Linear Regression

We observe $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where $X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. For notational simplicity, we will always assume that $X_i(1) = 1$.

Given a new pair (X, Y) we want to predict Y from X. The *conditional prediction risk* is

$$R(\widehat{m}) = \mathbb{E}[(Y - \widehat{m}(X))^2 | \mathcal{D}] = \int (y - \widehat{m}(x))^2 dP(x, y)$$

and the *prediction risk* of $\widehat{m}$ is

$$r(\widehat{m}) = \mathbb{E}(Y - \widehat{m}(X))^2 = \mathbb{E}[r(\widehat{m})]$$

where the expected value is over all random variables. The true regression function is

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

We have the following bias-variance decomposition:

$$r(\widehat{m}) = \sigma^2 + \int b_n^2(x) dP(x) + \int v_n(x) dP(x)$$

where

$$\sigma^2 = \mathbb{E}[Y - m(X)]^2, \quad b_n(x) = \mathbb{E}[\widehat{m}(x)] - m(x), \quad v_n(x) = \mathrm{Var}(\widehat{m}(x)).$$

Let $\epsilon = Y - m(X)$. Note that

$$\mathbb{E}[\epsilon] = \mathbb{E}[Y - m(X)] = \mathbb{E}[\mathbb{E}[Y - m(X) \,|\, X]] = 0.$$

A *linear predictor* has the form $g(x) = \beta^T x$. The *best linear predictor* minimizes $\mathbb{E}(Y - \beta^T X)^2$. (We do not assume that $m(x)$ is linear.) The minimizer, assuming that Σ is non-singular, is

$$\beta_* = \Sigma^{-1} \alpha$$

where $\Sigma = \mathbb{E}[XX^T]$ and $\alpha = \mathbb{E}(YX)$. **We will use linear predictors; but we should never assume that $m(x)$ is linear.** The *excess risk* is of the linear predictor $\beta^T x$ is

$$r(\beta) - r(\beta_*) = (\beta - \beta_*)^T \Sigma (\beta - \beta_*). \tag{1}$$

The *training error* is

$$\widehat{r}_n(\beta) = \frac{1}{n} \sum_i (Y_i - X_i^T \beta)^2$$

1 Low Dimensional Linear Regression

Recall that $\Sigma = \mathbb{E}[XX^T]$. The *least squares estimator* $\widehat{\beta}$ minimizes the training error $\widehat{r}_n(\beta)$. We then have that

$$\widehat{\beta} = \widehat{\Sigma}^{-1}\widehat{\alpha}$$

where

$$\widehat{\Sigma} = \frac{1}{n}\sum_i X_i X_i^T, \quad \widehat{\alpha} = \frac{1}{n}\sum_i Y_i X_i.$$

We want to show that $r(\widehat{\beta})$ is close to $r(\beta_*)$. For simplicity, we will assume that the distribution P of (Y_i, X_i) supported on a compact set. Also, for simplicity, we assume that $\widehat{\beta}$ is truncated by some large constant L.

Theorem 1 *Let $\mathcal{P}$ be the set of all distributions for $Z = (X, Y)$ supported on a compact set K. There exists constants c_1, c_2 such that the following is true. For any $\epsilon > 0$,*

$$\sup_{P\in\mathcal{P}} P^n\Bigl(r(\widehat{\beta}_n) > r(\beta_*(P)) + 2\epsilon\Bigr) \le c_1 e^{-nc_2\epsilon^2}. \tag{2}$$

Hence,

$$r(\widehat{\beta}_n) - r(\beta_*) = O_P\left(\sqrt{\frac{1}{n}}\right).$$

Proof. Given any β, define $\widetilde{\beta} = (-1, \beta)$ and $\Lambda = \mathbb{E}[ZZ^T]$ where $Z = (Y, X)$. Note that

$$r(\beta) = \mathbb{E}(Y - \beta^T X)^2 = \mathbb{E}[(Z^T\widetilde{\beta})^2] = \widetilde{\beta}^T \Lambda \widetilde{\beta}.$$

Similarly,

$$\widehat{r}_n(\beta) = \widetilde{\beta}^T \widehat{\Lambda}_n \widetilde{\beta}$$

where

$$\widehat{\Lambda}_n = \frac{1}{n}\sum_i Z_i Z_i^T.$$

So

$$|\widehat{r}_n(\beta) - r(\beta)| = |\widetilde{\beta}^T(\widehat{\Lambda}_n - \Lambda)\widetilde{\beta}| \le ||\widetilde{\beta}||_1^2\ \Delta_n$$

where

$$\Delta_n = \max_{j,k} |\widehat{\Lambda}_n(j,k) - \Lambda(j,k)|.$$

By Hoeffding's inequality and the union bound,

$$P\Bigl(\sup_{\beta\in B} |\widehat{r}_n(\beta) - r(\beta)| > \epsilon\Bigr) \le c_1 e^{-nc_2\epsilon^2}.$$

On the event $\sup_{\beta\in B}|\widehat{r}_n(\beta)-r(\beta)|<\epsilon$, we have

$$r(\beta_*)\leq r(\widehat{\beta}_n)\leq \widehat{r}_n(\widehat{\beta}_n)+\epsilon\leq \widehat{r}_n(\beta_*)+\epsilon\leq r(\beta_*)+2\epsilon.$$

□

The above result is not tight. Here is a more refined bound.

Theorem 2 (Theorem 11.3 of Gyorfi, Kohler, Krzyzak and Walk, 2002) *Let* $\sigma^2=\sup_x \mathrm{Var}(Y|X=x)<\infty$. *Assume that all the random variables are bounded by* $L<\infty$. *Then*

$$\mathbb{E}\int|\widehat{\beta}^T x-m(x)|^2dP(x)\leq 8\inf_\beta\int|\beta^T x-m(x)|^2dP(x)+\frac{Cd(\log(n)+1)}{n}.$$

The proof is straightforward but is very long. The strategy is to first bound $n^{-1}\sum_i(\widehat{\beta}^T X_i-m(X_i))^2$ using the properties of least squares. Then, using concentration of measure one can relate $n^{-1}\sum_i f^2(X_i)$ to $\int f^2(x)dP(x)$.

We have the following central limit theorem for $\widehat{\beta}$.

Theorem 3 *We have*

$$\sqrt{n}(\widehat{\beta}-\beta)\rightsquigarrow N(0,\Gamma)$$

where

$$\Gamma=\Sigma^{-1}\mathbb{E}\left[(Y-X^T\beta)^2XX^T\right]\Sigma^{-1}$$

The covariance matrix Γ *can be consistently estimated by*

$$\widehat{\Gamma}=\widehat{\Sigma}^{-1}\widehat{M}\widehat{\Sigma}^{-1}$$

where

$$\widehat{M}(j,k)=\frac{1}{n}\sum_{i=1}^n X_i(j)X_i(k)\widehat{\epsilon}_i^2$$

and $\widehat{\epsilon}_i=Y_i-\widehat{\beta}^T X_i$.

The matrix $\widehat{\Gamma}$ is called the *sandwich* estimator. The Normal approximation can be used to construct confidence intervals for β. For example, $\widehat{\beta}(j)\pm z_\alpha\sqrt{\widehat{\Gamma}(j,j)/n}$ is an asymptotic $1-\alpha$ confidence interval for $\beta(j)$. We can also get confidence intervals by using the bootstrap. Do **not** use the textbook formulas for the standard errors of $\widehat{\beta}$. These assume that the regression function itself is linear. See Buja et al (2015) for details.

2 High Dimensional Linear Regression

Now suppose that $d > n$. We can no longer use least squares. There are many approaches.

The simplest is to preprocess the data to reduce the dimension. For example, we can perform PCA on the $X's$ and use the first k principal components where $k < n$. Alternatively, we can cluster the covariates based on their correlations. We can the use one feature from each cluster or take the average of the covariates within each cluster. Another approach is to screen the variables by choosing the k features with the largest correlation with Y. After dimension reduction, we can the use least squares. These preprocessing methods can be very effective.

A different approach is to use all the covariates but, instead of least squares, we shrink the coefficients towards 0. This is called *ridge regression* and is discussed in the next section.

Yet another approach is model selection where we try to find a good subset of the covariates. Let S be a subset of $\{1, \ldots, d\}$ and let $X_S = (X(j) : \ j \in S)$. If the size of S is not too large, we can regress Y on X_S instead of S.

In particular, fix $k < n$ and let $\mathcal{S}_k$ denote all subsets of size k. For a given $S \in \mathcal{S}_k$, let β_S be the best linear predictor $\beta_S = \Sigma_S^{-1}\alpha_S$ for the subset S. We would like to choose $S \in \mathcal{S}_k$ to minimize

$$\mathbb{E}(Y - \beta_S^T X_S)^2.$$

This is equivalent to:

$$\text{minimize } \mathbb{E}(Y - \beta^T X)^2 \quad \text{subject to } ||\beta||_0 \leq k$$

where $||\beta||_0$ is the number of non-zero elements of β.

There will be a bias-variance tradeoff. As k increases, the bias decreases but the variance increases.

We can approximate the risk with the training error. But the minimization is over all subsets of size k. This minimization is NP-hard. So best subset regression is infeasible. We can approximate best subset regression in two different ways: a greedy approxmation or a convex relaxation. The former leads to forward stepwise regression. The latter leads to the lasso.

All these methods involve a tuning parameter which can be chosen by cross-validation.

3 Ridge Regression

In this case we minimize

$$\frac{1}{n}\sum_i (Y_i - X_i^T\beta)^2 + \lambda||\beta||^2$$

```
Forward Stepwise Regression

1. Input $k$. Let $S = \emptyset$.
2. Let $r_j = n^{-1}\sum_i Y_i X_i(j)$ denote the corrleation between $Y$ and the $j^{\rm th}$ feature.
   Let $J = \mathrm{argmax}_j |r_j|$. Let $S = S \cup \{J\}$.
3. Compute the regression of $Y$ on $X_S = (X(j) :\ j \in S)$. Compute the residuals
   $e = (e_1, \ldots, e_n)$ where $e_i = Y_i - \widehat{\beta}_S^T X_i$.
4. Compute the correlations $r_j$ between the residuals $e$ and the remaining features.
5. Let $J = \mathrm{argmax}_j |r_j|$. Let $S = S \cup \{J\}$.
6. Repeat steps 3-5 until $|S| = k$.
7. Output $S$.
```

Figure 1: Forward Stepwise Regression

where $\lambda \geq 0$. The minimizer is

$$\widehat{\beta} = (\widehat{\Sigma} + \lambda I)^{-1}\widehat{\alpha}.$$

As λ increases, the bias increases and the variance decreases.

Theorem 4 (Hsu, Kakade and Zhang 2014) *Suppose that* $||X_i|| \leq r$. *Let* $\beta^T x$ *be the best linear apprximation to* $m(x)$. *Then, with probability at least* $1 - 4e^{-t}$,

$$r(\widehat{\beta}) - r(\beta) \leq \left(1 + O\left(\frac{1 + \frac{r^2}{\lambda}}{n}\right)\right)\frac{\lambda||\beta||^2}{2} + \frac{\sigma^2}{n}\frac{\mathrm{tr}(\Sigma)}{2\lambda}.$$

Proposition 5 *If* $Y = X^T\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$ *and* $\beta \sim N(0, \tau^2 I)$. *Then the posterior mean is the ridge regression estimator with* $\lambda = \sigma^2/\tau^2$.

4 Forward Stepwise Regression (Greedy Regression)

Forward stepwise regression is a greedy approximation to best subset regression. In what follows, we will assume that the features have been standardized to have sample mean 0 and sample variance $n^{-1}\sum_i X_i^2(j) = 1$. The algorithm is in Fugure 1.

Now we will discuss the theory of forward stepwise regression. Let's start with a functional, noise-free version. We want to greedily approximate a function f using a dictionary of functions $\mathcal{D} = \{\psi_1, \psi_2, \ldots, \}$. The elements of $\mathcal{D}$ are called atoms. Assume that $||\psi|| = 1$ for all $\psi \in \mathcal{D}$. Assume that f and the atoms of the dictionary belong to a Hilbert space $\mathcal{H}$.

```
1. Input: $f$.
2. Initialize: $r_0 = f$, $f_0 = 0$, $V = \emptyset$.
3. Repeat: At step $N$ define
       $g_N = \text{argmax}_{\psi \in \mathcal{D}} |\langle r_{N-1}, \psi \rangle|$
   and set $V_N = V_{N-1} \cup \{g_N\}$. Let $f_N$ be the projection of $r_{N-1}$ onto $\text{Span}(V_N)$.
   Let $r_N = f - f_N$.
```

Figure 2: The Orthogonal Greedy Algorithm.

Let Σ_N denote all linear combinations of elements of $\mathcal{D}$ with at most N terms. Define the best N-term approximation error

$$\sigma_N(f) = \inf_{|\Lambda| \leq N} \inf_{g \in \text{Span}(\Lambda)} \|f - g\| \tag{3}$$

where Λ denotes a subset of $\mathcal{D}$ and $\text{Span}(\Lambda)$ is the set of linear combinations of functions in Λ.

Suppose first that f is in the span of the dictionary. The function may then have more than one expansion of the form $f = \sum_j \beta_j \psi_j$. We define the norm

$$\|f\|_{\mathcal{L}_p} = \inf \|\beta\|_p$$

where the infimum is over all expansions of f. The functional version of stepwise regression, known as the **Orthogonal Greedy Algorithm** (OGA), is also known as Orthogonal Matching Pursuit. The algorithm is given in Figure 2.

The algorithm produces a series of approximations f_N with corresponding residuals r_N. We have the following two theorems from Barron et al (2008), the first dating back to DeVore and Temlyakov (1996).

Theorem 6 *For all $f \in \mathcal{L}_1$, the residual r_N after N steps of OGA satsifies*

$$\|r_N\| \leq \frac{\|f\|_{\mathcal{L}_1}}{\sqrt{N+1}} \tag{4}$$

for all $N \geq 1$.

Proof. Note that f_N is the best approximation to f from $\text{Span}(V_N)$. On the other hand, the best approximation from the set $\{a\, g_N :\ a \in \mathbb{R}\}$ is $\langle f, g_N \rangle g_N$. The error of the former must be smaller than the error of the latter. In other words, $||f - f_N||^2 \leq ||f - f_{N-1} - \langle r_{N-1}, g_N\rangle g_N||^2$. Thus,

$$\begin{aligned} \|r_N\|^2 &\leq \|r_{N-1} - \langle r_{N-1}, g_N\rangle g_N\|^2 \\ &= \|r_{N-1}\|^2 + |\langle r_{N-1}, g_N\rangle|^2 \underbrace{\|g_N\|^2}_{=1} - 2|\langle r_{N-1}, g_N\rangle|^2 \\ &= \|r_{N-1}\|^2 - |\langle r_{N-1}, g_N\rangle|^2. \end{aligned} \tag{5}$$

Now, $f = f_{N-1} + r_{N-1}$ and $\langle f_{N-1}, r_{N-1}\rangle = 0$. So,

$$\begin{aligned} \|r_{N-1}\|^2 &= \langle r_{N-1}, r_{N-1}\rangle = \langle r_{N-1}, f - f_{N-1}\rangle = \langle r_{N-1}, f\rangle - \underbrace{\langle r_{N-1}, f_{N-1}\rangle}_{=0} \\ &= \langle r_{N-1}, f\rangle = \sum_j \beta_j \langle r_{N-1}, \psi_j\rangle \leq \sup_{\psi\in\mathcal{D}} |\langle r_{N-1}, \psi\rangle| \sum_j |\beta_j| \\ &= \sup_{\psi\in\mathcal{D}} |\langle r_{N-1}, \psi\rangle| \, \|f\|_{\mathcal{L}_1} = |\langle r_{N-1}, g_N\rangle| \, \|f\|_{\mathcal{L}_1}. \end{aligned}$$

Continuing from equation (5), we have

$$\begin{aligned} \|r_N\|^2 &\leq \|r_{N-1}\|^2 - |\langle r_{N-1}, g_N\rangle|^2 = \|r_{N-1}\|^2 \left(1 - \frac{\|r_{N-1}\|^2 |\langle r_{N-1}, g_N\rangle|^2}{\|r_{N-1}\|^4}\right) \\ &\leq \|r_{N-1}\|^2 \left(1 - \frac{\|r_{N-1}\|^2 |\langle r_{N-1}, g_N\rangle|^2}{|\langle r_{N-1}, g_N\rangle|^2 \, \|f\|^2_{\mathcal{L}_1}}\right) = \|r_{N-1}\|^2 \left(1 - \frac{\|r_{N-1}\|^2}{\|f\|^2_{\mathcal{L}_1}}\right). \end{aligned}$$

If $a_0 \geq a_1 \geq a_2 \geq \cdots$ are nonnegative numbers such that $a_0 \leq M$ and $a_N \leq a_{N-1}(1 - a_{N-1}/M)$ then it follows from induction that $a_N \leq M/(N+1)$. The result follows by setting $a_N = \|r_N\|^2$ and $M = \|f\|^2_{\mathcal{L}_1}$. □

If f is not in $\mathcal{L}_1$, it is still possible to bound the error as follows.

Theorem 7 *For all $f \in \mathcal{H}$ and $h \in \mathcal{L}_1$,*

$$\|r_N\|^2 \leq \|f - h\|^2 + \frac{4\|h\|^2_{\mathcal{L}_1}}{N}. \tag{6}$$

Proof. Choose any $h \in \mathcal{L}_1$ and write $h = \sum_j \beta_j \psi_j$ where $\|h\|_{\mathcal{L}_1} = \sum_j |\beta_j|$. Write $f = f_{N-1} + f - f_{N-1} = f_{N-1} + r_{N-1}$ and note that r_{N-1} is orthogonal to f_{N-1}. Hence, $\|r_{N-1}\|^2 =$

$\langle r_{N-1}, f\rangle$ and so

$$\begin{aligned} \|r_{N-1}\|^2 &= \langle r_{N-1}, f\rangle = \langle r_{N-1}, h + f - h\rangle = \langle r_{N-1}, h\rangle + \langle r_{N-1}, f - h\rangle \\ &\le \langle r_{N-1}, h\rangle + \|r_{N-1}\|\, \|f - h\| \\ &= \sum_j \beta_j \langle r_{N-1}, \psi_j\rangle + \|r_{N-1}\|\, \|f - h\| \\ &\le \sum_j |\beta_j|\, |\langle r_{N-1}, \psi_j\rangle| + \|r_{N-1}\|\, \|f - h\| \\ &\le \max_j |\langle r_{N-1}, \psi_j\rangle| \sum_j |\beta_j| + \|r_{N-1}\|\, \|f - h\| \\ &= |\langle r_{N-1}, g_k\rangle|\, \|h\|_{\mathcal{L}_1} + \|r_{N-1}\|\, \|f - h\| \\ &\le |\langle r_{N-1}, g_k\rangle|\, \|h\|_{\mathcal{L}_1} + \frac{1}{2}(\|r_{N-1}\|^2 + \|f - h\|^2). \end{aligned}$$

Hence,

$$|\langle r_{N-1}, g_k\rangle|^2 \ge \frac{(\|r_{N-1}\|^2 - \|f - h\|^2)^2}{4\|h\|_{\mathcal{L}_1}^2}.$$

Thus,

$$a_N \le a_{N-1}\left(1 - \frac{a_{N-1}}{4\|h\|_{\mathcal{L}_1}^2}\right)$$

where $a_N = \|r_N\|^2 - \|f - h\|^2$. By induction, the last displayed inequality implies that $a_N \le 4\|h\|_{\mathcal{L}_1}^2/k$ and the result follows. □

Corollary 8 *For each N,*

$$\|r_N\|^2 \le \sigma_N^2 + \frac{4\theta_N^2}{N}$$

where θ_N is the $\mathcal{L}_1$ norm of the best N-atom approximation.

In Figure 3 we re-express forward stepwise regression in a form closer to the notation we have been using. In this version, we have a finite dictionary $\mathcal{D}_n$ and a data vector $Y = (Y_1, \ldots, Y_n)^T$ and we use the empirical norm defined by

$$\|h\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n h^2(X_i)}.$$

We assume that the dictionary is normalized in this empirical norm.

By combining the previous results with concentration of measure arguments (see appendix for details) we get the following result, due to Barron, Cohen, Dahmen and DeVore (2008).

```
1. Input: $Y \in \mathbb{R}^n$.
2. Initialize: $r_0 = Y$, $\widehat{f}_0 = 0$, $V = \emptyset$.
3. Repeat: At step $N$ define
       $g_N = \text{argmax}_{\psi \in \mathcal{D}} |\langle r_{N-1}, \psi \rangle_n|$
   where $\langle a, b \rangle_n = n^{-1} \sum_{i=1}^n a_i b_i$. Set $V_N = V_{N-1} \cup \{g_N\}$. Let $f_N$ be the projection
   of $r_{N-1}$ onto Span($V_N$). Let $r_N = Y - f_N$.
```

Figure 3: The Greedy (Forward Stepwise) Regression Algorithm: Dictionary Version

Theorem 9 *Let* $h_n = \text{argmin}_{h \in \mathcal{F}_N} \|f_0 - h\|^2$. *Suppose that* $\limsup_{n\to\infty} \|h_n\|_{\mathcal{L}_{1,n}} < \infty$. *Let* $N \sim \sqrt{n}$. *Then, for every* $\gamma > 0$, *there exist* $C > 0$ *such that*

$$\|f - \widehat{f}_N\|^2 \leq 4\sigma_N^2 + \frac{C \log n}{n^{1/2}}$$

except on a set of probability $n^{-\gamma}$.

Let us compare this with the lasso which we will discuss next. Let $f_L = \sum_j \beta_j \psi_j$ minimize $\|f - f_L\|^2$ subject to $\|\beta\|_1 \leq L$. Then, we will see that

$$\|f - \widehat{f}_L\|^2 \leq \|f - f_L\|^2 + O_P\left(\frac{\log n}{n}\right)^{1/2}$$

which is the same rate.

The rate $n^{-1/2}$ is in fact optimal. It might be surprising that the rate is independent of the dimension. Why do you think this is the case?

4.1 The Lasso

The lasso approximates best subset regression by using a convex relaxation. In particular, the norm $||\beta||_0$ is replaced with $||\beta||_1 = \sum_j |\beta_j|$.

The lasso estimator $\widehat{\beta}$ is defined as the minimizer of

$$\sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1.$$

This is a convex problem so the estimator can be found efficiently. The estimator is sparse: for large enough λ, many of the components of $\widehat{\beta}$ are 0. This is proved in the course on convex optimization. Now we discuss some theoretical properties of the lasso.[1]

The following result was proved in Zhao and Yu (2006), Meinshausen and Bühlmann (2005) and Wainwright (2006). The version we state is from Wainwright (2006). Let $\beta = (\beta_1, \ldots, \beta_s, 0, \ldots, 0)$ and decompose the design matrix as $\mathbb{X} = (\mathbb{X}_S \ \mathbb{X}_{S^c})$ where $S = \{1, \ldots, s\}$. Let $\beta_S = (\beta_1, \ldots, \beta_s)$.

Theorem 10 (Sparsistency) *Suppose that:*

1. *The true model is linear.*
2. *The design matrix satisfies*
$$\|\mathbb{X}_{S^c}\mathbb{X}_S(\mathbb{X}_S^T\mathbb{X}_S)^{-1}\|_\infty \leq 1-\epsilon \quad \text{for some } 0 < \epsilon \leq 1. \tag{7}$$
3. $\phi_n(d_n) > 0$.
4. *The ϵ_i are Normal.*
5. λ_n *satisfies*
$$\frac{n\lambda_n^2}{\log(d_n - s_n)} \to \infty$$
and
$$\frac{1}{\min_{1\leq j\leq s_n}|\beta_j|}\left(\sqrt{\frac{\log s_n}{n}} + \lambda_n \left\| \left(\frac{1}{n}\mathbb{X}^T\mathbb{X}\right)^{-1}\right\|_\infty\right) \to 0. \tag{8}$$

Then the lasso is sparsistent, meaning that $P(\text{support}(\widehat{\beta}) = \text{support}(\beta)) \to 1$ *where* $\text{support}(\beta) = \{j : \ \beta(j) \neq 0$.

The conditions of this theorem are very strong. They are not checkable and they are unlikely to ever be true in practice.

Theorem 11 (Consistency: Meinshausen and Yu 2006) *Assume that*

1. *The true regression function is linear.*
2. *The columns of $\mathbb{X}$ have norm n and the covariates are bounded.*

[1] The norm $||\beta||_1$ can be thought of as a measure of sparsity. For example, the vectors $x = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ and $y = (1, 0, \ldots, 1)$ have the same L_2 norm. But $||y||_1 = 1 < ||x||_1 = \sqrt{d}$.

3. $\mathbb{E}(\exp|\epsilon_i|) < \infty$ *and* $\mathbb{E}(\epsilon_i^2) = \sigma^2 < \infty$.

4. $\mathbb{E}(Y_i^2) \le \sigma_y^2 < \infty$.

5. $0 < \phi_n(k_n) \le \Phi_n(k_n) < \infty$ *for* $k_n = \min\{n, d_n\}$.

6. $\liminf_{n\to\infty} \phi_n(s_n \log n) > 0$ *where* $s_n = \|\beta_n\|_0$.

Then

$$\|\beta_n - \widehat{\beta}_n\|^2 = O_P\left(\frac{\log n}{n}\frac{s_n \log n}{\phi_n^2(s_n \log n)}\right) + O\left(\frac{1}{\log n}\right) \tag{9}$$

If

$$s_n \log d_n \left(\frac{\log n}{n}\right) \to 0 \tag{10}$$

and

$$\lambda_n = \sqrt{\frac{\sigma_y^2 \Phi_n(\min n, d_n) n^2}{s_n \log n}} \tag{11}$$

then $\|\widehat{\beta}_n - \beta_n\|^2 \xrightarrow{P} 0$.

Once again, the conditions of this theorem are very strong. They are not checkable and they are unlikely to ever be true in practice.

The next theorem is the most important one. It does not require unrealistic conditions. We state the theorem for bounded covariates. A more general version appears in Greenshtein and Ritov (2004).

Theorem 12 *Let* $Z = (Y, X)$. *Assume that* $|Y| \le B$ *and* $\max_j |X(j)| \le B$. *Let*

$$\beta_* = \underset{||\beta||_1 \le L}{\operatorname{argmin}}\, r(\beta)$$

where $r(\beta) = \mathbb{E}(Y - \beta^T X)^2$. *Thus,* $x^T\beta_*$ *is the best, sparse linear predictor (in the* L_1 *sense). Let* $\widehat{\beta}$ *be the lasso estimator:*

$$\widehat{\beta} = \underset{||\beta||_1 \le L}{\operatorname{argmin}}\, \widehat{r}(\beta)$$

where $\widehat{r}(\beta) = n^{-1}\sum_{i=1}^n (Y_i - X_i^T\beta)^2$. *With probabilty at least* $1 - \delta$,

$$r(\widehat{\beta}) \le r(\beta_*) + \sqrt{\frac{16(L+1)^4 B^2}{n} \log\left(\frac{\sqrt{2}\, d}{\sqrt{\delta}}\right)}.$$

Proof. Let $Z = (Y, X)$ and $Z_i = (Y_i, X_i)$. Define $\gamma \equiv \gamma(\beta) = (-1, \beta)$. Then

$$r(\beta) = \mathbb{E}(Y - \beta^T X)^2 = \gamma^T \Lambda \gamma$$

where $\Lambda = \mathbb{E}[ZZ^T]$. Note that $||\gamma||_1 = ||\beta||_1 + 1$. Let $\mathcal{B} = \{\beta : \ ||\beta||_1 \leq L\}$. The training error is

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

where $\widehat{\Lambda} = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T$. For any $\beta \in \mathcal{B}$,

$$\begin{aligned}
|\widehat{r}(\beta) - r(\beta)| &= |\gamma^T(\widehat{\Lambda} - \Lambda)\gamma| \\
&\leq \sum_{j,k} |\gamma(j)|\,|\gamma(k)|\ |\widehat{\Lambda}(j,k) - \Lambda(j,k)| \leq ||\gamma||_1^2 \delta_n \\
&\leq (L+1)^2 \Delta_n
\end{aligned}$$

where

$$\Delta_n = \max_{j,k} |\widehat{\Lambda}(j,k) - \Lambda(j,k)|.$$

So,

$$r(\widehat{\beta}) \leq \widehat{r}(\widehat{\beta}) + (L+1)^2\Delta_n \leq \widehat{r}(\beta_*) + (L+1)^2\Delta_n \leq r(\beta_*) + 2(L+1)^2\Delta_n.$$

Note that $|Z(j)Z(k)| \leq B^2 < \infty$. By Hoeffding's inequality,

$$\mathbb{P}(\Delta_n(j,k) \geq \epsilon) \leq 2e^{-n\epsilon^2/(2B^2)}$$

and so, by the union bound,

$$\mathbb{P}(\Delta_n \geq \epsilon) \leq 2d^2 e^{-n\epsilon^2/(2B^2)} = \delta$$

if we choose $\epsilon = \sqrt{(4B^2/n)\log\left(\frac{\sqrt{2}\,d}{\sqrt{\delta}}\right)}$. Hence,

$$r(\widehat{\beta}) \leq r(\beta_*) + \sqrt{\frac{16(L+1)^4 B^2}{n}\log\left(\frac{\sqrt{2}\,d}{\sqrt{\delta}}\right)}.$$

with probability at least $1 - \delta$. □

Problems With Sparsity. Sparse estimators are convenient and popular but they can some problems. Say that $\widehat{\beta}$ is **weakly sparsistent** if, for every β,

$$P_\beta\big(I(\widehat{\beta}_j = 1) \leq I(\beta_j = 1) \ \text{ for all } j\big) \to 1 \tag{12}$$

as $n \to \infty$. In particular, if $\widehat{\beta}_n$ is sparsistent, then it is weakly sparsistent. Suppose that d is fixed. Then the least squares estimator $\widehat{\beta}_n$ is minimax and satisfies

$$\sup_\beta E_\beta(n\|\widehat{\beta}_n - \beta\|^2) = O(1). \tag{13}$$

But sparsistent estimators have much larger risk:

Theorem 13 (Leeb and Pötscher (2007)) *Suppose that the following condiitons hold:*

1. *d is fixed.*
2. *The covariariates are nonstochastic and $n^{-1}\mathbb{X}^T\mathbb{X} \to Q$ for some positive definite matrix Q.*
3. *The errors ϵ_i are independent with mean 0, finite variance σ^2 and have a density f satisfying*
$$0 < \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx < \infty.$$

If $\widehat{\beta}$ is weakly sparsistent then

$$\sup_\beta E_\beta(n\|\widehat{\beta}_n - \beta\|^2) \to \infty. \tag{14}$$

More generally, if ℓ is any nonnegative loss function then

$$\sup_\beta E_\beta(\ell(n^{1/2}(\widehat{\beta}_n - \beta))) \to \sup_s \ell(s). \tag{15}$$

Proof. Choose any $s \in \mathbb{R}^d$ and let $\beta_n = -s/\sqrt{n}$. Then,

$$\begin{aligned}\sup_\beta E_\beta(\ell(n^{1/2}(\widehat{\beta} - \beta)) &\geq E_{\beta_n}(\ell(n^{1/2}(\widehat{\beta} - \beta)) \geq E_{\beta_n}(\ell(n^{1/2}(\widehat{\beta} - \beta))I(\widehat{\beta} = 0)) \\ &= \ell(-\sqrt{n}\beta_n)P_{\beta_n}(\widehat{\beta} = 0) = \ell(s)P_{\beta_n}(\widehat{\beta} = 0).\end{aligned}$$

Now, $P_0(\widehat{\beta} = 0) \to 1$ by assumption. It can be shown that we also have $P_{\beta_n}(\widehat{\beta} = 0) \to 1$.[2] Hence, with probability tending to 1,

$$\sup_\beta E_\beta(\ell(n^{1/2}(\widehat{\beta} - \beta)) \geq \ell(s).$$

Since s was arbitrary the result follows. □

It follows that, if R_n denotes the minimax risk then

$$\sup_\beta \frac{R(\widehat{\beta}_n)}{R_n} \to \infty.$$

The implication is that when d is much smaller than n, sparse estimators have poor behavior. However, when d_n is increasing and $d_n > n$, the least squares estimator no longer satisfies (13). Thus we can no longer say that some other estimator outperforms the sparse estimator. In summary, sparse estimators are well-suited for high-dimensional problems but not for low dimensional problems.

[2] This follows from a property called contiguity.

5 Inference?

Is it possible to do inference after model selection? Do we need to? I'll discuss this in class.

References

Buja, Berk, Brown, George, Pitkin, Traskin, Zhao and Zhang (2015). Models as Apprximations — A Conspiracy of Random Regressors and Model Deviations Against Classical Inference in Regression. *Statistical Science.*

Hsu, Kakade and Zhang (2014). Random design analysis and ridge regression. arXiv:1106.2363.

Gyorfi, Kohler, Krzyzak and Walk. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer.

Appendix: L_2 Boosting

Define estimators $\widehat{m}_n^{(0)}, \ldots, \widehat{m}_n^{(k)}, \ldots,$ as follows. Let $\widehat{m}^{(0)}(x) = 0$ and then iterate the following steps:

1. Compute the residuals $U_i = Y_i - \widehat{m}^{(k)}(X_i)$.
2. Regress the residuals on the Y_i's: $\widehat{\beta}_j = \sum_i U_i X_{ij} / \sum_i X_{ij}^2$, $j = 1, \ldots, d$.
3. Find $J = \text{argmin}_j RSS_j$ where $RSS_j = \sum_i (U_i - \widehat{\beta}_J X_{iJ})^2$.
4. Set $\widehat{m}^{(k+1)}(x) = \widehat{m}^{(k)}(x) + \widehat{\beta}_J x_J$.

The version above is called L_2 **boosting** or **matching pursuit**. A variation is to set $\widehat{m}^{(k+1)}(x) = \widehat{m}^{(k)}(x) + \nu \widehat{\beta}_J x_J$ where $0 < \nu \leq 1$. Another variation is to set $\widehat{m}^{(k+1)}(x) = \widehat{m}^{(k)}(x) + \nu \text{sign}(\widehat{\beta}_J) x_J$ which is called **forward stagewise regression**. Yet another variation is to set $\widehat{m}^{(k)}$ to be the linear regression estimator based on all variables selected up to that point. This is **forward stepwise regression** or **orthogonal matching pursuit**.

Theorem 14 *The matching pursuit estimator is linear. In particular,*

$$\widehat{Y}^{(k)} = B_k Y \tag{16}$$

where $\widehat{Y}^{(k)} = (\widehat{m}^{(k)}(X_1), \ldots, \widehat{m}^{(k)}(X_n))^T$,

$$B_k = I - (I - H_k)(I - H_{k-1}) \cdots (I - H_1), \tag{17}$$

and

$$H_j = \frac{\mathbb{X}_j \mathbb{X}_j^T}{\|\mathbb{X}_j\|^2}. \tag{18}$$

Theorem 15 (Bühlmann 2005) *Let* $m_n(x) = \sum_{j=1}^{d_n} \beta_{j,n} x_j$ *be the best linear approximation based on* d_n *terms. Suppose that:*

(A1 Growth) $d_n \leq C_0 e^{C_1 n^{1-\xi}}$ *for some* $C_0, C_1 > 0$ *and some* $0 < \xi \leq 1$.

(A2 Sparsity) $\sup_n \sum_{j=1}^{d_n} |\beta_{j,n}| < \infty$.

(A3 Bounded Covariates) $\sup_n \max_{1 \leq j \leq d_n} \max_i |X_{ij}| < \infty$ *with probability 1.*

(A4 Moments) $\mathbb{E}|\epsilon|^s < \infty$ *for some* $s > 4/\xi$.

Then there exists $k_n \to \infty$ *such that*

$$\mathbb{E}_X |\widehat{m}_n(X) - m_n(x)|^2 \to 0 \tag{19}$$

as $n \to 0$.

We won't prove the theorem but we will outline the idea. Let $\mathcal{H}$ be a Hilbert space with inner product $\langle f, g \rangle = \int f(x) g(x) dP(x)$. Let $\mathcal{D}$ be a dictionary, that is a set of functions, each of unit norm, that span $\mathcal{H}$. Define a functional version of matching pursuit, known as the **weak greedy algorithm**, as follows. Let $R_0(f) = f$, $F_0 = 0$. At step k, find $g_k \in \mathcal{D}$ so that

$$|\langle R_{k-1}(f), g_k \rangle| \geq t_k \sup_{h \in \mathcal{D}} |\langle R_{k-1}(f), h \rangle|$$

for some $0 < t_k \leq 1$. In the weak greedy algorithm we take $F_k = F_{k-1} + \langle f, g_k \rangle g_k$. In the weak orthogonal greedy algorithm we take F_k to be the projection of $R_{k-1}(f)$ onto $\{g_1, \ldots, g_k\}$. Finally set $R_k(f) = f - F_k$.

Theorem 16 (Temlyakov 2000) *Let* $f(x) = \sum_j \beta_j g_j(x)$ *where* $g_j \in \mathcal{D}$ *and* $\sum_{j=1}^{\infty} |\beta_j| \leq B < \infty$. *Then, for the weak orthogonal greedy algorithm*

$$\|R_k(f)\| \leq \frac{B}{\left(1 + \sum_{j=1}^k t_j^2\right)^{1/2}} \tag{20}$$

and for the weak greedy algorithm

$$\|R_k(f)\| \leq \frac{B}{\left(1 + \sum_{j=1}^k t_j^2\right)^{t_k/(2(2+t_k))}}. \tag{21}$$

L_2 boosting essentially replaces $\langle f, X_j\rangle$ with $\langle Y, X_j\rangle_n = n^{-1}\sum_i Y_i X_{ij}$. Now $\langle Y, X_j\rangle_n$ has mean $\langle f, X_j\rangle$. The main burden of the proof is to show that $\langle Y, X_j\rangle_n$ is close to $\langle f, X_j\rangle$ with high probability and then apply Temlyakov's result. For this we use Bernstein's inequality. Recall that if $|Z_j|$ are bounded by M and Z_j has variance σ^2 then

$$\mathbb{P}(|\overline{Z} - \mathbb{E}(Z_j)| > \epsilon) \leq 2\exp\left\{-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + M\epsilon/3}\right\}. \tag{22}$$

Hence, the probability that any empirical inner products differ from their functional counterparts is no more than

$$d_n^2 \exp\left\{-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + M\epsilon/3}\right\} \to 0 \tag{23}$$

because of the growth condition.

Appendix: Proof of Theorem 9

The $\mathcal{L}_1$ norm depends on n and so we denote this by $\|h\|_{\mathcal{L}_{1,n}}$. For technical reasons, we assume that $\|f\|_\infty \leq B$, that $\widehat{f}_n$ is truncated to be no more than B and that $\|\psi\|_\infty \leq B$ for all $\psi \in \mathcal{D}_n$.

Theorem 17 *Suppose that $p_n \equiv |\mathcal{D}|_n \leq n^c$ for some $c \geq 0$. Let $\widehat{f}_N$ be the output of the stepwise regression algorithm after N steps. Let $f(x) = \mathbb{E}(Y|X = x)$ denote the true regression function. Then, for every $h \in \mathcal{D}_n$,*

$$\mathbb{P}\left(\|f - \widehat{f}_N\|^2 > 4\|f - h\|^2 + \frac{8\|h\|^2_{\mathcal{L}_{1,n}}}{N} + \frac{CN\log n}{n}\right) < \frac{1}{n^\gamma}$$

for some positive constants γ and C.

Before proving this theorem, we need some preliminary results. For any $\Lambda \subset \mathcal{D}$, let $S_\Lambda = \mathrm{Span}(\Lambda)$. Define

$$\mathcal{F}_N = \bigcup \left\{ S_\Lambda : \ |\Lambda| \leq N \right\}.$$

Recall that, if $\mathcal{F}$ is a set of functions then $N_p(\epsilon, \mathcal{F}, \nu)$ is the L_p covering entropy with respect to the probability measure ν and $N_p(\epsilon, \mathcal{F})$ is the supremum of $N_p(\epsilon, \mathcal{F}, \nu)$ over all probability measures ν.

Lemma 18 *For every $t > 0$, and every $\Lambda \subset \mathcal{D}_n$,*

$$N_1(t, S_\Lambda) \leq 3\left(\frac{2eB}{t}\log\left(\frac{3eB}{t}\right)\right)^{|\Lambda|+1}, \qquad N_2(t, S_\Lambda) \leq 3\left(\frac{2eB^2}{t^2}\log\left(\frac{3eB^2}{t^2}\right)\right)^{|\Lambda|+1}.$$

Also,

$$N_1(t,\mathcal{F}_N) \le 12p^N \left(\frac{2eB}{t}\log\left(\frac{3eB}{t}\right)\right)^{N+1}, \qquad N_2(t,\mathcal{F}_N) \le 12p^N \left(\frac{2eB^2}{t^2}\log\left(\frac{3eB^2}{t^2}\right)\right)^{N+1}.$$

Proof. The first two equation follow from standard covering arguments. The second two equations follow from the fact that the number of subsets of Λ of size at most N is

$$\sum_{j=1}^{N}\left(\begin{array}{c} p \\ j \end{array}\right) \le \sum_{j=1}^{N}\left(\frac{ep}{j}\right)^j \le N\left(\frac{ep}{N}\right)^N \le p^N \max_N N\left(\frac{p}{N}\right)^N \le 4p^N.$$

□

The following lemma is from Chapter 11 of Gyorfi et al. The proof is long and technical and we omit it.

Lemma 19 *Suppose that $|Y| \le B$, where $B \ge 1$, and $\mathcal{F}$ is a set of real-valued functions such that $\|f\|_\infty \le B$ for all $f \in \mathcal{F}$. Let $f_0(x) = \mathbb{E}(Y|X=x)$ and $\|g\|^2 = \int g^2(x)dP(x)$. Then, for every $\alpha, \beta > 0$ and $\epsilon \in (0, 1/2]$,*

$$\mathbb{P}\Bigg((1-\epsilon)\|f-f_0\|^2 \ge \|Y-f\|_n^2 - \|Y-f_0\|_n^2 + \epsilon(\alpha+\beta) \quad \text{for some } f \in \mathcal{F}\Bigg)$$
$$\le 14 N_1\left(\frac{\beta\epsilon}{20B}, \mathcal{F}\right)\exp\left\{-\frac{\epsilon^2(1-\epsilon)\alpha n}{214(1+\epsilon)B^4}\right\}.$$

Proof of Theorem 17. For any $h \in \mathcal{F}_n$ we have

$$\begin{aligned}\|\widehat{f}-f_0\|_n^2 &= \underbrace{\|\widehat{f}-f_0\|^2 - 2\Bigg(\|Y-\widehat{f}\|_n^2 - \|Y-f_0\|_n^2\Bigg)}_{A_1} \\ &\quad + \underbrace{2\Bigg(\|Y-\widehat{f}\|_n^2 - \|Y-h\|_n^2\Bigg)}_{A_2} + \underbrace{2\Bigg(\|Y-h\|_n^2 - \|Y-f_0\|_n^2\Bigg)}_{A_3}.\end{aligned}$$

Apply Lemma 19 with $\epsilon = 1/2$ together with Lemma 18 to conclude that, for $C_0 > 0$ large enough,

$$\mathbb{P}\left(A_1 > \frac{C_0 N \log n}{n} \quad \text{for some } f\right) < \frac{1}{n^\gamma}.$$

To bound A_2, apply Theorem 7 with norm $\|\cdot\|_n$ and with Y replacing f. Then,

$$\|Y-\widehat{f}\|_n^2 \le \|Y-h\|_n^2 + \frac{4\|h\|_{1,n}^2}{k}$$

and hence $A_2 \leq \frac{8\|h\|_{1,n}^2}{k}$. Next, we have that

$$\mathbb{E}(A_3) = \|f_0 - h\|^2$$

and for large enough C_1,

$$\mathbb{P}\left(A_3 > \|f_0 - h\|^2 + \frac{C_1 N \log n}{n} \quad \text{for some } f\right) < \frac{1}{n^\gamma}.$$

□