Lecture 17: Maximum Likelihood Estimation (MLE)

Examples:

1. $X_1, \ldots, X_n$ i.i.d. Bernoulli:

$$\Pr(X_i = 1) = p, \qquad \Pr(X_i = 0) = 1 - p$$

$S_n = X_1 + \ldots + X_n$ — Binomial (n, p)

How to estimate p based on observed $X_1, \ldots, X_n$? Natural estimate is $\hat{p} = \frac{S_n}{n} = \bar{X}_n$. **$\hat{p}$ is actually an MLE.**

2. $X_1, \ldots, X_n$ i.i.d.

$$P(X_i = j) = p_j, \qquad j = 1, \ldots, m \qquad \sum_{j=1}^{m} p_j = 1.$$

$$N_j = \sum_{i=1}^{n} 1_{\{X_i = j\}} \qquad \#j^{th} \text{ class in data} \qquad \sum_{j=1}^{m} N_j = n$$

$(N_1, N_2, \ldots, N_m) \sim$ Multinomial $(n, p_1, \ldots, p_m)$.

Natural estimator for p_j is $\hat{p}_j = \dfrac{N_j}{n}$

$\{\hat{p}_j\}$ **are MLE of** $\{p_j\}$.

$m = 2 \longrightarrow$ Back to Binomial.

$$\begin{aligned} \text{Var}(S_n) &= np(1-p) \\ E(S_n) &= np \end{aligned}$$

because $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$.

In Multinomial notation,

$$N_1 = S_n, \qquad N_0 = n - N_1 = n - S_n$$

which leads to

$$\begin{aligned} \text{Cov}(N_1, N_0) &= \text{Cov}(S_n, n - S_n) \\ &= -\text{Cov}(S_n, S_n) \\ &= -\text{Var}(S_n) \\ &= -np(1-p) \end{aligned}$$

In general?

$$E(N_j) = np_j, \qquad \mathrm{Var}(N_j) = np_j(1-p_j)$$
$$\mathrm{Cov}(N_k, N_j) = -np_kp_j \qquad k \neq j$$

Proof: $E(N_j) = np_j$, $\mathrm{Var}(N_j) = np_j(1-p_j)$ follow from Binomial as $N_j \sim \mathrm{Bin}(n, p_j)$.

$$\begin{aligned}
\mathrm{Cov}(N_k, N_j) &= \mathrm{Cov}\left(\sum_{i=1}^n 1_{\{X_i=k\}}, \sum_{t=1}^n 1_{\{X_t=j\}}\right) \\
&= \sum_{i=1}^n \sum_{t=1}^n \mathrm{Cov}\left(1_{\{X_i=k\}}, 1_{\{X_t=j\}}\right) \qquad k \neq j \\
&= \sum_{i=1}^n \mathrm{Cov}\left(1_{\{X_i=k\}}, 1_{\{X_i=j\}}\right) \\
&= \sum_{i=1}^n \left[E(1_{\{X_i=k\}} \cdot 1_{\{X_i=j\}}) - p_kp_j\right] \\
&= -np_kp_j \qquad k \neq j
\end{aligned}$$

When $n \longrightarrow \infty$, $p \longrightarrow 0$, $np \longrightarrow \lambda > 0$, $\qquad X \sim \mathrm{Bin}(n, p)$

$$P(X = j) \longrightarrow e^{-\lambda}\frac{\lambda^j}{j!} \text{ — Poisson}(\lambda), \qquad j = 0, 1, \ldots$$

3. $X_1, \ldots, X_n$ i.i.d. Poisson (λ)

$$EX_1 = \lambda, \qquad \mathrm{Var}(X_1) = \lambda, \qquad \mathrm{SD}(X_1) = \sqrt{\lambda}$$

How to estimate λ?: $\hat{\lambda} = \frac{1}{n}\sum X_i$ is again MLE.

Normal Approximation to Poisson, when λ is large:

$$X_1 \sim N(\lambda, (\sqrt{\lambda})^2)$$

e.g. Lab 1 (Q3): $X_1(k) = \#$ purines in a block of size k bps.

$$X_1(k) \sim \mathrm{Poisson}(\lambda_0 k), \qquad \lambda_0 = \text{purine rate per bps}$$

$$EX_1(k) = \lambda_0 k, \qquad \mathrm{SD}(X_1(k)) = \sqrt{\lambda_0}\sqrt{k}$$

The Method of Maximum Likelihood

Suppose $X_1, \ldots, X_n$ have a joint density $f(x_1, \ldots, x_n|\theta)$. Given $X_i = x_i$, the likelihood of θ as a function of $x_1, \ldots, x_n$ is defined as

$$\text{lik}(\theta) = f(x_1, \ldots, x_n|\theta)$$

MLE $\hat{\theta}$ is the maximizer of lik(θ) over parameter domain of θ.

If $X_1, \ldots, X_n$ are i.i.d.

$$\text{lik}(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

and log-likelihood is given by

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Facts:

1. $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ i.e. $\theta = (\mu, \sigma^2)$

$$\text{MLE of } \mu \ : \ \hat{\mu} = \frac{1}{n}\sum X_i = \bar{X}_n$$

$$\text{MLE of } \sigma^2 \ : \ \hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \bar{X}_n)^2$$

Proof:

$$f(x_1, \ldots, x_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma}\right]^2\right\}$$

$$l(\mu, \sigma) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\begin{cases} \dfrac{\partial l}{\partial \mu} = -\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^{n} -2(x_i - \mu) = 0 \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2}\dfrac{1}{\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}$$

$$\Longleftrightarrow \quad \begin{aligned} \hat{\mu} &= \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \end{aligned}$$

Note: $E\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2$ — biased

$$Es^2 = \sigma^2, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Therefore, MLE of σ^2 is **biased**.

2. $(N_1, \ldots, N_m) \sim \text{Multinomial}(n; p_1, \ldots, p_m)$. MLE of p_j is $\hat{p}_j = \frac{N_j}{n}$, $j = 1, \ldots, m$

   Proof: Rice p. 259. Given $N_j = n_j$ (different notation $j \longrightarrow i$, $n_j \longrightarrow x_i$)

$$\begin{aligned} f(n_1, \ldots, n_m | p_1, \ldots, p_m) &= \frac{n!}{\prod_{j=1}^m n_j!} \prod_{j=1}^m p_j^{n_j} \\ l(p_1, \ldots, p_m) &= \log n! - \sum_{j=1}^m \log n_j! + \sum_{j=1}^m n_j \log p_j \end{aligned}$$

   maximize $l(p_1, \ldots, p_m)$ subject to $\sum_{j=1}^m p_j = 1$ by Lagrange multiplier.

$$L(p_1, \ldots, p_m; \lambda) = l(p_1, \ldots, p_m) + \lambda \left(\sum_{j=1}^m p_j - 1 \right)$$

   Set partial derivatives to zero,

$$\begin{aligned} \hat{p}_j &= -\frac{n_j}{\lambda} \\ 1 = \sum \hat{p}_j &= -\frac{\sum n_i}{\lambda} = -\frac{n}{\lambda} \\ \Longrightarrow \hat{p}_j &= \frac{n_j}{n}, \qquad j = 1, \ldots, m \end{aligned}$$

3. $X_1, \ldots, X_n$ i.i.d Poisson (λ) (cf Rice p. 254)

   MLE of λ: $\hat{\lambda} = \bar{X} = \frac{1}{n} \sum X_i$

   Proof:

$$\begin{aligned} P(X_1, \ldots, X_n | \lambda) &= \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ l(\lambda) &= \log P(X_1, \ldots, X_n | \lambda) \\ &= \sum_{i=1}^n x_i \log \lambda - n\lambda - \sum_{i=1}^n \log x_i! \end{aligned}$$

$$\frac{\partial l}{\partial \lambda} = 0 \iff \sum_{i=1}^n x_i \frac{1}{\lambda} - n = 0 \quad \Longrightarrow \hat{\lambda} = \frac{1}{n} \sum x_i = \bar{X}$$

   In lab 1, you have also been concerned with 1st and 2nd order Markov chains (stationary). For example, we have $m = 4$ bases. $b_1, \ldots, b_n$ is the observed DNA sequence.

4. $b_i \in \{A, G, C, T\}$, $i = 1, \ldots, n$. Let π_A, π_G, π_C, π_T be the stationary distribution with $\sum \pi = 1$ and let

$$P = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} A \quad G \quad C \quad T \\ \left(\begin{array}{c} \\ P_{ts} \\ \\ \end{array} \right) \end{array}$$

   be the transition matrix. Row sums are 1.

Or $A \longrightarrow 1$, $G \longrightarrow 2$, $C \longrightarrow 3$, $T \longrightarrow 4$,

$$\sum \pi_j = 1, \qquad j = 1, \ldots, 4 \qquad\qquad \sum_{j=1}^{4} P_{kj} = 1 \qquad \forall\, k = 1, \ldots, 4$$

Likelihood of $b_1, \ldots, b_n$, $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ is

$$\begin{aligned} P(b_1, \ldots, b_n | \pi, \mathbf{P}) &= \prod_{i=1}^{n-1} P_{b_i, b_{i+1}} \pi_{b_1} \\ &= \prod_{k=1}^{4} \prod_{j=1}^{4} P_{kj}^{N_{kj}} \pi_{b_1} \text{ where } N_{kj} = \# \text{ pairs } (k, j) \text{ in the sequence} \end{aligned}$$

$$l\mathbf{P}) \cong \sum_{i=1}^{4} \sum_{i=1}^{4} N_{kj} \log P_{kj}$$

Maximize $l\mathbf{P})$ subject to $\sum_{j=1}^{4} P_{kj} = 1$, $k = 1, \ldots, 4$ by Lagrange multipliers:

$$L(\mathbf{P}) = \sum_{k=1}^{4} \sum_{j=1}^{4} N_{kj} \log P_{kj} + \sum_{k=1}^{4} \lambda_k \left(\sum_{j=1}^{4} P_{kj} - 1 \right)$$

Setting partial derivatives to zero

$$\frac{\partial L}{\partial P_{kj}} = \frac{N_{kj}}{P_{kj}} + \lambda_k = 0$$

This leads to

$$\begin{aligned} \hat{P}_{kj} &= -\frac{N_{kj}}{\lambda_k} \\ 1 = \sum_{j=1}^{4} \hat{P}_{kj} &= -\frac{N_k}{\lambda_k} \qquad N_k = \sum_{j=1}^{4} N_{kj} = \# k'\text{s} \\ \lambda_k = -N_k &\implies \hat{P}_{kj} = \frac{N_{kj}}{N_k} \end{aligned}$$

π's can be determined uniquely by

$$\pi_j = \sum_{k=1}^{4} P_{kj} \pi_k \qquad j = 1, \ldots, 4$$

Plug in $\hat{\pi}_k = \frac{N_k}{n}$ to verify the equations hold.

5. 2nd order Stationary Markov

   Similarly, MLE $\hat{P}(k|ij) = \dfrac{N_{ijk}}{N_{ij}}$ i.e. conditioning on ij

6. Hardy-Weinberg equilibrium - cf. Rice.