36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 1: January 14

Lecturer: Siva Balakrishnan

More so than in previous iterations of this class, we will assume you have taken 36-705 (ideally from me). If at any point you feel like the class is moving too quickly, or we should cover more background material then please feel free to let me know (by email or in class).

We will begin today by discussing Chapter 1 of the Wainwright book.

1.1 Classical Versus High-Dimensional Analyses

In 36-705 we spent a lot of time on studying classical asymptotics of popular estimators (in particular, the MLE). By classical asymptotics, we mean that we made some regularity assumptions, assumed that the target of inference $\theta^* \in \mathbb{R}^d$ was fixed, and allowed the sample-size $n \to \infty$. In this setting, we argued about the consistency, asymptotic normality and rates of convergence of our estimators to θ^*.

In this class, we will use high-dimensional problems of various kinds as an excuse to introduce several tools for non-asymptotic analysis. In a high-dimensional analysis, we typically either:

- Consider a completely non-asymptotic viewpoint producing guarantees for fixed (n, d) that hold with high-probability.
- Consider a high-dimensional asymptotic viewpoint where we still allow $n \to \infty$, but now we consider settings where d also grows and some aspect ratio remains fixed. For instance, we might suppose that $d/n \to \alpha < 1$, or that $\log(d)/n \to 0$.

High-dimensional asymptotics might seem strange at first sight (Why are we doing asymptotics at all? Why is problem size growing with n?).

From a motivation standpoint, there are multiple reasons to study high-dimensional problems:

1. Many modern datasets are relatively high-dimensional (i.e. typically d will be comparable or larger than n) and so we might expect our classical intuitions to breakdown. Consequently, we need to understand/develop the tools and techniques to guide us toward better estimators (if/when they exist).

This of course still does not quite answer the question why would we want to do high-dimensional asymptotics? The answer is similar to why we ever do asymptotics – unlike

in finite-sample analysis, we will often be able to make very precise theoretical predictions through asymptotics (think of a CLT instead of a tail bound).

The focus of our class will be on learning deeply tools and techniques of non-asymptotic statistics. These tools and techniques are useful in a surprisingly broad range of problems, beyond the examples we will use to illustrate them.

1.2 Some Examples

Our next goal for this lecture will be to, somewhat superficially, discuss some vignettes with a high-dimensional flavor, in part to understand can go wrong and how we might hope to fix it. We will focus on three examples: a vector example, a matrix example and a non-parametric function example. This will in some approximate sense mimic the structure of the class where we will dig deeper into each of these types of estimation problems.

High-dimensional Classification: Suppose we have the following setup, we observe samples:

$$\begin{aligned} \text{(Class 1)} \quad & x_1, \ldots, x_{n_1} \sim N(\mu_1, \Sigma), \\ \text{(Class 2)} \quad & x_{n_1+1}, \ldots, x_{n_1+n_2} \sim N(\mu_2, \Sigma), \end{aligned}$$

where μ_1, μ_2, Σ are unknown and our goal is to build a classifier to distinguish these classes. This is the setup of Fisher's Linear/Quadratic Discriminant Analysis.

If you knew (μ_1, μ_2, Σ) (and suppose that the two classes were equally likely) then the optimal classifier builds on the likelihood ratio test, and labels a new sample x 1 if:

$$(x - \mu_1)^T \Sigma^{-1} (x - \mu_1) \leq (x - \mu_2)^T \Sigma^{-1} (x - \mu_2),$$

i.e. the sample is closer to μ_1 than μ_2 in the appropriate Mahalanobis sense. Alternatively, we classify a sample as 1 if:

$$\langle x - (\mu_1 + \mu_2)/2, \, \Sigma^{-1}(\mu_1 - \mu_2) \rangle \geq 0.$$

Given a sample, Fisher's LDA/QDA is a plug-in method, where we estimate:

$$\begin{aligned} \widehat{\mu}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} x_i \\ \widehat{\mu}_2 &= \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} x_i \\ \widehat{\Sigma} &= \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \widehat{\mu}_1)(x_i - \widehat{\mu}_1)^T + \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \widehat{\mu}_2)(x_i - \widehat{\mu}_2)^T, \end{aligned}$$

and then use the plug-in rule of declaring the label as 1 if:

$$\phi(x) := \langle x - (\widehat{\mu}_1 + \widehat{\mu}_2)/2, \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_2)\rangle \geq 0.$$

Implicitly, we have assumed that $n_1 + n_2 \geq d$ so that the covariance matrix constructed is actually invertible. A natural question is how good is this classifier?

Let us first try to understand how we might evaluate the classifier, and what the best we might hope for is. We can evaluate the classifier by the probability that it will mis-label a future example. In our case, sticking with the assumption that the classes are equally balanced we can define the mis-classification error of our classifier to be:

$$\Psi = \frac{1}{2}\left[P_{N(\mu_1,\Sigma)}(\phi(x) < 0) + P_{N(\mu_2,\Sigma)}(\phi(x) \geq 0)\right].$$

Using symmetry the best possible classifier makes an error with probability,

$$\begin{aligned}
\Psi &= P_{N(\mu_1,\Sigma)}(\langle x, \Sigma^{-1}(\mu_1 - \mu_2\rangle < \langle (\mu_1 + \mu_2)/2, \Sigma^{-1}(\mu_1 - \mu_2)\rangle) \\
&= P_{Z\sim N(0,I_d)}(\langle Z, \Sigma^{-1/2}(\mu_1 - \mu_2)\rangle < \frac{1}{2}(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)) \\
&= \Phi(-\sqrt{(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)}/2).
\end{aligned}$$

So in some sense, the best we can hope for is that we match the performance of this best classifier.

To simplify matters further, we might assume that $\Sigma = I_d$, and in this case a lengthy but straightforward calculation shows that if $d/n \to \alpha$ then the error of our estimated classifier,

$$\widetilde{\Psi} = P_{N(\mu_1,I_d)}(\langle x, \mu_1 - \mu_2\rangle < \langle (\mu_1 + \mu_2)/2, \mu_1 - \mu_2\rangle) \to \Phi(-\gamma^2/2\sqrt{\gamma^2 + 2\alpha}),$$

in probability, where $\gamma^2 = (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)$. On the other hand, in the low-dimensional setting when $d/n \to 0$, we obtain that $\widetilde{\Psi} \to \Psi$.

This illustrates one of the basic things that happens in high-dimensional calculations. In low-dimensions, we often are able to estimate unstructured {vectors, matrices, functions} consistently, and as a result many asymptotic calculations simplify. On the other hand, in high-dimensional calculations we need to be careful about this. We may not be able to consistently estimate unstructured {vectors, matrices, functions}, and we need to appropriately adjust our theoretical predictions/expectations.

Covariance Matrix Estimation: Suppose now that we observe $X_1, \ldots, X_n$ each in $\mathbb{R}^d$, which are drawn i.i.d. from some 0 mean distribution, with an unknown covariance Σ. We would like to estimate the covariance matrix and a natural (unbiased) estimate is the sample covariance matrix.

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

We might ask how close is $\widehat{\Sigma}$ to Σ. There are various matrix norms in which we might choose to measure this. For instance, we might hope that the operator norm,

$$\Delta = \|\widehat{\Sigma} - \Sigma\|_{\text{op}} := \sup_{u:\|u\|_2=1} u^T(\widehat{\Sigma} - \Sigma)u,$$

decays to 0 as $n \to \infty$ (with high-probability).

In the low-dimensional setup (where d is held fixed, and $n \to \infty$) it is not hard to see that the usual law of large numbers implies this fact, i.e. that the sample-covariance matrix is an operator-norm consistent estimator of the true covariance. This is relatively strong notion of consistency and in turn implies for instance that PCA is consistent (we'll be clearer about what we mean by this but roughly the principal subspaces of $\widehat{\Sigma}$ and Σ will be close to each other). In this low-dimensional setting the spectrum of the difference Δ is asymptotically just a point mass at 0.

On the other hand, in the high-dimensional setting this is no longer the case. In particular, suppose that $X_1, \ldots, X_n \sim N(0, \Sigma)$, and suppose that $d/n \to \alpha \in (0, 1)$. It turns out that in this case, the spectrum of Δ is non-trivial asymptotically. In particular, the largest and smallest eigenvalues concentrate around:

$$(1 - \sqrt{\alpha})^2 - 1 \quad \text{and} \quad (1 + \sqrt{\alpha})^2 - 1,$$

respectively. A lot more is known – in particular the density of the eigenvalues of $\widehat{\Sigma}$ asymptotically follows a distribution called the Marcenko-Pastur law. In this setting, when α does not tend to 0, PCA can be inconsistent (and if d/n diverges it can be very "strongly" inconsistent).

One might wonder if matrix estimation is really that different from vector estimation? In particular, the covariance matrix is after all an $O(d^2)$ dimensional vector. As is often the case, the norm in which we choose to estimate the matrix, and the notion of structure we will choose to impose, will be much more natural when the estimand is viewed as a matrix (as opposed to a long vector).

This is just the beginning of a large area of random matrix theory – which has also already studied in high-dimensional asymptotics in the 1950s (to put this in context this is roughly as old as the field of Statistics...). We will cover various non-asymptotic analogues of the Marcenko-Pastur law, discuss various matrix concentration inequalities and discuss implications for the estimation of low-rank, and other similarly structured matrices.

Non-parametric Function Estimation: This example we have covered to some extent in 705, so we will be somewhat brief. In non-parametric regression, our goal given paired samples $\{(X_1, y_1), \ldots, (X_n, y_n)\}$ is to estimate the conditional expectation

$$f(x) = \mathbb{E}[y|X = x].$$

Suppose for simplicity that our Xs belong to $[0, 1]^d$. In order to estimate this function well, we might hope that for each query point $x \in [0, 1]^d$ we have a training sample somewhere

close by (we might only need this for a typical x as opposed to every x, or only need this in some average sense but we will ignore this). Concretely, fix some small $\delta > 0$. For any $x_0 \in [0,1]^d$ we would like to ensure that there is a sample $x \in \{X_1, \ldots, X_n\}$ such that,

$$\|x_0 - x\|_\infty \leq \delta.$$

A straightforward computation shows that to ensure this condition, we need:

$$n \gtrsim \left(\frac{1}{\delta}\right)^d,$$

which grows exponentially with the dimension d.

1.3 What can help us?

The short answer is (often hidden) low-dimensional structure. This will come in different forms:

1. Vectors might be sparse, have small ℓ_p-norm, have entries which exhibit an appropriate type of decay, have entries which are shape-constrained (increasing, decreasing, convex, ...).
2. Matrices might be low-rank, sparse, have non-zero entries only near the diagonal, ...
3. Functions might be smooth, multivariate functions might be additive (i.e. $f(x) \approx \sum_{j=1}^d f_j(x)$), the data might be concentrated near a lower dimensional manifold (so that the function is easier to estimate near this lower-dimensional manifold), ...

Our goal is to develop some general tools to study structured estimation problems. Towards this, we would like to identify interesting/useful notions of structure, understand how this structure might help in estimation and then design computationally-efficient estimators to exploit the structure.