Lecture Notes 23
36-705

We begin with a quick review of low dimensional linear regression, before turning our attention to high-dimensional regression with the LASSO.

1 Low Dimensional Linear Regression – Review

Linear regression is a tool to approximate the conditional expectation

$$\mu(x) = \mathbb{E}[Y|X = x]$$

with a linear function of X. We observe iid pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we use the model

$$Y_i = \beta^T X_i + \epsilon_i,$$

where $Y_i \in \mathbb{R}, X_i \in \mathbb{R}^d$ and $\epsilon_i \sim N(0, \sigma^2)$. Usually the first coordinate of X_i is 1 for all i which means that β_1 is the intercept. The design matrix is the $n \times d$ matrix $\mathbb{X}$ where $\mathbb{X}_{ij}$ is the j^{th} coordinate of X_i. We let

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T = \frac{1}{n}\mathbb{X}^T\mathbb{X}.$$

Least Squares: The least squares estimator is

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2$$

and is given by

$$\widehat{\beta} = \widehat{\Sigma}^{-1}\widehat{\alpha} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

where $\widehat{\alpha} = \frac{1}{n}\sum_i X_i Y_i$ and $\mathbb{Y} = (Y_1, \ldots, Y_n)$.

Exercise. Show that the least squares estimator is also the (conditional) maximum likelihood estimator, that is, $\widehat{\beta}$ maximizes $\prod_i p(Y_i|X_i; \beta)$.

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$. Many of the calculations are done conditional on $\mathcal{X}$. This simplifies the calculations but it leads to valid inferences. For example, suppose that C is a confidence set for β conditional on $\mathcal{X}$. So $P(\beta \in C|\mathcal{X}) = 1 - \alpha$. Then

$$P(\beta \in C) = \mathbb{E}[P(\beta \in C|\mathcal{X})] = 1 - \alpha.$$

We can write

$$\mathbb{Y} = \mathbb{X}\beta + \epsilon$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. So, conditional on $\mathcal{X}$,

$$\begin{aligned}\widehat{\beta} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\beta + \epsilon)\\ &= \beta + (\mathbb{X}^T\mathbb{X})^{-1}\epsilon \overset{d}{=} N(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}).\end{aligned}$$

This implies that, conditional on $\mathcal{X}$, $\widehat{\beta}_j \sim N(\beta_j, \tau_j^{)}$ where τ_j^2 is the j^{th} diagonal element of $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$. A consistent estimaator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-d}\sum_i \widehat{\epsilon}_i^2$$

where $\widehat{\epsilon}_i = Y_i - \widehat{\beta}^T X_i$. Hence, an asymptotic $1-\alpha$ confidence interval is $\widehat{\beta}_j \pm z_{\alpha/2}\widehat{\tau}$. The fitted values are

$$\widehat{\mathbb{Y}} = \mathbb{X}\widehat{\beta} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T$$

where $\widehat{Y}_i = \widehat{\beta}^T X_i$. Note that

$$\widehat{\mathbb{Y}} = HY$$

where

$$H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$$

is called the hat matrix. This means that $\widehat{\mathbb{Y}}$ is the projection of $\mathbb{Y}$ onto the column space of $\mathbb{X}$.

There are several other quantities of interest in linear regression:

1. The in-sample prediction error:

$$\frac{1}{n}\sum_i (\widehat{Y}_i - Y_i)^2 = \frac{1}{n}||\widehat{\mathbb{Y}} - \mathbb{Y}||^2.$$

2. The out-of-sample prediction error:

$$\mathbb{E}[(\widehat{Y} - Y)^2]$$

   where (X, Y) is a new point and the expectation is over both the randomness in $\widehat{\beta}$ and in the new sample.

3. The ℓ_2 error $\mathbb{E}[\|\widehat{\beta} - \beta\|_2^2]$.

4. The support recovery error (makes most sense when β^* is sparse):

$$\mathbb{P}(\text{supp}(\widehat{\beta}) \neq \text{supp}(\beta^*)).$$

Let us review these quantities for low-dimensional regression.

In-sample prediction error. Note that

$$\frac{1}{n}||\widehat{\mathbb{Y}} - \mathbb{Y}||^2 = \frac{1}{n}||\mathbb{X}\widehat{\beta} - \mathbb{X}\beta - \epsilon||^2 = \frac{1}{n}||\mathbb{X}\widehat{\beta} - \mathbb{X}\beta||^2 + \frac{1}{n}||\epsilon||^2 - 2\frac{1}{n}\langle \epsilon, \mathbb{X}(\widehat{\beta} - \beta)\rangle.$$

The second term is the unavoidable error: our estimate $\widehat{\beta}$ has no effect on it. The last term has mean 0 and concentrates to 0 quickly. So people focus on the first term. Recall that, conditional on $\mathcal{X}$,

$$\widehat{\beta}\ N(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$$

so that

$$\mathbb{X}\widehat{\beta}\ N(\mathbb{X}\beta, \sigma^2 H).$$

that is

$$\Delta \equiv \mathbb{X}\widehat{\beta} - \mathbb{X}\beta \sim N(0, \sigma^2 H).$$

Now we use the following fact: if $W \sim N(\mu, \Sigma)$ then $\mathbb{E}[W^T A W] = \mathrm{trace}(A\Sigma) + \mu^T A \mu$ where tr denotes the trace (sum of the diagonal elements). Hence,

$$\begin{aligned}\mathbb{E}[\Delta^T\Delta] = \sigma^2\mathrm{tr}(H) = \sigma^2\mathrm{tr}(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X})&\\ &= \sigma^2\mathrm{tr}(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)\\ \sigma^2\mathrm{tr}(\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}) = \sigma^2\mathrm{tr}(I) = \sigma^2 d.&\end{aligned}$$

Finally,

$$\mathbb{E}\left[\frac{\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2}{n}\ \middle|\ \mathcal{X}\right] = \frac{\sigma^2 d}{n}.$$

It follows that

$$\mathbb{E}\left[\frac{\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2}{n}\right] = \frac{\sigma^2 d}{n}.$$

ℓ_2 **error**. Again, under our assumptions we know that, conditional on $\mathcal{X}$,

$$\widehat{\beta} \sim N(\beta^*, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}).$$

Thus

$$\mathbb{E}\Big[||\widehat{\beta} - \beta^*||^2\ \ \Big|\ \ \mathcal{X}\Big] = \sigma^2\mathrm{tr}(\mathbb{X}^T\mathbb{X})^{-1})$$

and so

$$\mathbb{E}\Big[||\widehat{\beta} - \beta^*||^2\ \Big] = \sigma^2\mathbb{E}[\mathrm{tr}(\mathbb{X}^T\mathbb{X})^{-1}] = \frac{\sigma^2}{n}\mathbb{E}[(\widehat{\Sigma})^{-1}]$$

where we recall that $\widehat{\Sigma} = n^{-1}\mathbb{X}^T\mathbb{X}$. Suppose that $\widehat{\Sigma}$ has eigenvalues bounded from below by $c > 0$. Then $\widehat{\Sigma}$ has eigenvalues bounded from above by $1/c$. Recall that the trace is equal to the sum of the eigenvalues. Therefore,

$$\mathbb{E}\Big[||\widehat{\beta} - \beta^*||^2\Big] \leq \frac{\sigma^2 d}{cn}.$$

2 High-dimensional Regression

In high-dimensional regression, we are interested in the setting where the covariate distribution has dimension $d \gg n$. The first thing to observe is that even if our old analysis worked (it does not) the prediction error and ℓ_2 error both scale as $\sigma^2 d/n$ which does not go to 0 as we increase the sample-size, which would mean that our methods are inconsistent. From a minimax perspective, it turns out that this is unavoidable, i.e. it is impossible to consistently estimate the regression vector β, when $d \gg n$, and we need to make structural assumptions to make progress.

Also, the least-squares estimator is no longer well-defined. To see this, observe that the assumption that $\widehat{\Sigma}$ is invertible (which is completely benign in low-dimensions) can never hold in high-dimensions. In particular the matrix,

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T,$$

has rank at most n (it is a sum of rank 1 matrices) and is a $(d \times d)$ matrix, so is clearly not invertible if $d > n$. The way to picture this is that in high-dimensions there will be many vectors β such that, $Y = \mathbb{X}\beta$ which have least squares error of 0 (i.e. exactly pass through all the samples).

This is a form of over-fitting, and one way to avoid this is to use regularization. This is roughly equivalent to imposing some type of structure on the unknown β and then attempting to recover β by leveraging this structure. We will focus on versions of sparsity, i.e. settings where β is either exactly sparse (i.e. has s non-zero entries) or is approximately sparse (i.e. has bounded ℓ_1 norm).

Analogous to the Gaussian sequence model there are two estimators that one might consider:

1. **Hard-Thresholding type estimator:** The analog of hard thresholding is:

   $$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \frac{t^2}{2}\sum_{i=1}^{d} \mathbb{I}(\beta_i \neq 0).$$

   This is usually called best-subset regression. The best way to think about the nomenclature is to consider a closely related estimator:

   $$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2,$$

   $$\text{subject to } \sum_{i=1}^{d} \mathbb{I}(\beta_i \neq 0) \leq k,$$

   where now we have a different tuning parameter $k > 0$ (instead of t). You should be able to (with some effort) convince yourself of the fact that these two programs are exactly equivalent, i.e. if you fix any $t > 0$ and solve the first program, then there is some k for which you obtain exactly the same solution. The first form is sometimes called the penalized-form and the second is called the constrained-form.

The natural way to implement the second estimator would be to enumerate all subsets of size k, fit a regression on this subset and then pick the subset, and estimate β that has lowest mean -squared error. Hence the name, "best-subset regression". But this is computationally infeasible.

2. **Soft-Thresholding type estimator:** The analog of soft thresholding is known as the LASSO, i.e. the Least Absolute Selection and Shrinkage Operator,

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + t\sum_{i=1}^{d} |\beta_i|.$$

Analogous to the above, one can consider a closely related estimator:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2,$$

$$\text{subject to } \sum_{i=1}^{d} |\beta_i| \leq k,$$

again there is an equivalence, i.e. every value of t corresponds to some value of k. This program is a convex program, and simple methods (roughly, gradient descent with tweaks) can be used to solve it quite fast. There is typically no closed-form solution but that is not a huge problem.

This brings us to an important distinction between the Gaussian sequence model and regression. In the Gaussian sequence model (no X) both of these programs had simple closed-form solutions, whereas now this is no longer the case. More importantly, best-subset is computationally intractable but the LASSO is not.

With this motivation in place, let us study the prediction error of the LASSO. We begin with some assumptions, for simplicity we will study the constrained form of the LASSO, and further we will just assume that the tuning parameter k is chosen to be exactly $\|\beta^*\|_1$ where β^* is the true value of β. In practice, one might choose this tuning parameter by cross-validation or some other method.

To simplify our calculations we will also assume the design matrix X is column-normalized, i.e. for each column j of the matrix:

$$\sum_{i=1}^{n} \mathbb{X}_{ij}^2 = n.$$

You can ensure this by re-normalizing every column of X. This does change β^* (and its ℓ_1 norm).

Theorem 1 *Suppose we consider the constrained-LASSO with* $k = \|\beta^*\|_1$ *where* β^* *denotes the true value. Then, with probability at least* $1 - \delta$*:*

$$\frac{1}{n}\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq 4\sigma\|\beta^*\|_1\sqrt{\frac{2\log(2d/\delta)}{n}}.$$

This bound is exactly analogous to the bound on the error of the hard/soft-thresholding estimator in the Gaussian sequence model when we assumed that the ℓ_1 norm of the mean vector θ^* was bounded. Notice again, that the prediction error goes to 0 with n, even in settings where $d \gg n$.

This result is due to Greenshtein and Ritov and really kicked off the wave of high-dimensional statistics. It showed that high-dimensional prediction was possible (at least in the linear model). Several later works showed that under stronger assumptions, one could achieve small ℓ_2 error and even exactly identify the non-zero components of β^* (i.e. do feature selection) in the high-dimensional setting. Furthermore, most of these phenomena generalize to general parametric models (for instance, high-dimensional logistic regression, high-dimensional graphical model estimation and so on).

Proof: To prove this we note that, since we selected the tuning parameter to be equal to $\|\beta^*\|_1$, the vector β^* is feasible for the program and $\widehat{\beta}$ is optimal, so we have the so-called basic inequality:

$$\frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\widehat{\beta}\|_2^2 \leq \frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\beta^*\|_2^2,$$

where we divided both sides by n for convenience. Re-arranging this inequality we obtain that,

$$\frac{1}{2n}\|\mathbb{E}X\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq \frac{1}{n}\langle\epsilon, \mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\rangle = \langle\frac{\mathbb{X}^T\epsilon}{n}, \widehat{\beta} - \beta^*\rangle,$$

where ϵ is the noise in the linear model. Holder's inequality tells us that for any two vectors $a, b \in \mathbb{R}^d$,

$$\langle a, b\rangle \leq \left(\max_{i=1} a_i\right)\left(\sum_{i=1}^{d}|b_i|\right).$$

Applying this inequality we obtain,

$$\frac{1}{n}\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq 2\|\widehat{\beta} - \beta^*\|_1 \max_{i=1}\frac{X_i^T\epsilon}{n}$$

where X_i denotes the i-th column of the design matrix. Now, by the triangle inequality, $\|\widehat{\beta} - \beta^*\|_1 \leq 2\|\beta^*\|_1$ (recall that we constrained our optimal solution to have ℓ_1 norm at most $\|\beta^*\|_1$), so it only remains to bound $\max_{i=1}^d \frac{X_i^T\epsilon}{n}$.

Each entry here, conditional on $\mathcal{X}$, has a Gaussian distribution with mean 0 and variance $\sigma^2\|X_i\|_2^2/n^2 \leq \sigma^2/n$, using our column normalization assumption. So with probability at least $1-\delta$, we have that,

$$\max_{i=1} \frac{X_i^T\epsilon}{n} \leq \sigma\sqrt{\frac{2\log(2d/\delta)}{n}},$$

and combining these facts we obtain the desired bound.