An introduction to high-dimensional statistics

Rajen Shah

5th March 2012

In this note, we aim to give a very brief introduction to high-dimensional statistics. Rather than attempting to give an overview of this vast area, we will explain what is meant by high-dimensional data and then focus on two methods which have been introduced to deal with this sort of data. Many of the state of the art techniques used in high-dimensional statistics today are based on these two core methods.

We begin with a quick recap of least squares regression.

1 Introduction

We consider the setting where we have observed data $(y_1, x_1), \ldots, (y_n, x_n)$ with each y_i a realisation of a scalar random variable Y_i, and each $x_i = (x_{i1}, \ldots, x_{ip})^T$ a p-vector of explanatory variables. Let X be a matrix whose i^{th} row is given by x_i^T. Let us assume that

$$Y_i = \mu + (X\beta)_i + \epsilon_i,$$

where each ϵ_i is a random variable with $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2 > 0$ and $\mathbb{E}(\epsilon_i \epsilon_k) = 0$ for $i \neq k$. The parameters μ and β are unknown, and least squares regression gives us a way of estimating them. Without loss of generality, we may assume that the columns of X are centred (i.e. their components sum to 0).

We will require that X has full column rank. Note that this implies $p < n$. The least squares esimate of β is given by

$$(\hat{\mu}, \hat{\beta}) = \underset{(m,b)}{\arg\min} \|Y - m\mathbf{1} - Xb\|^2,$$

so

$$\hat{\mu} = \bar{Y}, \quad \hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (Y - \bar{Y}).$$

Why is this a sensible choice? First of all, $(\hat{\mu}, \hat{\beta})$ is *unbiased*. That is, $\mathbb{E}_{\mu,\beta}\{(\hat{\mu}, \hat{\beta})\} = (\mu, \beta)$. So, 'on average least squares regression does the right thing'. This is helpful, but much more is true.

Theorem 1 (Gauss–Markov Theorem)**.** *If $\tilde{\beta}$ is any linear, unbiased estimator of β, we have that* $\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta})$ *is positive semidefinite.*

Remark 1.

1. We have excluded $\hat{\mu}$ from consideration here simply to ease notation.

2. Note that, for any new vector of explanatory variables x_{new}, we have that $\mathbb{E}(x_{\text{new}}^T\beta - x_{\text{new}}^T\tilde{\beta})^2 \geq \mathbb{E}(x_{\text{new}}^T\beta - x_{\text{new}}^T\hat{\beta})^2$.

Proof. Let $\tilde{\beta} = \{(X^TX)^{-1}X^T + A^T\}Y$. As $\hat{\beta}$ is unbiased, we have

$$\mathbb{E}_{\mu,\beta}(\tilde{\beta}) = \beta + A^T(\mu\mathbf{1} + X\beta).$$

Thus in order for $\tilde{\beta}$ to be unbiased, we must have that the columns of A are centred, and $A^TX = 0$. Now we compute the variances.

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\{((X^TX)^{-1}X^T + A)\epsilon\} - \mathrm{Var}((X^TX)^{-1}X^T\epsilon) \\
&= \sigma^2((X^TX)^{-1}X^T + A)((X^TX)^{-1}X^T + A)^T - \sigma^2(X^TX)^{-1} \\
&= \sigma^2AA^T
\end{aligned}$$

□

One might think that this is the last word on linear regression, and that least squares is unbeatable. However, the theorem only compared least squares to other *unbiased* estimators of β. Can we do better by introducing some bias? It turns out that the answer is yes, and particularly with high-dimensional data, we can do much better. We now discuss what is meant by high-dimensional data.

2 High-dimensional data

The dimension of data refers to the number of explanatory variables, p. In classical statistics, we imagine data being composed of a handful of carefully chosen covariates, and we imagine that the number of observations n, greatly exceeds p. However, in many statistical applications today, the situation is quite the opposite, with p sometimes much larger than n. In these situations, least squares regression simply does not work.

In fact, even as p approaches n, things can go wrong. If p is large, some of the columns of X are likely to be nearly collinear, making X^TX ‘almost singular’. In this situation, the inverse $(X^TX)^{-1}$ will have some very large eigenvalues, and consequently the least squares estimator will exhibit high variance and thus be a poor estimator of β. By sacrificing some bias, we can reduce the variance and produce a better estimator, as we shall see in the next section.

3 Ridge regression (Hoerl and Kennard, 1970)

For a fixed $\lambda > 0$, the ridge regression estimator is $\hat{\beta}^{\mathrm{R}}_{\lambda}$, where

$$(\hat{\mu}, \hat{\beta}^{\mathrm{R}}_{\lambda}) = \arg\min_{m,b} \left\{ \|Y - m\mathbf{1} - Xb\|^2 + \lambda \|b\|^2 \right\}.$$

This estimator shrinks the estimated coefficients towards 0 by an amount determined by λ, thus reducing the variance of the estimator. We have that, as before, $\hat{\mu} = \bar{Y}$, and

$$\hat{\beta}^{\mathrm{R}}_{\lambda} = (X^T X + \lambda I)^{-1} X^T Y.$$

In this form, we can see how the addition of the λI term helps to stabilise the estimator. Note that when X does not have full column rank (such as in high-dimensional situations), we can still compute this estimator. On the other hand, when X does have full column rank, we have the following theorem.

Theorem 2. *There exists a λ such that*

$$\mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T - \mathbb{E}(\hat{\beta}^{\mathrm{R}}_{\lambda} - \beta)(\hat{\beta}^{\mathrm{R}}_{\lambda} - \beta)^T$$

is positive definite.

Remark 2. In particular, $\mathbb{E}(x^T_{\mathrm{new}}\beta - x^T_{\mathrm{new}}\hat{\beta})^2 > \mathbb{E}(x^T_{\mathrm{new}}\beta - x^T_{\mathrm{new}}\hat{\beta}^{\mathrm{R}}_{\lambda})^2$.

Proof. First we compute the bias of $\hat{\beta}^{\mathrm{R}}_{\lambda}$.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}^{\mathrm{R}}_{\lambda}) - \beta &= (X^T X + \lambda I)^{-1} X^T X \beta - \beta \\
&= (X^T X + \lambda I)^{-1} (X^T X + \lambda I - \lambda I)\beta - \beta \\
&= -\lambda (X^T X + \lambda I)^{-1} \beta.
\end{aligned}$$

Now we look at the variance of $\hat{\beta}^{\mathrm{R}}_{\lambda}$.

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}^{\mathrm{R}}_{\lambda}) &= \mathbb{E}\{(X^T X + \lambda I)^{-1} X^T \epsilon\}\{(X^T X + \lambda I)^{-1} X^T \epsilon\}^T \\
&= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}.
\end{aligned}$$

Thus $\mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T - \mathbb{E}(\hat{\beta}^{\mathrm{R}}_{\lambda} - \beta)(\hat{\beta}^{\mathrm{R}}_{\lambda} - \beta)^T$ is equal to

$$\sigma^2 (X^T X)^{-1} - \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} - \lambda^2 (X^T X + \lambda I)^{-1} \beta \beta^T (X^T X + \lambda I)^{-1}.$$

After some simplification, we see that this is equal to

$$\lambda (X^T X + \lambda I)^{-1} [\sigma^2 \{2I + \lambda (X^T X)^{-1}\} - \lambda \beta \beta^T] (X^T X + \lambda I)^{-1}.$$

Thus $\mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T - \mathbb{E}(\hat{\beta}^{\mathrm{R}}_\lambda - \beta)(\hat{\beta}^{\mathrm{R}}_\lambda - \beta)^T$ is positive definite for $\lambda > 0$ if and only if

$$\sigma^2\{2I + \lambda(X^T X)^{-1}\} - \lambda\beta\beta^T$$

is positive definite, which is true for $\lambda > 0$ sufficiently small. □

4 The Lasso (Tibshirani, 1996)

In very high dimensional situations, we often want to do variable selection. That is, we want to set some of the estimated coefficients to be exactly 0. This is essential for model interpretation, but often we also believe that the response depends on only a few of the covariates. The Lasso estimator replaces the penalty term $\|b\|^2$ in the ridge regression optimisation, by $\|b\|_1$. Crucially, the use of the ℓ_1 penalty results in sparse estimates for β. The Lasso estimator has been hugely influential in high-dimensional statistics and the original Lasso paper has over 5000 citations.