieve about as good results. These results question the need for explicitly working in infinite (or high) dimensional spaces for $L_2$ methods. In contrast, for sparsity based methods (including $L_1$ regularization), such non-adaptive projection methods significantly loose predictive power.

## 2 Ridge Regression and Dimensionality Reduction

This lecture will characterize the risk of ridge regression (in infinite dimensions) in terms of a bias-variance tradeoff. Furthermore, we will show that a simple dimensionality reduction scheme, simply based on PCA, along with just MLE estimates (in this projected space) performs nearly as well as ridge regression.

## 3 Risk and Fixed Design Regression

Let us now consider the 'normal means' problem, sometimes referred to as the fixed design setting. Here, we have a set of $n$ points $\mathcal{X} = \{X_i\} \subset \mathbb{R}^d$, and let $X$ denote the $\mathbb{R}^{n\times d}$ matrix where the $i$ row of $X$ is $X_i$. We also observe a output vector $Y \in \mathbb{R}^n$. We desire to learn $\mathbb{E}[Y]$. In particular, we seek to predict $\mathbb{E}[Y]$ as $X\hat{\beta}$.

The square loss of an estimator $w$ is:

$$L(w) = \frac{1}{n}\mathbb{E}_Y\|Y - Xw\|^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i - X_i w)^2$$

where the expectation is with respect to $Y$. Let $\beta$ be the optimal predictor:

$$\beta = \arg\min_w L(w)$$

The risk of an estimator $\hat{\beta}$ is defined as:

$$R(\hat{\beta}) = L(\hat{\beta}) - L(\beta) = \frac{1}{n}\|X\hat{\beta} - X\beta\|^2$$

(which is the fixed design risk). Denoting,

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

we can write the risk as:

$$R(\hat{\beta}) = (\hat{\beta} - \beta)^\top\Sigma(\hat{\beta} - \beta) := \|\hat{\beta} - \beta\|_\Sigma^2$$

Another interpretation of the risk is how well we accurately learn the parameters of the model.

Assume that $\hat{\beta}(Y)$ is an estimator constructed with the outcome $Y$ — we drop the explicit $Y$ dependence as this is clear from context. Let $\overline{\beta} = \mathbb{E}_Y \hat{\beta}$ be expected weight. We can decompose the expected risk as:

$$\begin{aligned}\mathbb{E}_Y[R(\hat{\beta})] &= \frac{1}{n}\mathbb{E}_Y\|X\hat{\beta} - X\overline{\beta}\|^2 + \frac{1}{n}\|X\overline{\beta} - X\beta\|^2 \\ &= \mathbb{E}_Y\|\hat{\beta} - \overline{\beta}\|_\Sigma^2 + \|\overline{\beta} - \beta\|_\Sigma^2\end{aligned}$$

where we have that:

$$\text{(average) variance} = \frac{1}{n}\mathbb{E}_Y\|X\hat{\beta} - X\overline{\beta}\|^2$$

and

$$\text{prediction bias vector} = X\overline{\beta} - X\beta$$

which shows a certain bias/variance decomposition of the error.

## 3.1 Risk Bounds for Ridge Regression

The ridge regression estimator using an outcome $Y$ is just:

$$\hat{\beta}_\lambda = \arg\min_w \frac{1}{n}\|Y - Xw\|^2 + \lambda\|w\|^2$$

The estimator is then:

$$\hat{\beta}_\lambda = (\Sigma + \lambda I)^{-1}(\frac{1}{n}X^\top Y) = (\Sigma + \lambda I)^{-1}(\frac{1}{n}\sum Y_i X_i^\top)$$

For simplicity, let us rotate $X$ such that:

$$\Sigma := \frac{1}{n}X^\top X = diag(\lambda_1, \lambda_2, \ldots \lambda_d)$$

(note this rotation does not alter the predictions of rotationally invariant algorithms). With this choice, we have that:

$$[\hat{\beta}_\lambda]_j = \frac{\frac{1}{n}\sum_{i=1}^n Y_i[X_i]_j}{\lambda_j + \lambda}$$

It is straightforward to see that:

$$\beta = E[\hat{\beta}_0]$$

and it follows that:

$$[\overline{\beta}_\lambda]_j := \mathbb{E}[\hat{\beta}_\lambda]_j = \frac{\lambda_j}{\lambda_j + \lambda}\beta_j$$

by just taking expectations.

**Lemma 3.1.** *(Risk Bound) If* $\mathrm{Var}(Y_i) \le 1$*, we have that:*

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \le \frac{1}{n}\sum_j(\frac{\lambda_j}{\lambda_j + \lambda})^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1 + \lambda_j/\lambda)^2}$$

*This holds with equality if* $\mathrm{Var}(Y_i) = 1$.

*Proof.* For the variance term, we have:

$$\begin{aligned}
\mathbb{E}_Y\|\hat{\beta}_\lambda - \overline{\beta}_\lambda\|_\Sigma^2 &= \sum_j \lambda_j \mathbb{E}_Y([\hat{\beta}_\lambda]_j - [\overline{\beta}_\lambda]_j)^2 \\
&= \sum_j \frac{\lambda_j}{(\lambda_j+\lambda)^2}\frac{1}{n^2}\mathbb{E}[\sum_{i=1}^n (Y_i - E[Y_i])[X_i]_j \sum_{i'=1}^n (Y_{i'} - E[Y_{i'}])[X_{i'}]_j] \\
&= \sum_j \frac{\lambda_j}{(\lambda_j+\lambda)^2}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Y_i)[X_i]_j^2 \\
&\leq \sum_j \frac{\lambda_j}{(\lambda_j+\lambda)^2}\frac{1}{n}\sum_{i=1}^n [X_i]_j^2 \\
&= \frac{1}{n}\sum_j \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}
\end{aligned}$$

This holds with equality if $\mathrm{Var}(Y_i) = 1$. For the bias term,

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

and the result follows from algebraic manipulations. □

There following bound characterizes the risk for two natural settings for $\lambda$.

**Corollary 3.2.** *Assume* $\mathrm{Var}(Y_i) \leq 1$

- *(Finite Dims) For* $\lambda = 0$*,*

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \leq \frac{d}{n}$$

*And if* $Var(Y_i) = 1$*, then* $\mathbb{E}_Y[R(\hat{\beta}_\lambda)] = \frac{d}{n}$.

- *(Infinite Dims) For* $\lambda = \frac{\sqrt{\|\Sigma\|_{trace}}}{\|\beta\|\sqrt{n}}$*, then:*

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \leq \frac{\|\beta\|\sqrt{\|\Sigma\|_{trace}}}{2\sqrt{n}} = \frac{\|\beta\|\sqrt{\frac{1}{n}\sum_i \|X_i\|^2}}{2\sqrt{n}} \leq \frac{\|\beta\|\|\mathcal{X}\|}{2\sqrt{n}}$$

*where the trace norm is the sum of the singular values and* $\|\mathcal{X}\| = \max_i \|X_i\|^2$*. Furthermore, for all* $n$ *there exists a distribution* $\Pr[Y]$ *and an* $X$ *such that the* $\inf_\lambda \mathbb{E}_Y[R(\hat{\beta}_\lambda)]$ *is* $\Omega^*(\frac{\|\beta\|\sqrt{\|\Sigma\|_{trace}}}{2\sqrt{n}})$ *(so the above bound is tight up to log factors).*

Conceptually, the second bound is 'dimension free', i.e. it does not depend explicitly on $d$, which could be infinite. And we are effectively doing regression in a large (potentially) infinite dimensional space.

*Proof.* The $\lambda = 0$ case follows directly from the previous lemma. Using that $(a+b)^2 \geq 2ab$, we can bound the variance term for general $\lambda$ as follows:

$$\frac{1}{n}\sum_j (\frac{\lambda_j}{\lambda_j+\lambda})^2 \leq \frac{1}{n}\sum_j \frac{\lambda_j^2}{2\lambda_j\lambda} = \frac{\sum_j \lambda_j}{2n\lambda}$$

Again, using that $(a+b)^2 \geq 2ab$, the bias term is bounded as:

$$\sum_j \beta_j^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2} \leq \sum_j \beta_j^2 \frac{\lambda_j}{2\lambda_j/\lambda} = \frac{\lambda}{2}||\beta||^2$$

So we have that:

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] \leq \frac{\|\Sigma\|_{\text{trace}}}{2n\lambda} + \frac{\lambda}{2}||\beta||^2$$

and using the choice of $\lambda$ completes the proof.

To see the above bound is tight, consider the following problem. Let $X_i = \sqrt{\frac{n}{i}}$ and $\beta_i = \sqrt{\frac{1}{i}}$ and let $Y = X\beta + \eta$ where $\eta$ is unit variance. Here, we have that $\lambda_i = \frac{1}{i}$ so $\sum_j \lambda_j \leq \log n$ and $\|\beta\|^2 \leq \log n$, so the upper is $\frac{\log n}{\sqrt{n}}$. Now one can write the risk as:

$$\begin{aligned}
\mathbb{E}_Y[R(\hat{\beta}_\lambda)] &= \frac{1}{n}\sum_j (\frac{\frac{1}{i}}{\frac{1}{i}+\lambda})^2 + \sum_j \frac{\frac{1}{i^2}}{(1+\frac{1}{i\lambda})^2} && (1)\\
&= \sum_j \frac{\frac{1}{n}+\lambda^2}{(1+i\lambda)^2} && (2)\\
&\geq \int_1^n \frac{\frac{1}{n}+\lambda^2}{(1+x\lambda)^2} dx && (3)\\
&= (\frac{1}{n}+\lambda^2)(\frac{1}{\lambda(1+\lambda)} - \frac{1}{\lambda(1+n\lambda)}) && (4)\\
&= (\frac{1}{n\lambda}+\lambda)(\frac{1}{1+\lambda} - \frac{1}{1+n\lambda}) && (5)\\
& && (6)
\end{aligned}$$

and this is $\Omega(\sqrt{n})$, for all $\lambda$. □

However, now we show that with $L_2$ complexity, we can effectively working in finite dimensions (where the dimension is chosen as a function of $n$).

# 4 PCA Projections and MLEs

Fix some $\lambda$. Consider the following 'keep or kill' estimator, which uses the MLE estimate if $\lambda_i \geq \lambda$ and $0$ otherwise:

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

where $\hat{\beta}_0$ is the MLE estimator. This estimator is $0$ for the small values of $\lambda_i$ (those in which we are effectively regularizing more anyways).

**Theorem 4.1.** *(Risk Inflation of $\hat{\beta}_{PCA,\lambda}$)*
*Assume* $\text{Var}(Y_i) = 1$*, then*

$$\mathbb{E}_Y[R(\hat{\beta}_{PCA,\lambda})] \leq 4\mathbb{E}_Y[R(\hat{\beta}_\lambda)]$$

Note that the the actual risk (not just an upper bound) of the simple PCA estimate is within a factor of $4$ of the ridge regression risk on a wide class of problems.

*Proof.* Recall that:

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] = \frac{1}{n}\sum_j (\frac{\lambda_j}{\lambda_j+\lambda})^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2}$$

Since we can write the risk as:

$$\mathbb{E}_Y[R(\hat{\beta})] = \mathbb{E}_Y\|\hat{\beta} - \overline{\beta}\|_\Sigma^2 + \|\overline{\beta} - \beta\|_\Sigma^2$$

we have that:

$$\mathbb{E}_Y[R(\hat{\beta}_{PCA,\lambda})] = \frac{1}{n}\sum_j \mathbb{I}(\lambda_j > \lambda) + \sum_{j:\lambda_j<\lambda} \lambda_j \beta_j^2$$

where $\mathbb{I}$ is the indicator function.

We now show that each term in the risk of $\hat{\beta}_{PCA,\lambda}$ is within a factor of $4$ for each term in $\hat{\beta}_\lambda$. If $\lambda_j > \lambda$, then the ratio of the $j-th$ terms is:

$$\begin{aligned}
\frac{\frac{1}{n}}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2 + \beta_j^2\frac{\lambda_j}{(1+\lambda_j/\lambda)^2}} &\leq \frac{\frac{1}{n}}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2} \\
&= \frac{(\lambda_j+\lambda)^2}{\lambda_j^2} \\
&\leq (1+\frac{\lambda}{\lambda_j})^2 \\
&\leq 4
\end{aligned}$$

Similarly, if $\lambda_j \leq \lambda$, then the ratio of the $j$-th terms is:

$$\begin{aligned}
\frac{\lambda_j\beta_j^2}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2 + \frac{\lambda_j\beta_j^2}{(1+\lambda_j/\lambda)^2}} &\leq \frac{\lambda_j\beta_j^2}{\frac{\lambda_j\beta_j^2}{(1+\lambda_j/\lambda)^2}} \\
&= (1+\lambda_j/\lambda)^2 \\
&\leq 4
\end{aligned}$$

Since each term is within a factor of $4$, the proof is completed. □

# References

The observation about the risk inflation of ridge regression vs. PCA was first pointed out to my by Dean Foster.