IEOR 290A – Lecture 5
Classical M-Estimators

1 Maximum Likelihood Estimators

A maximum likelihood estimator (MLE) is a method for parametric estimation that defines the estimate to be the set of parameters that is "most likely" given the data observed. This statement can be made more precise. If we define $f(x_i, y_i; \beta)$ to be the probability density function (pdf), then the MLE estimate is given by

$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} f(x_i, y_i; \beta),$$

since the likelihood for observing (x_i, y_i) that are i.i.d. is given by $\prod_i f(x_i, y_i; \beta)$. (Recall that the pdf for iid random variables is the product of the individual pdf's.) Because the exponential family of distributions is so ubiquitous, it is common to define the MLE estimate by taking the logarithm of the above objective function. The MLE estimate defined using the log-likelihood is

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \log f(x_i, y_i; \beta).$$

1.1 Example: Population Parameters of Normal Distribution

It is interesting to consider a few examples of MLE estimators. Suppose that $x_i \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown. Then the MLE estimator for μ and σ^2 are given by

$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu, \sigma^2} \sum_{i=1}^{n} \Big(-(x_i - \mu)^2/(2\sigma^2) - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi) \Big),$$

since the normal pdf is $f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x_i - \mu)^2/(2\sigma^2))$. Setting the gradient equal to zero:

$$\sum_{i=1}^{n} 2(x_i - \hat{\mu})/(2\hat{\sigma}^2) = 0$$
$$\sum_{i=1}^{n} \Big((x_i - \hat{\mu})^2/(2(\hat{\sigma}^2)^2) - 1/(2\hat{\sigma}^2) \Big) = 0.$$

The first equation of the gradient can be manipulated as follows:

$$\sum_{i=1}^{n} 2(x_i - \hat{\mu})/(2\sigma^2) = 0 \Rightarrow \sum_{i=1}^{n} \Big(x_i\Big) - n\hat{\mu} = 0$$
$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This is just the normal sample average. Next we manipulate the second equation of the gradient:

$$\sum_{i=1}^{n} \Big((x_i - \hat{\mu})^2/(2(\hat{\sigma}^2)^2) - 1/(2\hat{\sigma}^2)\Big) = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Interestingly, this estimate is biased but consistent.

1.2 Example: Linear Function of Normal Distribution

As another example, suppose that $x_i \sim \mathcal{N}(0, \Sigma)$ and $y_i = x_i'\beta + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with unknown σ^2. The MLE estimate of β is given by

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \Big(-(y_i - x_i'\beta)^2/(2\sigma^2) - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi)\Big).$$

Simplifying this, we have that

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} -(y_i - x_i'\beta)^2 = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - x_i'\beta)^2 = \arg\min_{\beta} \|Y - X\beta\|_2^2.$$

This is equivalent to the OLS estimate!

2 M-Estimator

An M-estimator (the M stands for maximum likelihood-type) is an estimate that is defined as the minimizer to an optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(x_i, y_i; \beta),$$

where ρ is some suitably chosen function. We have already seen examples of M-estimators. Another example is a nonlinear least squares estimator

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - f(x_i; \beta))$$

for some parametric nonlinear model $y_i = f(x_i; \beta) + \epsilon$. There are in fact many different aspects of M-estimators, including:

- MLE estimators;
- robust estimators;
- estimates with Bayesian priors.

One of the strengths of M-estimators is that these various components can be mixed and matched. We have already discussed MLE estimators, and so we will next discuss robust estimators and Bayesian priors.

2.1 Robust Estimators

One of the weaknesses of a squared loss function $L(u) = u^2$ is that it can overly emphasize outliers. As a result, other loss functions can be used. For instance the absolute loss $L(u) = |u|$ could be used. Because it is not differentiable at the origin, one could instead use the Huber loss function

$$L_\delta(u) = \begin{cases} u^2/2, & \text{if } |u| < \delta \\ \delta(|u| - \delta/2), & \text{otherwise} \end{cases}$$

which is quadratic for small values and linear for large values. As a result, it is differentiable at all values. A final example is Tukey's biweight loss function

$$L_\delta(u) = \begin{cases} u^2/2, & \text{if } |u| < \delta \\ \delta^2/2, & \text{otherwise} \end{cases}$$

which is quadratic for small values and constant for large values. One "weakness" of this loss function is that it is not convex, which complicates the optimization problem since finding the minimizer is more difficult for non-convex optimization problems.

2.2 Bayesian Priors

Consider the problem of regression for a linear model $y_i = x_i'\beta + \epsilon_i$, and suppose that the errors ϵ_i are Gaussian with zero mean and finite variance. Now imagine that we have some Bayesian prior density for the model parameters β. There are two important cases:

- Suppose that the β_i have a Bayesian prior given by a normal distribution with zero mean and finite variance. Then the corresponding estimate is given by

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

- Suppose that the β_i have a Bayesian prior given by a Laplace distribution, which has pdf

$$f(\beta_i) = \frac{\lambda}{2}\exp(-\lambda|\beta_i|).$$

  Then the corresponding estimate is given by

$$\hat{\beta} = \arg\min_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$