36-705: Intermediate Statistics **Fall 2017**

Lecture 30: November 15

Lecturer: Siva Balakrishnan

30.1 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 turn to structural assumptions to make progress.

A perhaps even more alarming aspect of high-dimensional regression is that 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, $y = X\beta$ 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 again 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:

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

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

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2,$$
$$\text{subject to } \sum_{i=1}^{d} \mathbb{I}(\beta_i \neq 0) \leq k,$$

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".

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

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

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

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2,$$
$$\text{subject to } \sum_{i=1}^{d} |\beta_i| \leq k,$$

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 $\|\beta^*\|_1$. 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:

$$\sum_{i=1}^{n} X_{ij}^2 \leq n.$$

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

Theorem 30.1 *Suppose we consider the constrained-LASSO with $k = \|\beta^*\|_1$, then the prediction error of our estimator, with probability at least $1-\delta$, satisfies:*

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

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

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

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

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

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

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

Applying this inequality we obtain,

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

where X_i denotes the i-th column of the design. 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, has a Gaussian distribution with mean 0 and variance $\sigma^2 \|X_i\|_2^2 / n \leq \sigma^2/n$, using our column normalization assumption. So with probability at least $1 - \delta$, we have that,

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

and combining these facts we obtain the desired bound.

30.2 Bayesian Inference

We have already talked about the mechanics of Bayesian inference when we discussed constructing point estimates by treating the parameters as random with a prior, and the computing and summarizing the posterior. What we really did not talk about yet, and will spend a small amount of time talking about now is the philosophy of Bayesian inference.

The philosophical distinction between Bayes and frequentists is deep. We have so far followed the frequentist framework, where, to us a probability is representing some type of long run frequency, i.e. when we say the probability that our estimator is close to some unknown "true" parameter with probability at least $1 - \delta$ we are really imagining repeating this (or some other) experiment many many times and then our guarantees will be correct for at least $1 - \delta$ of these experiments. Similarly, with confidence intervals, we imagine many people across the world construct confidence intervals and our guarantee is that 95% of those intervals would trap the true parameter, i.e. **the goal of frequentist inference is to create procedures with long run guarantees.**

Moreover, the guarantees should be uniform over θ if possible. For example, a confidence interval traps the true value of θ with probability $1 - \alpha$, no matter what the true value of θ is. **In frequentist inference, procedures are random while parameters are fixed, unknown quantities.**

In the *Bayesian approach*, probability is regarded as a measure of **subjective degree of belief**. One can view the Bayesian approach as a way to manipulate beliefs. Beliefs are then assumed to follow the rules of normal probabilities by a notion called *coherence*. In this framework, everything, including parameters, is regarded as random. These procedures do not have to satisfy frequency guarantees.

Here is a table from Larry:

A summary of the main ideas is in Table 30.1.

	Bayesian	Frequentist
Probability	subjective degree of belief	limiting frequency
Goal	analyze beliefs	create procedures with frequency guarantees
θ	random variable	fixed
X	random variable	random variable
Use Bayes' theorem?	Yes. To update beliefs.	Yes, if it leads to procedure with good frequentist behavior. Otherwise no.

Table 30.1: Bayesian versus Frequentist Inference

30.3 The mechanics of Bayesian Inference

Roughly, the setup in Bayesian Inference is exactly the same as in frequentist inference: we begin by specifying a statistical model, i.e. a collection of distributions $\{P_\theta : \theta \in \Theta\}$.

The main distinction is that we now treat the parameter θ as random, and encode our "prior beliefs" about the value of the parameter in a distribution π.

We assume that the observed data, is from the *conditional distribution*, conditional on some realization of the random parameter, i.e. the setup is:

$$\theta \sim \pi$$
$$\{X_1, \ldots, X_n\} | \theta \sim P_\theta.$$

We do not observe θ but can compute our "posterior belief" using Bayes' rule, i.e.:

$$\pi(\theta | X_1, \ldots, X_n) = \frac{\mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)}{\int_\theta \mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)},$$

i.e. while the frequentist treats the likelihood as just a function of θ, the Bayesian (weights and) normalizes the likelihood and interprets it as a distribution over Θ.

We have seen examples of this whole thing before, except rather than treat the posterior as an object of interest, we used it to obtain point estimates.

30.4 The goals of Bayesian inference

In frequentist inference the goal was somehow part of the definition: create procedures that have good frequency properties.

In Bayesian inference the goal was not very clearly articulated, i.e. when is a Bayesian analysis considered successful. There are two camps here:

1. **Pure Bayesian viewpoint:** Once you write down a prior that captures your prior belief and compute the posterior, you are essentially done, i.e. you have succeeded.
2. **The frequentist viewpoint:** You are successful if you are successful in the frequentist viewpoint, i.e. treat the parameter as fixed, and define success as your posterior concentrating around the true parameter (i.e. some equivalent of consistency) and confidence intervals computed using the posterior have frequentist coverage guarantees.

30.5 Bayesian confidence sets and Frequentist guarantees

Once we have a posterior distribution, we can construct what are called credible sets: they are the Bayesian analogue of confidence sets but are quite different.

A $1-\alpha$ credible set/interval is simply any set C_α to which the posterior assigns $1-\alpha$ mass, i.e.

$$\int_{C_\alpha} \pi(\theta|X_1,\ldots,X_n)d\theta = 1-\alpha.$$

Once again notice that the thing that is random is θ, the data is conditioned on (i.e. fixed). We could write this as:

$$\mathbb{P}_{\theta\sim\pi(\theta|X_1,\ldots,X_n)}(\theta \in C_\alpha|X_1,\ldots,X_n) = 1-\alpha.$$

The set C_α is fixed (i.e. not random) here, unlike in a frequentist confidence interval. These intervals do not typically have frequency guarantees, and we will see examples of this.

If one is interested in the frequency properties of Bayesian inference then one might also be interested in some notion of frequentist consistency and rates of convergence. The typical way to formulate frequentist consistency is via something called *posterior contraction*, i.e. in the frequentist setup (where θ^* is fixed, unknown) we want that our posterior concentrates around the true value of the parameter (consistency) and does so quickly (rates of convergence).

Formally, consistency says that for any fixed $\epsilon > 0$,

$$\pi(\{\theta : \|\theta - \theta^*\| \geq \epsilon\}|X_1,\ldots,X_n) \to 0,$$

when $X_1, \ldots, X_n \sim P_{\theta^*}$. We would also say that the rate of convergence is ϵ_n if for some $\delta_n \to 0$ we have that,

$$\pi(\{\theta : \|\theta - \theta^*\| \leq \epsilon_n\} | X_1, \ldots, X_n) \geq 1 - \delta_n,$$

as $n \to \infty$ (again, $X_1, \ldots, X_n \sim P_{\theta^*}$).

30.6 Bernstein-von Mises theorem

At a high-level the Bernstein-von Mises theorem guarantees us that in fixed-dimensional problems, under the assumption that the prior is continuous, and (strictly) positive in a neighborhood around θ^*, the posterior is close to a Gaussian, i.e.

$$\left\| \pi(\theta | X_1, \ldots, X_n) - N\left(\widehat{\theta}_n, \frac{I(\widehat{\theta}_n)^{-1}}{n}\right) \right\|_{\text{TV}} \to 0,$$

where $\widehat{\theta}_n$ is the MLE and the distance between the two distributions is the total-variation distance, i.e. for two distributions with densities p, q:

$$\|p - q\|_{\text{TV}} = \frac{1}{2} \int |p(x) - q(x)| dx.$$

We might discuss this further at some point but for now you should take away that the posterior is very close to a Gaussian centered at the MLE with rapidly shrinking variance.

One immediate consequence of the BvM result is that credible intervals will be roughly identical to the usual Wald interval (based on the MLE) as $n \to \infty$. The key take away: in fixed dimension, large sample-size problems, under some conditions Bayesian procedures will have strong frequency guarantees.

The condition that "(strictly) positive in a neighborhood around θ^*" is extremely strong in high-dimensions or in a non-parametric problem. In general, when the parameter space is large you should be suspicious of this assumption.

30.7 Where do priors come from?

The "correct" answer is that the prior should truly be an encoding of your prior beliefs. Often we choose priors by convenience (recall our examples from the minimax lecture). Some might argue that in many cases the priors do not matter (see below) but this is only rigorously true in low-dimensional, parametric problems.

Some might also choose priors based on the data, this is known as *empirical Bayes.* This is often a good idea but some would consider it to not strictly adhere to the Bayesian philosophy.

Some have argued for what are called non-informative priors, i.e. priors that somehow capture complete ignorance about the parameter. The natural first attempt would be to say that we take $\pi(\theta) \propto 1$, however this has some drawbacks. If we have no information about the parameter θ then presumably we should also have no information about some transformation about the parameter, i.e. say θ^2. However, if you transform the flat prior to a prior on θ^2 it will not be flat. A prior that is in fact invariant under transformations is called Jeffreys prior where we choose $\pi(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information for the model under consideration. You will verify this in your HW.

There are also so-called hierarchical priors, i.e. we can parameterize the prior, treat these parameters as random variables and then place priors on those parameters as well. There is some folklore intuition that results are less sensitive to the parameters of the higher-level priors but this is somewhat difficult to make precise.

Finally, perhaps all of this is missing the point? There is a sense in which many pragmatic Bayesians believe that the prior is not the important piece of Bayesian inference, it is the complicated averaging that happens in Bayesian inference. Again this is difficult to make precise. There are frequentist versions of the model averaging idea (look up exponentially weighted aggregation or mirror averaging) that lead to aggregated models with great properties.

30.8 Priors = Regularizers?

A slightly different viewpoint that is often articulated is that one can view priors as regularizers.

Most regularized frequentist estimators, for instance estimating Bernoulli probabilities with "Laplace smoothing" (i.e. adding psuedo-counts) is just the posterior mean with a Beta prior. The LASSO regression estimator or Ridge regression estimator are just the posterior mode with either a Laplace prior or a Gaussian prior (respectively). Relatedly, one can derive model complexity regularizers with appropriate model complexity dependent priors.

The argument is that "many sensible frequentists procedures (ones with strong guarantees) are just posterior summaries with particular priors."

This argument should be taken with a grain of salt: lets focus on the LASSO, which is the posterior mode with a Laplace prior. As we have discussed previously the LASSO has some very desirable properties (high-dimensional prediction consistency) and so one might wonder does the (full) posterior have nice properties in a high-dimensional setting?

The answer turns out to be no: *once you leave the realm of the Bernstein-von Mises theorem (fixed d, growing n) things can break down.* In particular, in the high-dimensional regression problem the posterior itself does not meaningfully concentrate and sampling from the posterior will lead to completely meaningless inference (from a frequentist point of view). Essentially only the posterior mode has nice properties, the rest of the posterior is useless.