IEOR 265 – Lecture 5
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

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

for some parametric nonlinear model $y_i=f(x_i;\beta)+\epsilon$. There are in fact many different cases 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

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

3 Ridge Regression

The M-estimator which had the Bayesian interpretation of a linear model with Gaussian prior on the coefficients

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

has multiple names: ridge regression, $L2$-regularization, and Tikhonov regularization.

3.1 Computation of Ridge Regression Estimate

Computation becomes straightforward if we rewrite the objective. Observe that

$$\begin{aligned}\|Y - X\beta\|_2^2 + \|\beta\|_2^2 &= \|Y - X\beta\|_2^2 + \|0 - \sqrt{\lambda}\mathbb{I}_p\beta\|_2^2 \\ &= \left\| \begin{bmatrix} Y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda}\mathbb{I}_p \end{bmatrix} \beta \right\|_2^2.\end{aligned}$$

Thus, the objective in the optimization used to compute the ridge regression estimate is the same as the objective in OLS, but with "pseudo-measurements" corresponding to $\tilde{X} = \sqrt{\lambda}\mathbb{I}_p$ and $\tilde{Y} = 0$ added. From the solution to OLS, we have that the ridge regression estimate is

$$\begin{aligned}\hat{\beta} &= \left(\begin{bmatrix} X \\ \sqrt{\lambda}\mathbb{I}_p \end{bmatrix}' \begin{bmatrix} X \\ \sqrt{\lambda}\mathbb{I}_p \end{bmatrix} \right)^{-1} \begin{bmatrix} X \\ \sqrt{\lambda}\mathbb{I}_p \end{bmatrix}' \begin{bmatrix} Y \\ 0 \end{bmatrix} \\ &= (X'X + \lambda\mathbb{I}_p)^{-1}X'Y.\end{aligned}$$

3.2 Proportional Shrinkage

The ridge regression estimate has an important interpretation in the bias-variance context. Suppose that we compute the singular value decomposition of $X \in \mathbb{R}^{n\times p}$:

$$X = USV'$$

where $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $S \in \mathbb{R}^{n\times p}$ is a rectangular diagonal matrix $S = \begin{bmatrix}\text{diag}(s_1, \ldots, s_p)' & 0\end{bmatrix}'$ where $s_1 \geq \ldots \geq s_p$. Then, the ridge regression estimate can be rewritten as

$$\begin{aligned}\hat{\beta} &= (X'X + \lambda\mathbb{I}_p)^{-1}X'Y \\ &= (VS'U'USV' + \lambda\mathbb{I}_p)^{-1}VS'U'Y \\ &= (V\text{diag}(s_1^2, \ldots, s_p^2)V' + \lambda VV')^{-1}VS'U'Y \\ &= V\text{diag}\left(\frac{1}{s_1^2 + \lambda}, \ldots, \frac{1}{s_p^2 + \lambda}\right)V'VS'U'Y \\ &= V\begin{bmatrix}\text{diag}\left(\frac{s_1}{s_1^2+\lambda}, \ldots, \frac{s_p}{s_p^2+\lambda}\right) & 0\end{bmatrix}U'Y.\end{aligned}$$

If $\lambda = 0$, then the estimate is just the OLS estimate. So one interpretation of the ridge regression estimate is that we are shrinking the inverse of the singular values towards zero. The shrinkage is proportional to the magnitude of s_i, meaning that the shrinkage is relatively smaller for s_i versus s_{i+1}.

4 Collinearity

The usage of SVD suggests a geometric interpretation may be valuable. Consider the standard linear model $y_i = x_i'\beta + \epsilon_i$, and further suppose that $x_i = z_i'B + \mu_i$, where $z_i \in \mathbb{R}^d$ is a vector of

"hidden" variables, $d < p$, $B \in \mathbb{R}^{p \times d}$ is a matrix of coefficients, and μ_i is zero mean noise with finite variance. The idea of this model is that we observe x_i, but x_i has smaller dimensionality due to these variables actually being an unknown function of z_i that are not measured. We will assume that z_i are Gaussian and have zero mean with finite variance. For notational convenience, we define $Z \in \mathbb{R}^{n \times d}$ to be the matrix whose i-th row is z_i'; similarly, we define $M \in \mathbb{R}^{n \times p}$ to be the matrix whose i-th row is μ_i'. Lastly, we define Σ to be the covariance matrix of X, Σ_z to be the covariance matrix of z_i, and $\sigma^2 \mathbb{I}$ to be the covariance matrix of μ_i.

To understand why this situation is problematic, consider the sample covariance matrix

$$\tfrac{1}{n} X'X = \tfrac{1}{n} B'Z'ZB + \tfrac{1}{n} M'M \xrightarrow{p} \Sigma = B'\Sigma_z B + \sigma^2 \mathbb{I}.$$

Now note that Σ_z has rank p, and $B'\Sigma_z B$ is positive semidefinite and so can be diagonalized. Specifically, we can write

$$\begin{aligned} \Sigma &= B'\Sigma_z B + \sigma^2 \mathbb{I} \\ &= U \mathsf{diag}(s_1, \ldots, s_d, 0, \ldots, 0) U' + \sigma^2 \mathbb{I} \\ &= U \mathsf{diag}(s_1 + \sigma^2, \ldots, s_d + \sigma^2, \sigma^2, \ldots, \sigma^2) U'. \end{aligned}$$

This is a problem for two reasons. First, the small σ^2 looks like signal, but it is actually noise. Second, the small σ^2 distorts our signal (though we cannot fix this issue without specifically considering errors-in-variables estimators).

The ridge regression estimate tries to shrink the σ^2 noise terms towards zero, while impacting the signal terms s_i less (i.e., proportional shrinkage). And so ridge regression can be interpreted in this geometrical context as trying to estimate the linear coefficients subject to a model in which the measured variables x_i as actually linear functions of a lower dimensional variable z_i that is not measured.

5 Exterior Derivative Estimator

Consider the collinearity model described above, and suppose that instead we only shrink the values that we believe are noise $s_{d+1}, \ldots, s_p$. Then we can define another estimator as

$$\hat{\beta} = V \left[\mathsf{diag}\left(\tfrac{1}{s_1}, \ldots, \tfrac{1}{s_d}, \tfrac{s_{d+1}}{s_{d+1}^2 + \lambda}, \ldots, \tfrac{s_p}{s_p^2 + \lambda} \right) \quad 0 \right] U'Y.$$

This estimator provides a different bias-variance tradeoff. It turns out that we can define this as the following M-estimator

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

where Π is a projection matrix that projects onto the $(p-d)$ smallest eigenvectors of the sample covariance matrix $\frac{1}{n} X'X$. We call this the exterior derivative estimator (EDE).

The name for this estimator is inspired by the following question: If we estimate the β coefficients in our model, then what is their interpretation? This question looks simple, but it is more complex than it seems at first glance. The coefficients β cannot be interpreted as a gradient because the x_i do not span the whole space. It turns out that the correct interpretation of the β in this model is that of an exterior derivative, which is an extension of gradients to differentiable manifolds. The intuition is that the β only gives derivative information in the directions of the manifold, but we do not get derivative information in other directions. This is important because if we interpret, say ridge regression, in a geometric context then it means that we have only been able to estimate derivative information in some "measured" directions. The EDE estimate makes this intuition clear because we are penalizing for deviations of our estimate from the "measured" directions.

5.1 Principal Component Regression

There is another regression method known as principal component regression (PCR) in which the estimate is

$$\hat{\beta} = V \left[\text{diag}\left(\tfrac{1}{s_1}, \ldots, \tfrac{1}{s_d}, 0, \ldots, 0 \right) \quad 0 \right] U'Y.$$

The normal way for writing this estimate is as a change of coordinates that converts x_i into some scaled variables in a lower dimension $\tilde{z}_i$, builds a linear model with inputs $\tilde{z}_i$ and output y_i, and then performs the inverse coordinate change to get the linear model in the x_i space. We can also interpret PCR as a special case of the EDE. This can be seen by defining the PCR estimate as

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \lim_{\lambda \to \infty} \|Y - X\beta\|_2^2 + \lambda \|\Pi\beta\|_2^2 \\ &= \lim_{\lambda \to \infty} \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\Pi\beta\|_2^2. \end{aligned}$$

Note that swapping the limit and minimization is not allowed in every situation, but it is allowed in this situation. The reasons for this are technical and will be discussed later in the course.