7. Linear Models and Maximum Likelihood Estimation

ECE 830 & CS 761, Spring 2016

Primary Goal

General problem statement:

We observe

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

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

Given a collection of observations $y_1, ..., y_n$ and a probability model

$$p(y_1, ..., y_n|\theta)$$

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

Estimation Using the Likelihood

Definition: Likelihood function

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

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

Definition: Maximum Likelihood Estimation

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

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

Maximum Likelihood Estimation

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

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

Computing the MLE

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

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

If multiple solutions exist, then the MLE is the solution that maximizes $\log p(y|\theta)$. 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

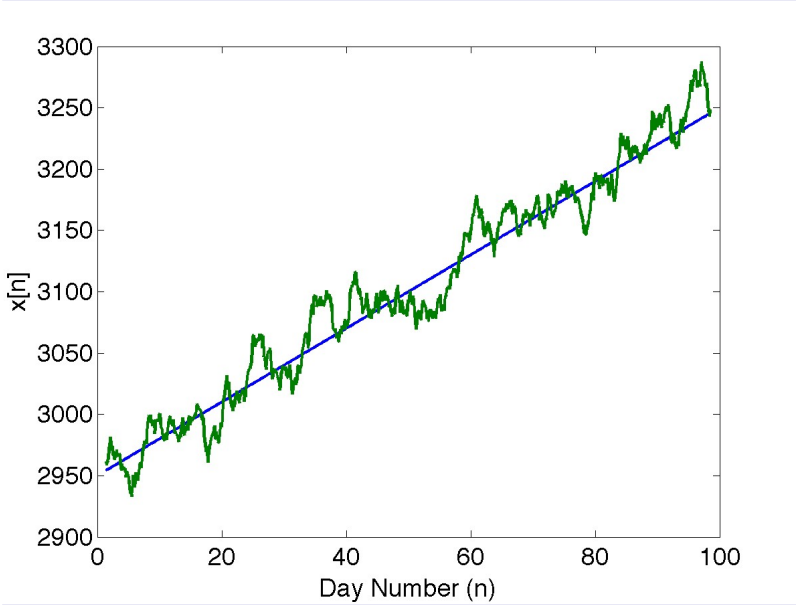
$$y_i = A + \nu_i, \qquad \nu_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, \cdots, n$$

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

$$\begin{aligned}
\frac{\partial \log p(y|\theta)}{\partial A} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - A) \\
\frac{\partial \log p(y|\theta)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - A)^2 \\
\Rightarrow \widehat{A} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\Rightarrow \widehat{\sigma^2} &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{A})^2
\end{aligned}$$

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:

$$y_i = A + Bi + \nu_i$$

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

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

Example: A modern Example - Imaging

Image processing can involve complicated estimation problems. For example, suppose we observe a moving object with noise. This image is blurry and noisy and our goal is to restore this image by debarring and denoising.

Example: A modern Example - Imaging

We can model the moving part of the observed image as

$$y = \underbrace{\theta}_{\text{noise-free blurry image}} + \underbrace{\nu}_{\text{noise}}$$
$$= \underbrace{x * w}_{\text{motion blur (convolution)}} + \nu$$

where the parameter of interest w is the ideal (non-blurry) image and x represents the blur model or point spread function.

Observation model:

$$y = Xw + \nu, \qquad \nu \sim N(0, \sigma^2 I)$$
$$y \sim \mathcal{N}(Xw, \sigma^2 I)$$

where X is a circulant matrix with the first row corresponding to x.

Linear Models and Subspaces

The basic linear model is:

$$\theta = Xw$$

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

$$\theta = \sum_{i=1}^{p} w_i x_i$$

where w_i is the i^{th} element of w and x_i is the i^{th} column of X.

We say that θ is in a subspace spanned by the columns of X

Subspaces

Consider a set of vectors $x_1, x_2, \ldots, x_p \in \mathbb{R}^n$. The span of these vectors is the set of all vectors $\theta \in \mathbb{R}^n$ that can be generated from linear combinations of the set

$$\text{span}\left(\{x_i\}_{i=1}^p\right) := \left\{\theta : \theta = \sum_{i=1}^p w_i x_i, \quad w_1, \ldots, w_p \in \mathbb{R}\right\}$$

p-dimensional subspace

If the $x_1, \ldots, x_p \in \mathbb{R}^n$ are linearly independent, then their span is a p-**dimensional subspace** of $\mathbb{R}^n$.

Hence, even though the signal θ is a length-n vector, the fact that it lies in the subspace $\mathcal{X}$ means that it is actually a function of only $p \leq n$ free parameters or "degrees of freedom".

Example: $n = 3$

$$x_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\mathcal{X} = \text{span}\,(x_1, x_2) = \left\{ \begin{bmatrix} a \\ b \\ 0 \end{bmatrix} : a, b \in \mathbb{R} \right\}$$

Example: $n = 3$

$$x_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\mathcal{X} = \text{span}\,(x_1, x_2) = \left\{ \begin{bmatrix} a \\ a \\ b \end{bmatrix} : a, b \in \mathbb{R} \right\}$$

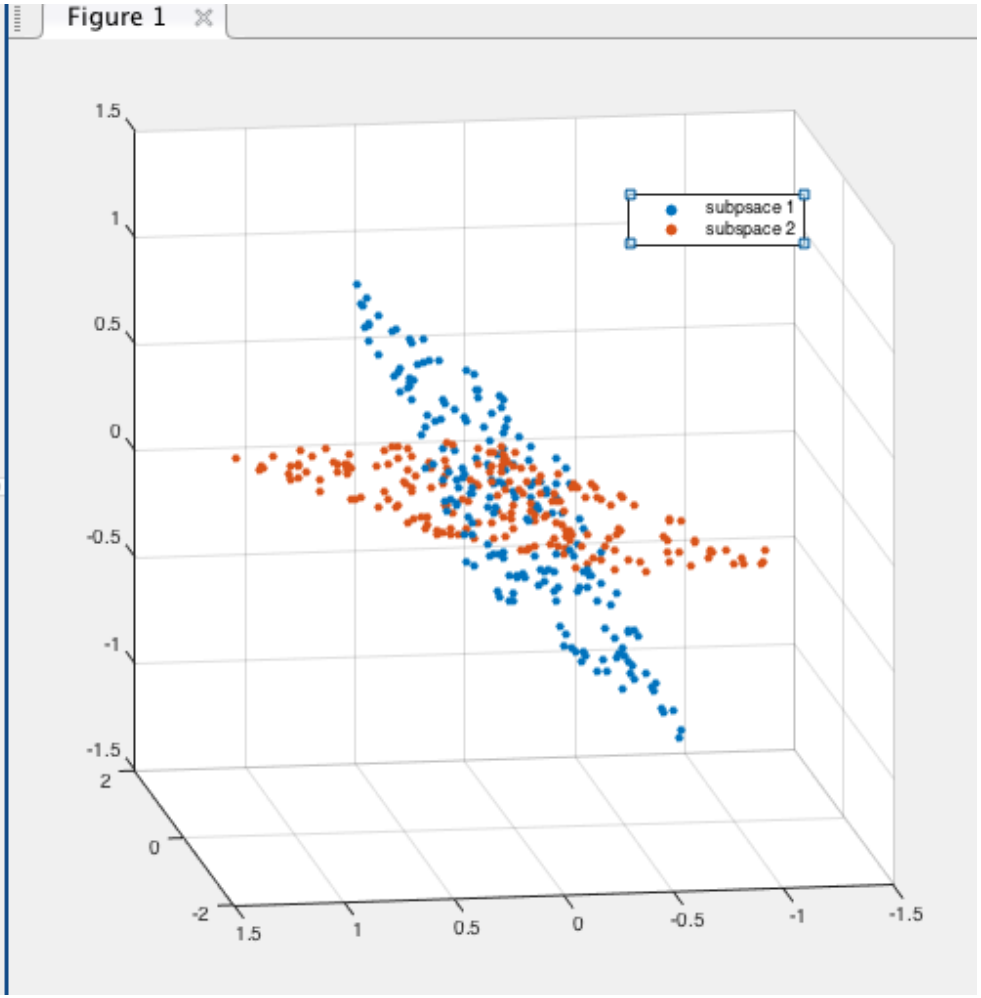
subspace.m

```
N = 200;
n = 3;
p = 2;

X = orth(randn(n,p));
w = 2*rand(p,N)-1;
theta = X*w;

X2 = orth(randn(n,p));
w2 = 2*rand(p,N)-1;
theta2 = X2*w2;

figure(1);clf;
scatter3(theta(1,:),theta(
   theta(3,:),20,'filled')
hold on
scatter3(theta2(1,:),theta
   theta2(3,:),20,'filled'
legend('subpsace 1','subsp
```



Examples

Example: "Constant" subspace

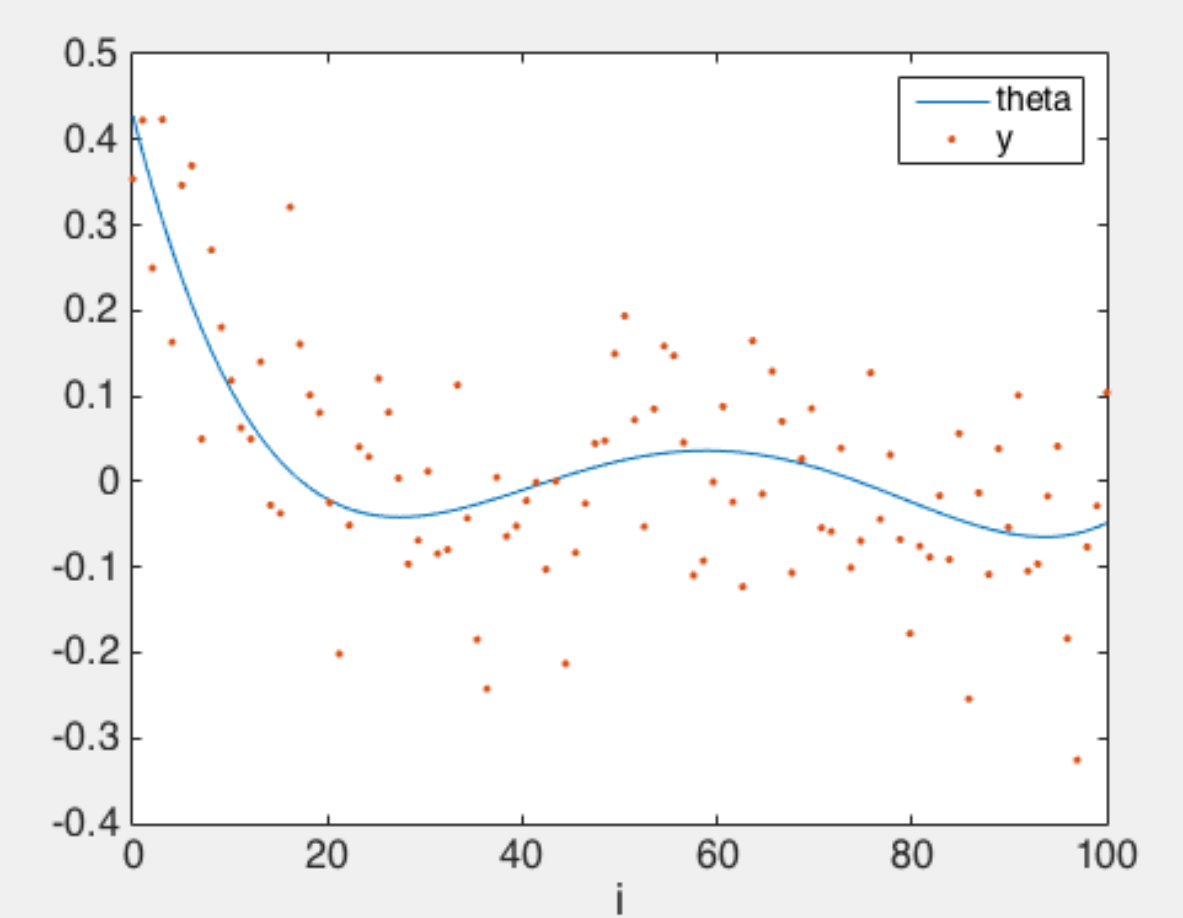
$$p = 1, X = x_1 = (1/\sqrt{n}) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

$\theta = x_1 w_1$ is a constant signal with unknown amplitude.

Example: Recommender system

We have n movies in a database, and for a new client want to estimate how much they will like each movie, denoted $\theta \in \mathbb{R}^n$. We have p existing archetype clients for whom we know their movie preferences; the preferences of the i^{th} archetype client are denoted $x_i \in \mathbb{R}^n$, $i = 1, \ldots, p$. We will assume that θ is a weighted combination of these archetype profiles.

Example: Polynomial regression



$$\theta_i = w_1 + w_2 i + w_3 i^2 + \cdots + w_p i^{p-1}$$

$$X = \text{Vandermonde matrix}$$

Using subspace models

- Imagine we measure a function or signal in noise,

$$y = \theta + \nu,$$

where elements of ν are independent realizations from a Gaussian density and

$$\theta = Xw$$

- Question: How can we use knowledge of X to estimate θ from y?
- Let $\mathcal{X}$ be the subspace spanned by the columns of X, and let $\mathcal{X}_\perp$ be its orthogonal subspace. Then we can uniquely write

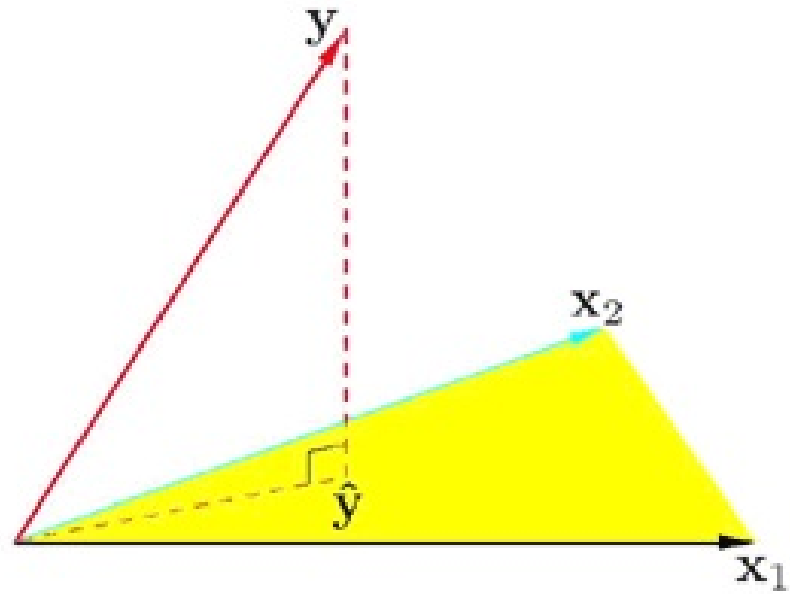
$$y = u + v,$$

where $u \in \mathcal{X}$ and $v \in \mathcal{X}_\perp$. Since we know $\theta \in \mathcal{X}$, v can be considered pure "noise" and it makes sense to remove it.
- Answer: Orthogonal projection – find the point in $u \in \mathcal{X}$ which is closest to the observed y.
- In general $\theta \neq u$ because ν has some component in $\mathcal{X}$.

Orthogonal projection

Given a point $y \in \mathbb{R}^n$ and subspace $\mathcal{X}$, let's say we want to find the point $\hat{y} \in \mathcal{X}$ which is closest to y:

$$\begin{aligned}
\hat{y} &= \arg\min_{\theta \in \mathcal{X}} \|\theta - y\|^2 \\
&= \arg\min_{\theta = Xw, w \in \mathbb{R}^p} \|\theta - y\|^2 \\
&= \arg\min_{\theta = Xw, w \in \mathbb{R}^p} \|Xw - y\|^2 \\
&= \underbrace{X \quad \underbrace{(X^\top X)^{-1} X^\top}_{\text{pseudoinverse of } X}}_{=: P_X, \text{ projection matrix}} \quad y.
\end{aligned}$$



Essentially, we want to keep the component of y that lies in the subspace $\mathcal{X}$ and remove or kill the component in the orthogonal subspace.

Sinusoid subspace

Example: Sinusoid subspace

Let

$$\theta_i = A\cos(2\pi f i + \phi), \quad i = 1, \ldots, n$$

where frequency f is known but amplitude A and phase ϕ are unknown. Goal: estimate A and ϕ from

$$y = \theta + \nu$$

where ν is white Gaussian noise.

First we need to express $\theta = [\theta_1, \ldots, \theta_n]^\top$ as an element in a 2-dimensional subspace; i.e. write $\theta = Xw$ where X is a known $n \times 2$ matrix and w is unknown. Then we can apply orthogonal projection.

Real solution:

First note that $\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$, so that

$$\theta_i = A\cos(2\pi f i + \phi) = A\cos(2\pi f i)\cos(\phi) - A\sin(2\pi f i)\sin(\phi).$$

Let $w_1 := A\cos(\phi)$ and $w_2 := A\sin(\phi)$, and

$$X = \begin{bmatrix} \cos(2\pi f) & \sin(2\pi f) \\ \cos(4\pi f) & \sin(4\pi f) \\ \vdots & \vdots \\ \cos(2n\pi f) & \sin(2n\pi f) \end{bmatrix}.$$

Note $\theta = Xw$ where $X \in \mathbb{R}^{n\times 2}$ and $w \in \mathbb{R}^2$.
Use pseudoinverse to compute $\hat{w}$, then set

$$\widehat{A} = \sqrt{\hat{w}_1^2 + \hat{w}_2^2}$$
$$\widehat{\phi} = \arctan(-\hat{w}_2/\hat{w}_1)$$

MLE and Linear Models

When we observe

$$y = \theta + \nu = Xw + \nu$$

and $\nu \sim \mathcal{N}(0, \sigma^2 I_n)$, then the MLE of w (and hence 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 y and w. The basic idea is that we define a link function g such that

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

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

Example: Colored Gaussian Noise

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

The value of $\widehat{w}$ is given by,

$$
\begin{aligned}
\hat{w} &= \arg\min_w -\log p(y|w) \\
&= \arg\min_w (y - Xw)^\top \Sigma^{-1}(y - Xw) \\
&= (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y
\end{aligned}
$$

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 $y = [y_1, \cdots, y_n]^\top$ where $y_i \sim \text{Poisson}(\lambda)$. Given y, find the MLE of the probability that $y \sim \text{Poisson}(\lambda)$ exceeds the mean λ.

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

The MLE of g is

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

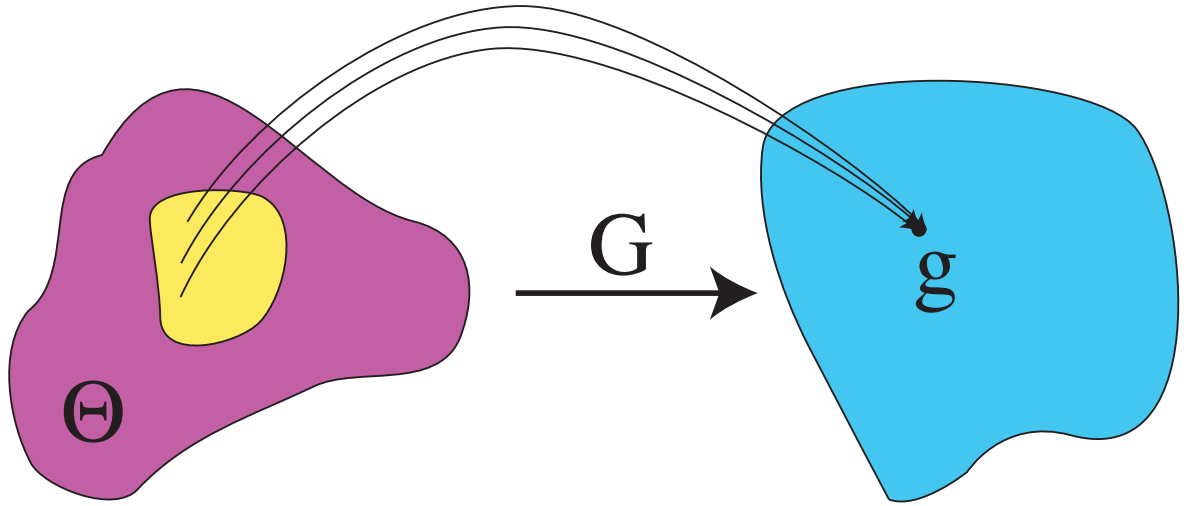
where $\widehat{\lambda}$ is the MLE of λ:

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

Sketch Proof:

Define the "induced" log likelihood function:

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



The MLE of g is

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

Terminology in Estimation Theory

Define

$$\epsilon(\widehat{\theta}) := \widehat{\theta} - \theta$$

Recall that $\widehat{\theta} = \widehat{\theta}(y)$ is a function of data $\Longrightarrow \epsilon(\widehat{\theta})$ is a statistic!
Mean Squared Error:

$$\mathbb{E}[\epsilon^\top \epsilon] := \mathsf{MSE}(\widehat{\theta})$$

Bias:

$$\mathsf{Bias}(\widehat{\theta}) := |\mathbb{E}[\widehat{\theta}] - \theta|$$

Variance / Covariance:

$$\mathsf{Var}(\widehat{\theta}) := \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^\top]$$

Bias-variance decomposition

Key fact:

$$\text{Bias}^2(\widehat{\theta}) + \text{Var}(\widehat{\theta}) = \text{MSE}[\widehat{\theta}]$$

$$\begin{aligned}
\text{Bias}^2(\widehat{\theta}) + \text{Var}(\widehat{\theta}) =& (\theta - E[\widehat{\theta}])^2 + E[\widehat{\theta}^2] - (E[\widehat{\theta}])^2 \\
=& \theta^2 - 2\theta E[\widehat{\theta}] + (E[\widehat{\theta}])^2 + E[\widehat{\theta}^2] - (E[\widehat{\theta}])^2 \\
=& E[\widehat{\theta}^2] - 2\theta E[\widehat{\theta}] + \theta^2 \\
=& E[\widehat{\theta}^2] - 2\theta E[\widehat{\theta}] + E[\theta^2] \\
=& E[\widehat{\theta}^2 - 2\theta\widehat{\theta} + \theta^2] \\
=& E[(\widehat{\theta} - \theta)^2] \\
=& \text{MSE}[\widehat{\theta}]
\end{aligned}$$

Asymptotics

Estimators are often studied as a function of the number of observations:

$$\widehat{\theta}(y) = \widehat{\theta}_n \text{ where } n = \dim y$$

$\widehat{\theta}$ is asymptotically unbiased if

$$\lim_{n\to\infty} \text{Bias}(\widehat{\theta}_n) = 0$$

An estimator is consistent if

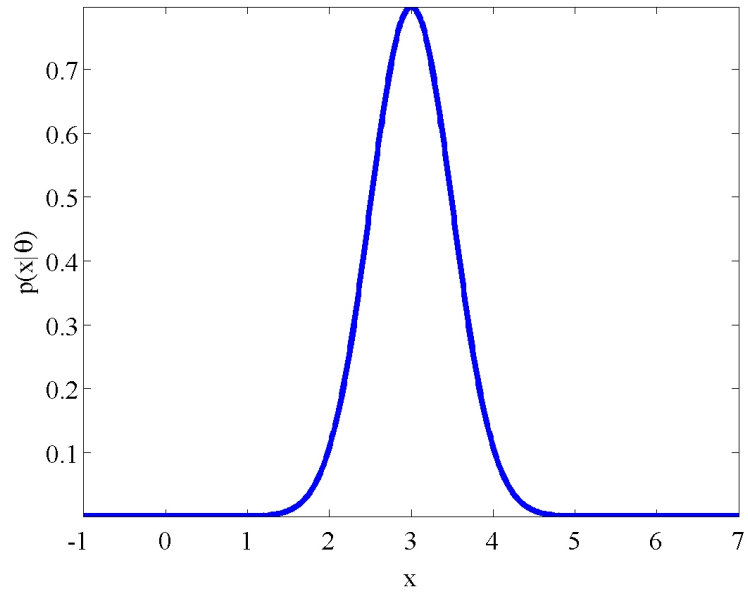
$$\lim_{n\to\infty} \text{MSE}(\widehat{\theta}_n) = 0$$

A consistent estimator is *at least* asymptotically unbiased. Some estimators are unbiased, but inconsistent.

The latter basically means that our estimation does not improve as the number of data increase. Inconsistent estimators can provide reasonable estimates when we have a small number of data. However, consistent estimators are usually favored in practice.

Estimator Accuracy

Consider the likelihood function $p(y|\theta)$ 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:

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

Thus an average measure of curvature is more appropriate.

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

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

Definition: Fisher Information

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

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

Asymptotic Distribution of MLE

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

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

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" $\mathbb{E}\left[\frac{\partial \log p(y|\theta)}{\partial \theta}\right] = 0$ 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

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

Regularity condition

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

$$\mathbb{E}\left[\frac{\partial \log p(y|\theta)}{\partial \theta}\Big|_{\theta=\theta^*}\right] = \int \frac{\partial \log p(y|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} p(y|\theta^*)dx$$
$$= \int \frac{1}{p(y|\theta^*)} \frac{\partial p(y|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} p(y|\theta^*)dx = \int \frac{\partial p(y|\theta)}{\partial \theta}\Big|_{\theta=\theta^*} dx$$
$$= \frac{\partial}{\partial \theta}\left[\int p(y|\theta)dx\right]\Big|_{\theta=\theta^*} = 0,$$

since $\int p(y|\theta)dx = 1$ 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 $p(y|\theta)$. As long as $p(y|\theta)$ 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 y depends on θ (e.g. $y \sim \text{Unif}(0, \theta)$).

Note:

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

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

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

Example:

$$y \sim \mathcal{N}(A \cdot \mathbf{1}_{n\times 1}, \sigma^2 I_n)$$

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

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

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

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

Example: (cont.)

For large n, the CLT 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,

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

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,

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