# Model Inference

by

AINDRILA GARAI

MSC STATISTICS, IIT KANPUR

aindrilag22@iitk.ac.in

## Introduction:

Here we provide a general exposition of the maximum likelihood approach and Bayesian method for inference.

### *A Smoothing Example:*

We illustrate the bootstrap in a simple one-dimensional smoothing problem, and show its connection to maximum likelihood. Denote the training data by $Z = z_1, z_2, \ldots, z_N$, with $z_i = (x_i, y_i), i = 1,2, \ldots, N$. So,

$$\mu(x) = \sum_{j=1}^{7} \beta_j \, h_j(x).$$

This is a seven-dimensional linear space of functions and $\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$ obtained by minimizing the squared error over the training set. The corresponding fit is $\hat{\mu}(x) = \sum_{j=1}^{7} \hat{\beta}_j \, h_j(x)$.

If we simulate new responses by adding Gaussian noise to the predicted values: $\hat{\mu}^*(x) \sim N(\hat{\mu}(x), h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}h(x)\hat{\sigma}^2)$.

Now if we use maximum likelihood approach and Z has a normal distribution with mean $\mu$ and variance $\sigma^2$ ,then $\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$ which agrees with the least squares estimate.

## Bayesian Methods:

In the Bayesian approach to inference we specify a sampling model $Pr(Z|\theta)$ and a prior distribution for the parameters $Pr(\theta)$, the posterior distribution

$$\Pr(\theta \mid \mathbf{Z}) = \frac{\Pr(\mathbf{Z} \mid \theta) \cdot \Pr(\theta)}{\int \Pr(\mathbf{Z} \mid \theta) \cdot \Pr(\theta) d\theta}$$

which represents our updated knowledge about $\theta$. The function $\mu(x)$ should be smooth, and have guaranteed this by expressing $\mu$ in a smooth low-dimensional basis of $B$splines.

### Relationship Between the Bootstrap and Bayesian Inference:

Let $z \sim N(\theta, 1)$ and $\theta \mid z \sim N\left(\frac{z}{1+1/\tau}, \frac{1}{1+1/\tau}\right)$. Now the larger we take $\tau$, the more concentrated the posterior becomes around the maximum likelihood estimate $\hat{\theta} = z$. his is the same as a parametric bootstrap distribution in which we generate bootstrap values $z^*$ from the maximum likelihood estimate of the sampling density $N(z, 1)$

### The EM Algorithm:

The EM algorithm is a popular tool for simplifying difficult maximum likelihood problems.

1. Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$.
2. Expectation Step: compute the responsibilities $\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, i = 1,2, \ldots, N$
3. Maximization Step: compute the weighted means and variances:
$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \quad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$
$\hat{\mu}_2 = \frac{\sum_{i-1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \quad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$ and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$
4. Iterate steps 2 and 3 until convergence.

### *MCMC for Sampling from the Posterior:*

One would like to draw samples from the resulting posterior distribution, in order to make inferences about the parameter. **Gibbs sampling** is an MCMC procedure.

1. Take some initial values $\theta^{(0)} = \left(\mu_1^{(0)}, \mu_2^{(0)}\right)$.
2. Repeat for $t = 1,2, \ldots,$.
   (a) For i = 1, 2,... ,N generate $\Delta_i^{(t)} \in \{0,1\}$ with $\Pr\left(\Delta_i^{(t)} = 1\right) = \hat{\gamma}_i\left(\theta^{(t)}\right)$
   (b) Set $\hat{\mu}_1 = \frac{\sum_{i=1}^N \left(1-\Delta_i^{(t)}\right) \cdot y_i}{\sum_{i=1}^N \left(1-\Delta_i^{(t)}\right)}$ and $\hat{\mu}_2 = \frac{\sum_{i=1}^N \Delta_i^{(t)} \cdot y_i}{\sum_{i=1}^N \Delta_i^{(t)}}$ and generate $\mu_1^{(t)} \sim N(\hat{\mu}_1, \hat{\sigma}_1^2)$ and $\mu_2^{(t)} \sim N(\hat{\mu}_2, \hat{\sigma}_2^2)$
3. Continue step 2 until the joint distribution of $\left(\mathbf{\Delta}^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}\right)$ doesn't change.