Lecture 17: Maximum Likelihood Estimation (MLE)

Examples:

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

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

   $$ S_n = X_1 + \ldots + X_n \sim \text{Binomial} \ (n, p) $$

   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.

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

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

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

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

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

   $m = 2 \rightarrow$ Back to Binomial.

   $$ \text{Var}(S_n) = np(1 - p) $$

   $$ E(S_n) = np $$

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

In Multinomial notation,

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

which leads to

$$ \text{Cov}(N_1, N_0) = \text{Cov}(S_n, n - S_n) $$

$$ = -\text{Cov}(S_n, S_n) $$

$$ = -\text{Var}(S_n) $$

$$ = -np(1 - p) $$
In general?
\[ E(N_j) = np_j, \quad \text{Var}(N_j) = np_j(1 - p_j) \]
\[ \text{Cov}(N_k, N_j) = -np_k p_j \quad k \neq j \]

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

\[
\text{Cov}(N_k, N_j) = \text{Cov}\left(\sum_{i=1}^{n} 1_{\{X_i = k\}}, \sum_{i=1}^{n} 1_{\{X_i = j\}}\right) \\
= \sum_{i=1}^{n} \sum_{i=1}^{n} \text{Cov}\left(1_{\{X_i = k\}}, 1_{\{X_i = j\}}\right) \quad k \neq j \\
= \sum_{i=1}^{n} \text{Cov}\left(1_{\{X_i = k\}}, 1_{\{X_i = j\}}\right) \\
= \sum_{i=1}^{n} \left[E(1_{\{X_i = k\}} \cdot 1_{\{X_i = j\}}) - p_k p_j\right] \\
= -np_k p_j \quad k \neq j
\]

When \( n \to \infty, p \to 0, np \to \lambda > 0, \quad X \sim \text{Bin}(n, p) \)

\[
P(X = j) \to e^{-\lambda} \frac{\lambda^j}{j!} \sim \text{Poisson}(\lambda), \quad j = 0, 1, \ldots
\]

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

\[
EX_1 = \lambda, \quad \text{Var}(X_1) = \lambda, \quad \text{SD}(X_1) = \sqrt{\lambda}
\]

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

Normal Approximation to Poisson, when \( \lambda \) 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.

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

\[
EX_1(k) = \lambda_0 k, \quad \text{SD}(X_1(k)) = \sqrt{\lambda_0 \sqrt{k}}
\]
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 $\theta$ 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 $\text{lik}(\theta)$ over parameter domain of $\theta$.

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. $\theta = (\mu, \sigma^2)$

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

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

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

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

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

$\Rightarrow \hat{\mu} = \bar{x} \quad \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$

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

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

Therefore, MLE of $\sigma^2$ 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 \(j \rightarrow i, n_j \rightarrow x_i\))

\[
f(n_1, \ldots, n_m | p_1, \ldots, p_m) = \frac{n!}{\prod_{j=1}^{m} n_j!} \prod_{j=1}^{m} p_j^{n_j}
\]

\[
l(p_1, \ldots, p_m) = \log n! - \sum_{j=1}^{m} \log n_j! + \sum_{j=1}^{m} n_j \log p_j
\]

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,

\[
\frac{\partial L}{\partial p_j} = \frac{n_j}{\lambda} = 0 \quad \Rightarrow \quad \hat{p}_j = \frac{n_j}{n}, \quad j = 1, \ldots, m
\]

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

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

Proof: \(P(X_1, \ldots, X_n | \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\)

\[
l(\lambda) = \log P(X_1, \ldots, X_n | \lambda) = \sum_{i=1}^{n} x_i \log \lambda - n \lambda - \sum_{i=1}^{n} \log x_i!
\]

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

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 \(\pi_A, \pi_G, \pi_C, \pi_T\) be the stationary distribution with \(\sum \pi = 1\) and let

\[
P = \begin{pmatrix}
A & G & C & T \\
G & P_{ts} & & \\
C & & & \\
T & & & 
\end{pmatrix}
\]

be the transition matrix. Row sums are 1.
Or $A \rightarrow 1, G \rightarrow 2, C \rightarrow 3, T \rightarrow 4,$

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

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

$$P(b_1, \ldots, b_n | \pi, P) = \prod_{i=1}^{n-1} P_{b_i, b_{i+1}} \pi_{b_i}$$

$$= \prod_{k=1}^{4} \prod_{j=1}^{4} P_{kj}^{N_{kj}} \pi_{b_i} \text{ where } N_{kj} = \# \text{ pairs } (k, j) \text{ in the sequence}$$

$$l(P) \simeq \sum_{i=1}^{4} \sum_{j=1}^{4} N_{kj} \log P_{kj}$$

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

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

Setting partial derivatives to zero

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

This leads to

$$\hat{P}_{kj} = -\frac{N_{kj}}{\lambda_k}$$

$$1 = \sum_{j=1}^{4} \hat{P}_{kj} = -\frac{N_k}{\lambda_k} \quad N_k = \sum_{j=1}^{4} N_{kj} = \# k's$$

$$\lambda_k = -N_k \implies \hat{P}_{kj} = \frac{N_{kj}}{N_k}$$

$\pi$'s can be determined uniquely by

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

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

5. 2nd order Stationary Markov

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