IEOR 265 – Lecture 1
Convex Geometry

1 Classifying Regression Methods

Suppose we have a *system* in which an input (also known as predictors) $x \in \mathbb{R}^k$ gives an output $y \in \mathbb{R}$, and suppose there is a static relationship between x and y that is given by $y = f(x) + \epsilon$, where ϵ is zero mean noise with finite variance (i.e., $\mathbb{E}(\epsilon) = 0$ and var$(\epsilon) = \sigma^2 < \infty$). We will also assume that ϵ is independent of x; otherwise our model has what is known as *endogeneity*, which is a common topic of study in econometrics.

The process of modeling involves using measured data to identify the relationship between x and y, meaning identify $\mathbb{E}[y|x] = f(x)$. This is a huge topic of inquiry, but in this course we will categorize this *regression* problem into three classes:

1. Parametric Regression – The unknown function $f(x)$ is characterized by a finite number of parameters. It is common to think of $f(x;\beta)$, where $\beta \in \mathbb{R}^p$ is a vector of unknown parameters. The simplest example is a linear model, in which we have

   $$f(x;\beta) = \sum_{j=1}^{p} \beta_j x^j.$$

   This approach is used when there is strong *a priori* knowledge about the structure of the system (e.g., physics, biology, etc.).

2. Nonparametric Regression – The unknown function $f(x)$ is characterized by an infinite number of parameters. For instance, we might want to represent $f(x)$ as an infinite polynomial expansion

   $$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots.$$

   This approach is used when there is little *a priori* knowledge about the structure of the system. Though it might seem that this approach is superior because it is more flexible than parametric regression, it turns out that one must pay a statistical penalty because of the need to estimate a greater number of parameters.

3. Semiparametric Regression – The unknown function $f(x)$ is characterized by a component with a finite number of parameters and another component with an infinite number of parameters. In some cases, the infinite number of parameters are known as nuisance parameters; however, in other cases this infinite component might have useful information in and of itself. A classic example is a partially linear model:

   $$f(x) = \sum_{j=1}^{m} \beta_j x^j + g(x^{m+1}, \dots, x^k).$$

Here, the $g(x^{m+1}, \ldots, x^k)$ is represented non-parametrically, and the $\sum_{j=1}^m \beta_j x^j$ term is the parametric component.

This categorization is quite crude because in some problems the classes can blend into each other. For instance, high-dimensional parametric regression can be thought of as nonparametric regression. The key problem in regression is that of *regularization*. The idea of regularization is to improve the statistical properties of estimates by imposing additional structure onto the model.

2 Ordinary Least Squares

Suppose that we have pairs of independent measurements (x_i, y_i) for $i = 1, \ldots, n$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, and that the system is described by a linear model

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i = x_i' \beta + \epsilon_i.$$

Ordinary least squares (OLS) is a method to estimate the unknown parameters $\beta \in \mathbb{R}^p$ given our n measurements. Because the y_i are noisy measurements (whereas the x_i are not noisy in this model), the intuitive idea is to choose an estimate $\hat{\beta} \in \mathbb{R}^p$ which minimizes the difference between the measured y_i and the estimated $\hat{y}_i = x_i' \hat{\beta}$.

There are a number of ways that we could characterize this difference. For mathematical and computational reasons, a popular choice is the *squared loss*: This difference is quantified as $\sum_i (y_i - \hat{y}_i)^2$, and the resulting problem of choosing $\hat{\beta}$ to minimize this difference can be cast as the following (unconstrained) optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i' \beta)^2.$$

For notational convenience, we will define a matrix $X \in \mathbb{R}^{n \times p}$ and a vector $Y \in \mathbb{R}^n$ such that the i-th row of X is x_i' and the i-th row of Y is y_i. With this notation, the OLS problem can be written as

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2,$$

where $\|\cdot\|_2$ is the usual L^2-norm. (Recall that for a vector $v \in \mathbb{R}^k$ the L^2-norm is $\|v\|_2 = \sqrt{(v^1)^2 + \ldots + (v^k)^2}$.)

Now given this notation, we can solve the above defined optimization problem. Because the problem is unconstrained, setting the gradient of the objective to zero and solving the resulting algebraic equation will give the solution. For notational convenience, we will use the function

$J(X, Y; \beta)$ to refer to the objective of the above optimization problem. Computing its gradient gives

$$\nabla_\beta J = 2X'(Y - X\hat{\beta}) = 0 \Rightarrow X'X\hat{\beta} = X'Y$$
$$\Rightarrow \hat{\beta} = (X'X)^{-1}(X'Y).$$

This is the OLS estimate of β for the linear model. In some cases, the solution is written as $\hat{\beta} = (\frac{1}{n}X'X)^{-1}(\frac{1}{n}X'Y)$. The reason for this will be discussed in future lectures.

2.1 Geometric Interpretation of OLS

Recall the optimization formulation of OLS,

$$\hat{\beta} = \arg\min_\beta \|Y - X\beta\|_2^2,$$

where the variables are as defined before. The basic tension in the problem above is that in general no exact solution exists to the linear equation

$$Y = X\beta;$$

otherwise we could use linear algebra to compute β, and this value would be a minimizer to the optimization problem written above.

Though no exact solution exists to $Y = X\beta$, an interesting question to ask is whether there is some related linear equation for which an exact solution exists. Because the noise is in Y and not X, we can imagine that we would like to pick some $\hat{Y}$ such that $\hat{Y} = X\hat{\beta}$ has an exact solution. Recall from linear algebra, that this is equivalent to asking that $\hat{Y} \in \mathcal{R}(X)$ (i.e., $\hat{Y}$ is in the range space of X). Now if we think of $\hat{Y}$ as true signal, then we can decompose Y as

$$Y = \hat{Y} + \Delta Y,$$

where ΔY represents orthogonal noise. Because – from Fredholm's theorem in linear algebra – we know that the range space of X is orthogonal to the null space of X' (i.e., $\mathcal{R}(X) \perp \mathcal{N}(X')$), it must be the case that $\Delta Y \in \mathcal{N}(X')$ since we defined $\hat{Y}$ such that $\hat{Y} \in \mathcal{R}(X)$. As a result, premultiplying $Y = \hat{Y} + \Delta Y$ by X' gives

$$X'Y = X'\hat{Y} + X'\Delta Y = X'\hat{Y}.$$

The intuition is that premultiplying by X' removes the noise component. And because $\hat{Y} \in \mathcal{R}(X)$ and $\hat{Y} = X\hat{\beta}$, we must have that

$$X'Y = X'\hat{Y} = X'X\hat{\beta}.$$

Solving this gives $\hat{\beta} = (X'X)^{-1}(X'Y)$, which is our regular equation for the OLS estimate.

2.2 Challenges with High Dimensionality

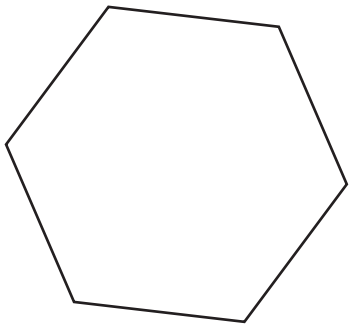
In modern data sets, it is common for p (number of predictors/parameters) to be the same order of magnitude (or even larger) than n (number of measurements). This is highly problematic because the statistical error of OLS is linearly increasing in p, and inversely proportional to n. (More formally, the mean squared error of OLS scales as $O_p(p/n)$; this notation will be explained later in the course.) Consequently, OLS is not statistically well-behaved in this modern high-dimensional setting. To be able to develop estimators that work in this setting, we are required to first better understand the geometery of high-dimensional convex bodies.

3 High-Dimensional Convex Bodies

Recall that $K \subseteq \mathbb{R}^p$ is a convex set if for all $x, y \in K$ and $\lambda \in [0, 1]$ it holds that

$$\lambda \cdot x + (1 - \lambda) \cdot y \in K.$$

In low dimensions, the volume of convex bodies is distributed evenly. A representative example is the two-dimensional polytope show below.



The situation is markedly different in high dimensions. Consider a convex body K, which is a convex set in $\mathbb{R}^p$ that is (i) closed, (ii) bounded, and (iii) has a nonempty interior. Furthermore, suppose K is isotropic, meaning that if a random variable X is uniformly distributed on K then it has the properties that

$$\begin{aligned} \mathbb{E}(X) &= 0 \\ \mathbb{E}(XX') &= \mathbb{I}, \end{aligned}$$

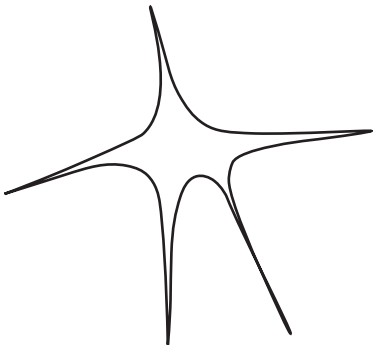
where $\mathbb{I}$ is the $p \times p$ identity matrix. It turns out that majority of the volume of K is concentrated about the ball with radius $\sqrt{p}$. More formally, for every $t \geq 1$ we have

$$\mathbb{P}(\|X\|_2 > t\sqrt{p}) \leq \exp(-ct\sqrt{p}).$$

Moreover, the majority of volume of K is found in a thin shell around the ball of radius $\sqrt{p}$: For every $\epsilon \in (0,1)$, we have

$$\mathbb{P}(|\|X\|_2 - \sqrt{p}| > \epsilon\sqrt{p}) \leq C\exp(-c\epsilon^3 p^{1/2}).$$

Note that in these two above results, C, c are positive absolute constants. As a result, a more intuitive picture of a convex body in high dimensions is



This "hyperbolic" picture is more intuitive because it shows that the convex body can be characterized as consisting of a bulk and outliers, where (i) the bulk of volume is concentrated in a ball of small radius, (ii) the outliers have large radius, and (iiii) the volume of the outliers is exponentially decreasing away from the bulk.

4 More Reading

The material in the last section follows that of

R. Vershynin (2014) *Estimation in high dimensions: a geometric perspective*, in Sampling Theory, a Renaissance. Springer. To appear.

More details about high-dimensional convex bodies can be found in this book chapter and its accompanying references.