Maximum Likelihood Estimation

STAT 37710 / CMSC 35300
Rebecca Willett and Yuxin Chen

Recall parameter estimation problem

General problem statement:

We observe

$$z_i \overset{\text{iid}}{\sim} p_\theta,\ \theta \in \Theta$$

and the goal is to determine the θ that produced $\{z_i\}_{i=1}^n$.

Given a collection of observations $z_1, ..., z_n$ and a probability model

$$p(z_1, ..., z_n|\theta)$$

parameterized by the parameter θ, determine the value of θ that best matches the observations.

Estimation using the likelihood

Definition: Likelihood function

$p(z|\theta)$ as a function of θ with z fixed is called the "likelihood function".

If the likelihood function carries the information about θ brought by the observations $z := \{z_i\}_i$, how do we use it to obtain an estimator?

Definition: Maximum Likelihood Estimation

$$\widehat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta \in \Theta} p(z|\theta)$$

is the value of θ that maximizes the density at z. Intuitively, we are choosing θ to maximize the probability of occurrence for z.

Maximum likelihood estimation (MLE)

MLEs are a very important type of estimator for the following reasons:

- The MLE is often simple and easy to compute
- MLEs are invariant under reparameterization
- MLEs often have asymptotic optimal properties (e.g. consistency (MSE $\to 0$ as $n \to \infty$)
- MLE occurs naturally in composite hypothesis testing (i.e., GLRT)

Computing the MLE

If the likelihood function is differentiable and concave, then $\widehat{\theta}$ is found from

$$\frac{\partial \log p(z|\theta)}{\partial \theta} = 0$$

If multiple solutions exist, then the MLE is the solution that maximizes $\log p(z|\theta)$. That is, take the global maximizer.

Note: It is possible to have multiple global maximizers that are all MLEs!

Example: Estimating the mean and variance of a Gaussian

$$z_i = A + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, \cdots, n$$

$$\theta = [A, \sigma^2]^\top$$

$$\frac{\partial \log p(z|\theta)}{\partial A} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (z_i - A)$$

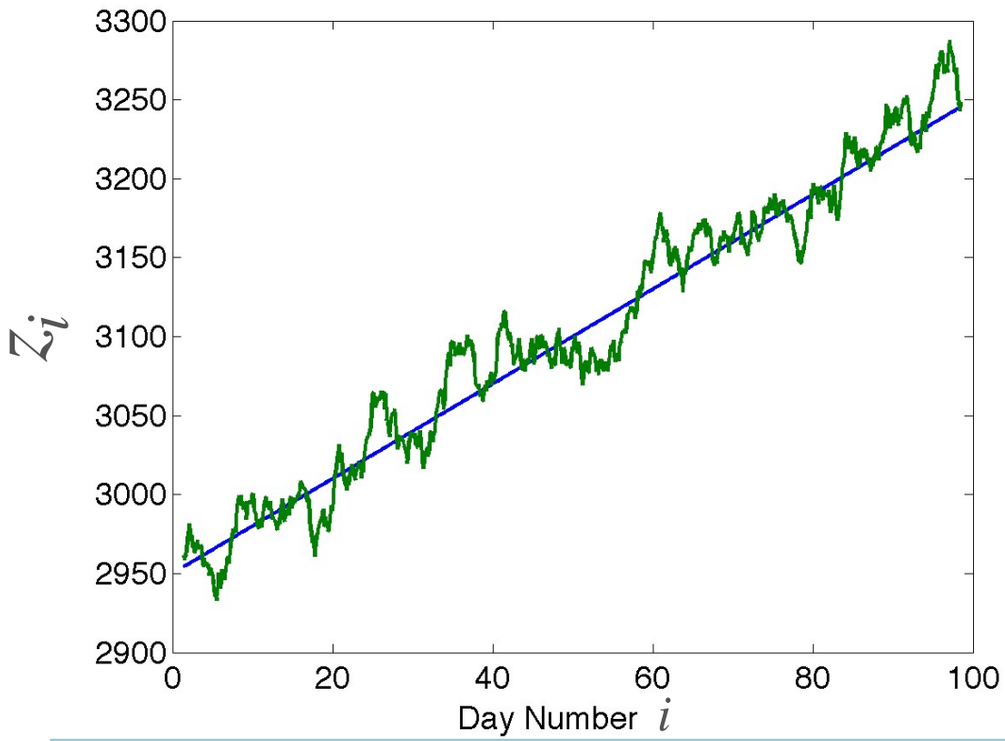
$$\frac{\partial \log p(z|\theta)}{\partial \sigma^2} = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^{n} (z_i - A)^2$$

$$\Rightarrow \widehat{A} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

$$\Rightarrow \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (z_i - \widehat{A})^2$$

Note: $\widehat{\sigma^2}$ is biased!

Example: Stock Market (Dow-Jones Industrial Avg.)



Based on this plot we might conjecture that the data is "on average" increasing. Probability model:

$$z_i = A + Bi + \epsilon_i$$

A, B are unknown parameters, ϵ_i white Gaussian noise to model fluctuations.

$$p(\{z_i\}_i | A, B) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - A - Bi)^2 \right\}$$

Linear models

$z_i = (x_i, y_i)$	training samples for $i = 1, \ldots, n$
$x_i \in \mathbb{R}^p$	feature vector
$y_i \in \mathbb{R}$	label
$\mathbb{E} y_i = x_i^\top \theta$	model

Goal: estimate θ so that for a new (test) feature vector x we can predict a label $\hat{y} = x^\top \theta$

Note: many generalizations of this model exist throughout machine learning

Linear models

The basic linear model is:

$$\mathbb{E}y = X\theta$$

where $\theta \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ with $n > p$ is known and full rank, and $\theta \in \mathbb{R}^p$ is a p-dimensional parameter or weight vector. In other words, we can write

$$\mathbb{E}y_i = x_i^\top \theta = \sum_{j=1}^{p} x_{i,j}\theta_j$$

where θ_j is the j^{th} element of θ, x_i is the i^{th} feature vector, and $x_{i,j}$ is the j^{th} element of x_i, and $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$.

Example: Colored Gaussian Noise

$$
\begin{aligned}
y &\sim \mathcal{N}(X\theta, \Sigma), \quad \theta \in \mathbb{R}^p, \quad \Sigma, X \text{ known} \\
p(y|\theta) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(y - X\theta)^\top \Sigma^{-1}(y - X\theta)\}
\end{aligned}
$$

The value of $\widehat{\theta}_{\mathsf{MLE}}$ is given by,

$$
\begin{aligned}
\widehat{\theta}_{\mathsf{MLE}} &= \arg\min_\theta -\log p(y|\theta) \\
&= \arg\min_\theta (y - X\theta)^\top \Sigma^{-1}(y - X\theta) \\
&= (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y
\end{aligned}
$$

Generalized Linear Models

When we observe

$$y = X\theta + \epsilon$$

and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then the MLE of θ can be computed by projection onto the subspace spanned by the columns of X.
For more general probability models, the MLE may not have this form. Generalized linear models provide a unifying framework for modeling the relationship between z and θ. The basic idea is that we define a link function g such that

$$g(\mathbb{E}y) = X\theta.$$

Typically there is no closed-form expression for $\hat{\theta}_{\mathsf{MLE}}$ and it must be computed numerically.

Example: Logistic Regression

$$\theta \in \mathbb{R}^p, \quad x_i \in \mathbb{R}^p \text{ known}$$

$$y_i | x_i \sim \text{Bernoulli}(p_i)$$

$$p_i = (1 + \exp(-x_i^\top \theta))^{-1} \text{ or } \log\left(\frac{p_i}{1-p_i}\right) = x_i^\top \theta$$

$$p(y|\theta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$

$$p_i^{y_i}(1-p_i)^{1-y_i} = \begin{cases} (1+\exp(-x_i^\top\theta))^{-1}, & y_i = 1 \\ (1+\exp(x_i^\top\theta))^{-1}, & y_i = 0 \end{cases}$$

Define $\tilde{y}_i := \begin{cases} 1, & \text{if } y_i = 1 \\ -1, & \text{if } y_i = 0 \end{cases}$

$$p_i^{y_i}(1-p_i)^{1-y_i} = (1+\exp(-\tilde{y}_i x_i^\top \theta))^{-1}$$

Example: Logistic Regression (cont.)

$$p(y|\theta) = \prod_{i=1}^{n} (1 + \exp(-\tilde{y}_i x_i^\top \theta))^{-1}$$

The value of $\widehat{\theta}_{\mathsf{MLE}}$ is given by

$$\widehat{\theta}_{\mathsf{MLE}} = \arg\min_{\theta} -\log p(y|\theta) = \arg\min_{\theta} \sum_{i=1}^{n} \log(1 + \exp(-\tilde{y}_i x_i^\top \theta))$$

which can be computed numerically, e.g. via gradient descent.

Invariance of MLE

Suppose we wish to estimate the function $g = G(\theta)$ and not θ itself. Intuitively we might try

$$\widehat{g} = G(\widehat{\theta})$$

where $\widehat{\theta}$ is the MLE of θ.

Remarkably, it turns out that $\widehat{g}$ is the MLE of g.

This very special invariance principle is summarized in the following theorem.

Theorem: Invariance of the MLE

Let $\widehat{\theta}$ denote the MLE of θ. Then $\widehat{g} = G(\widehat{\theta})$ is the MLE of $g = G(\theta)$.

Example:

Let $z = [z_1, \cdots, z_n]^\top$ where $z_i \sim \text{Poisson}(\lambda)$. Given z, find the MLE of the probability that $y \sim \text{Poisson}(\lambda)$ exceeds the mean λ.

$$\begin{aligned} G(\lambda) &= \mathbb{P}(y > \lambda) \\ &= \sum_{k=\lfloor \lambda+1 \rfloor}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \quad ; \quad \lfloor a \rfloor = \text{ largest integer} \leq a \end{aligned}$$

The MLE of g is

$$\widehat{g} = \sum_{k=\lfloor \widehat{\lambda}+1 \rfloor}^{\infty} e^{-\widehat{\lambda}} \frac{\widehat{\lambda}^k}{k!}$$

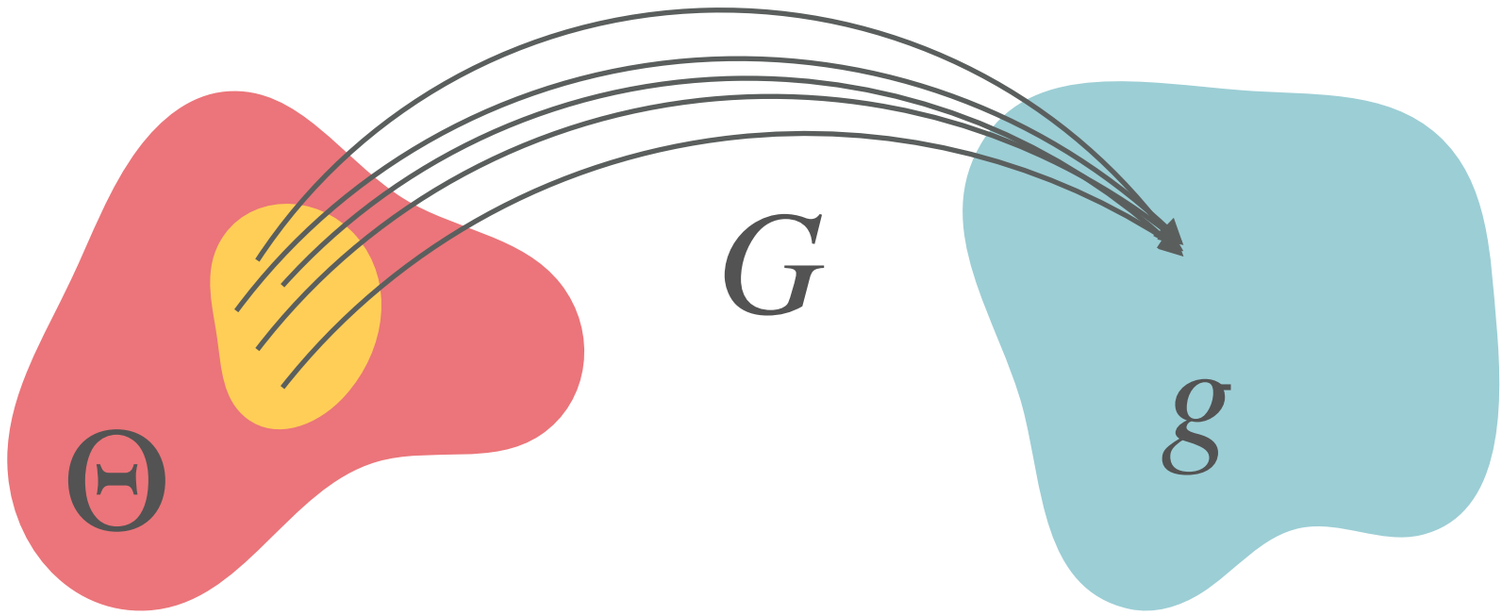
where $\widehat{\lambda}$ is the MLE of λ:

$$\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

Proof sketch

Define the "induced" log likelihood function:

$$L(z|g) \equiv \max_{\theta:G(\theta)=g} \log p(z|\theta)$$

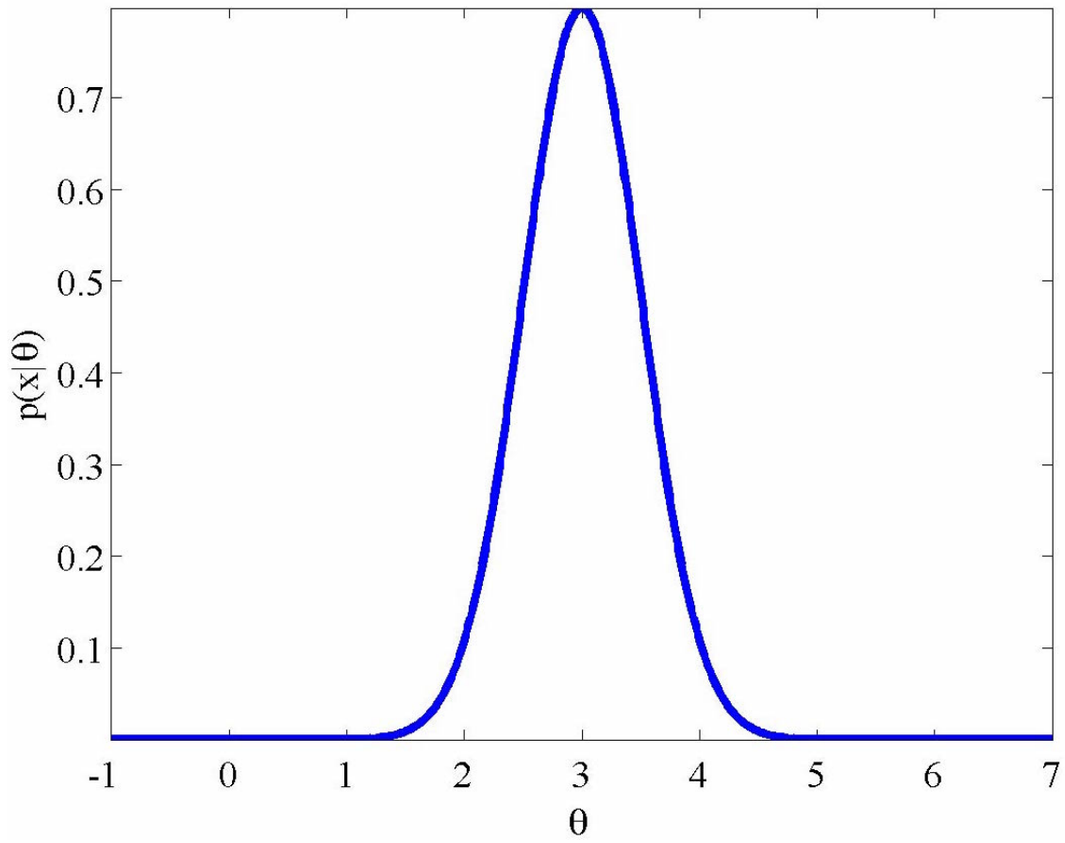


The MLE of g is

$$\begin{aligned}
\widehat{g} &= \arg\max_g L(z|g) \\
&= \arg\max_g \max_{\theta:G(\theta)=g} \log p(z|\theta) \\
&= G(\widehat{\theta}) \quad , \quad \text{where } \widehat{\theta} = \text{MLE of } \theta
\end{aligned}$$

Estimator accuracy

Consider the likelihood function $p(z|\theta)$ where θ is a scalar unknown (parameter).



We can plot the likelihood as a function of the unknown. The more "peaky" or "spiky" the likelihood function, the easier it is to determine the unknown parameter.

The peakiness is effectively measured by the negative of the second derivative of the log-likelihood at its peak.

Fisher information

In general, the curvature will depend on the observed data:

$$-\frac{\partial^2 \log p(z|\theta)}{\partial \theta^2} \quad \text{is a function of } z$$

Thus an average measure of curvature is more appropriate.

$$-\mathbb{E}\left[\frac{\partial^2 \log p(z|\theta)}{\partial \theta^2}\right]$$

The expectation averages out randomness due to the data and is a function of θ alone.

Definition: Fisher Information

$$I(\theta) := \mathbb{E}\left[\left(\frac{\partial \log p(z|\theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log p(z|\theta)}{\partial \theta^2}\right]$$

is the *Fisher Information*. Here the derivative is evaluated at the true value of θ and the expectation is with respect to $p(z|\theta)$.

Asymptotic Distribution of MLE

Let $z_i \overset{\text{iid}}{\sim} p_{\theta^*}$, $i = 1, \ldots, n$, where $\theta^* \in \mathbb{R}^p$,

$$L_n(\theta) := \sum_{i=1}^{n} \log p(z_i|\theta) \qquad \text{and} \qquad \widehat{\theta}_n = \arg\max_{\theta} \; L_n(\theta),$$

assume $\frac{\partial L_n(\theta)}{\partial \theta_j}$ and $\frac{\partial^2 L_n(\theta)}{\partial \theta_j \partial \theta_k}$ exist for all j, k, and the "regularity condition" $\mathbb{E}\left[\frac{\partial \log p(z|\theta)}{\partial \theta}\right] = 0$ for all θ holds. Then

$$\widehat{\theta}_n \overset{\text{asymp.}}{\sim} \mathcal{N}(\theta^*, n^{-1} I^{-1}(\theta^*))$$

where $I(\theta^*)$ is the Fisher-Information Matrix (FIM), whose elements are given by

$$[I(\theta^*)]_{j,k} = -\mathbb{E}\left[\frac{\partial^2 \log p(z|\theta)}{\partial \theta_j \partial \theta_k}\Big|_{\theta=\theta^*}\right]$$

Regularity condition

The regularity condition amounts to assuming that we can interchange order of differentiation and integration to compute

$$\mathbb{E}\left[\frac{\partial \log p(z|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}\right] = \int \frac{\partial \log p(z|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} p(z|\theta^*)dx$$

$$= \int \frac{1}{p(z|\theta^*)} \frac{\partial p(z|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} p(z|\theta^*)dx = \int \frac{\partial p(z|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} dx$$

$$= \frac{\partial}{\partial \theta}\left[\int p(z|\theta)dx\right]\Big|_{\theta=\theta^*} = 0,$$

since $\int p(z|\theta)dx = 1$ for all θ and the derivative of a constant is 0. The last line, where integration and differentiation are interchanged, is only possible for "regular" likelihood functions. This is simply the Fundamental Theorem of Calculus applied to $p(z|\theta)$. As long as $p(z|\theta)$ is absolutely continuous w.r.t. Lebesgue measure (i.e., when the derivative is well-defined), this is possible.

This is true for many distributions, but not true when the support of z depends on θ (e.g. $y \sim \text{Unif}(0, \theta)$).

Note:

$$\mathbb{E}[\widehat{\theta}] \to \theta^*$$

$$\text{Cov}(\widehat{\theta}) \to \frac{1}{n}\boldsymbol{I}^{-1}(\theta^*)$$

$\Rightarrow \widehat{\theta}$ is consistent and efficient asymptotically (i.e., asymptotically achieves CRLB)

Example:

$$z_i \overset{\text{iid}}{\sim} \mathcal{N}(A \cdot \mathbf{1}_{n\times 1}, \sigma^2 \boldsymbol{I}_n)$$

$$\theta = [A, \sigma^2]^\top$$

$$\widehat{A} = \frac{1}{n}\sum_{i=1}^{n} z_i \sim \mathcal{N}\left(A, \frac{\sigma^2}{n}\right)$$

$$s := \sum_{i=1}^{n} \frac{(z_i - \widehat{A})^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\widehat{\sigma^2} = \left(\frac{\sigma^2}{n}\right) s$$

Example: (cont.)

For large n, the central limit theorem tells us that

$$\chi^2_n \approx \mathcal{N}(n, 2n).$$

Therefore,

$$s \approx \mathcal{N}(n-1, 2(n-1)) \leftarrow \text{approximately distributed}$$

Hence,

$$\begin{aligned} \widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \widehat{A})^2 = \frac{\sigma^2}{n}s \\ &\approx \mathcal{N}\left(\frac{(n-1)}{n}\sigma^2, \frac{2(n-1)\sigma^4}{n^2}\right) \end{aligned}$$

Example: (cont.)

Moreover, for large n

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}\right] &= \begin{bmatrix} A \\ \frac{n-1}{n}\sigma^2 \end{bmatrix} \to \begin{bmatrix} A \\ \sigma^2 \end{bmatrix} = \theta^* \\
C_{\widehat{\theta}} &= \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \frac{2(n-1)\sigma^4}{n^2} \end{bmatrix} \to \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix} \\
&= I^{-1}(\theta^*) \leftarrow \text{ inverse Fisher Info. Matrix}
\end{aligned}$$

Hence,

$$\widehat{\theta} \sim \mathcal{N}(\theta^*, \boldsymbol{I}^{-1}(\theta^*)) \text{ for large } n$$