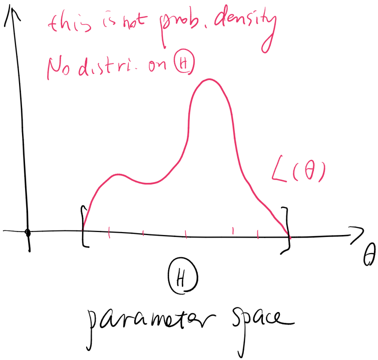
# Maximum Likelihood

Zhentao Shi

The Chinese University of Hong Kong

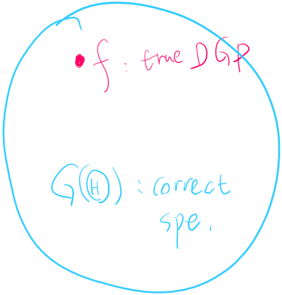
# Likelihood

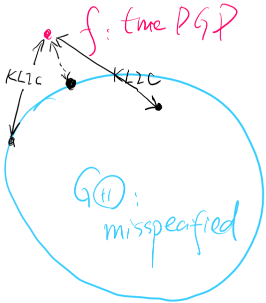
- The most likely outcome
- Distributional assumption



# Model Specification

- Nature: Data $z$ is drawn from a parameter model $f$
- Human: specify a family of models $g\left(z;\theta\right)$ and a parameter space $\Theta$, which span a **model space** $G\left(\Theta\right)=\left\{g\left(z;\theta\right):\theta\in\Theta\right\}$.



# Model and Specification

**Parametric model**. The distribution of the data $\mathbf{Y} = (Y_1, ..., Y_N)$ is known up to a finite dimensional parameter.

- **Semiparametric model**: If we know $Y \sim i.i.d. \left(\mu, \sigma^2\right)$, we can estimate $\mu, \sigma^2$ by method of moments.
- **Parametric model**: If we assume $Y \sim N\left(\mu, \sigma^2\right)$, the model has only two parameters $\mu$ and $\sigma^2$.

# Likelihood Function

- For simplicity, let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be i.i.d.
- The **likelihood** of the sample under a hypothesized value of $\theta \in \Theta$ is

$$L(\theta; \mathbf{Y}) = f(\mathbf{Y}; \theta) = \prod_{i=1}^{N} f(Y_i; \theta)$$

- Two perspectives:
  - (Probabilist) $f(\mathbf{Y}; \theta)$ is a function of $\mathbf{Y}$ given the parameter $\theta$
  - (Statistician) $L(\theta; \mathbf{Y})$ is a function of $\theta$ given the data $\mathbf{Y}$

Section 1

# Correct Specification

# Log-likelihood

- log-likelihood

$$\ell_N(\theta) = \log L(\theta; \mathbf{Y}) = \sum_{i=1}^{N} \log f(Y_i; \theta)$$

  is easier to compute.
- $\log(\cdot)$ is a monotonically increasing function
- The MLE estimator

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell_N(\theta)$$

# Why Maximization: Deep Justification

**Theorem**

*If the model is correctly specified, then $\theta_0$ is the maximizer.*

- Kullback-Leibler information criterion (KLIC):

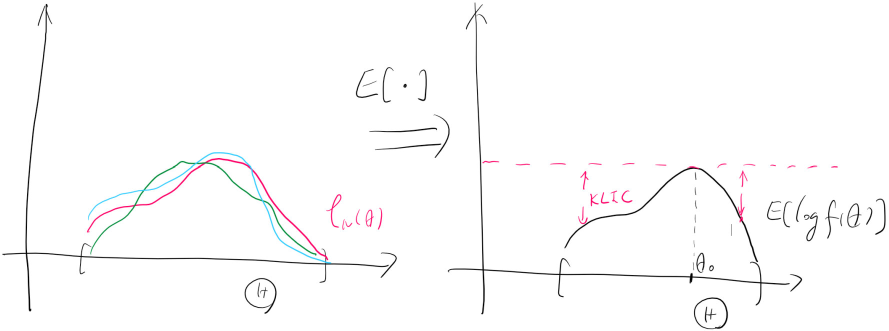
$$KLIC(f,g) = \int f(z) \log \frac{f(z)}{g(z)} dz$$

- $KLIC \geq 0$ because

$$\begin{aligned} & E[\log f(Y;\theta_0)] - E[\log f(Y;\theta)] \\ =& E[\log(f(Y;\theta_0)/f(Y;\theta))] \\ =& -E[\log(f(Y;\theta)/f(Y;\theta_0))] \\ \geq& -\log E[f(Y;\theta)/f(Y;\theta_0)] = 0 \end{aligned}$$

by the Jensen's inequality.

# KLIC



# Score and Hessian

- Score $s_N(\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial\theta} \log f(Y_i;\theta)$ is a function of $\theta$
- Efficient score $s_{i0} = \frac{\partial}{\partial\theta} \log f\left(Y_i;\theta_0\right)$ is evaluated at the true value $\theta_0$

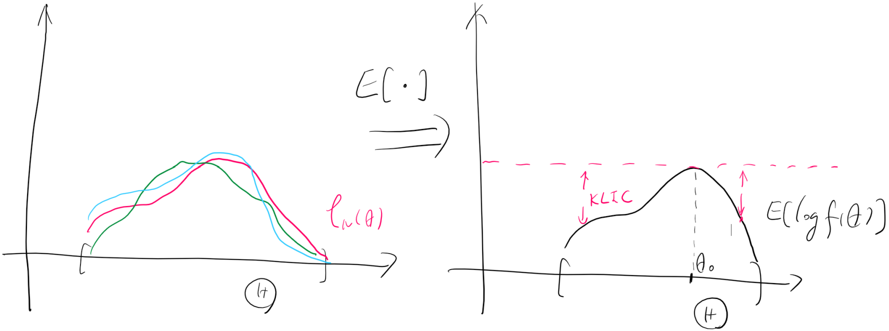
**Theorem**

*If the model is correctly specified, the support of $Y$ does not depend on $\theta$, and $\theta_0$ is in the interior of $\Theta$, then $E\left[s_{i0}\right] = 0$.*

MLE is equivalent to looking for roots of $s_N(\theta) = 0$.

- Hessian: $H_N(\theta) = -\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y_i;\theta)$
- Expected Hessian: $H_0 = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(Y;\theta_0\right)\right]$

# Score and Hessian: Illustration



# Information Equality

- **Fisher Information Matrix**: $I_0 = E[s_{i0}s_{i0}']$

**Theorem**

*If the model is correctly specified, the support of $\Upsilon$ does not depend on $\theta$, and $\theta_0$ is in the interior of $\Theta$, then*

$$I_0 = H_0.$$

- Information equality fails when the model is misspecified

# Cramér-Rao Lower Bound

**Theorem**

*Suppose the model is correctly specified, the support of $\Upsilon$ does not depend on $\theta$, and $\theta_0$ is in the interior of $\Theta$. If $\widetilde{\theta}$ is unbiased estimator, then*

$$var(\widetilde{\theta}) \geq (NI_0)^{-1}.$$

- More general than "BLUE"
- A lower bound for variance of unbiased estimator
- When reached, an estimator is called **Cramér-Rao efficient**.

# Normal Regression

- The normal regression models is

$$Y_i = X_i'\beta + \varepsilon_i$$

- Under the assumption $\varepsilon_i \mid X_i \sim N(0, \gamma)$, the conditional distribution is

$$Y_i \mid X_i \sim N\left(X_i'\beta, \gamma\right).$$

- Parameter $\theta = (\beta, \gamma)$
- The joint likelihood

$$f(Y_i, X_i) = f(Y_i|X_i)f(X_i),$$

where the specification of $f(X_i)$ is irrelevant to $\theta$.

# Asymptotic Normality

- Under regularity conditions, $\hat{\theta} \xrightarrow{p} \theta_0$, and

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, H_0^{-1} I_0 H_0^{-1}\right)$$

- When the information equality holds, we have

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, I_0^{-1}\right),$$

or equivalently

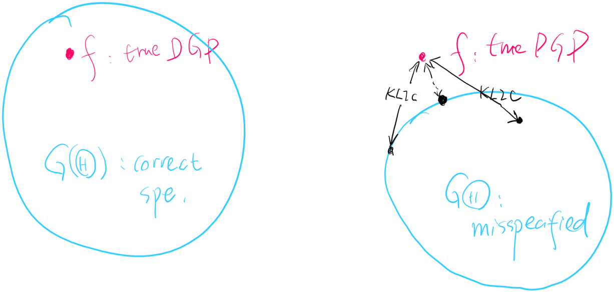
$$\hat{\theta} - \theta_0 \overset{a}{\sim} N\left(0, \frac{I_0^{-1}}{N}\right),$$

- The variance $(NI_0)^{-1}$ is efficient!

Section 2

# Mispecification

# KLIC for Misspecified Models




- If $f \notin G(\Theta)$, the model is misspecified.

$$
\begin{aligned}
KLIC\left(f, g\left(z;\theta\right)\right) &= \int f\left(z\right)\log f\left(z\right)dz - \int f\left(z\right)\log g\left(z;\theta\right)dz \\
&= E\left[\log f\left(z\right)\right] - E\left[\log g\left(z;\theta\right)\right] > 0
\end{aligned}
$$

# Misspecified Model

- Misspecified: $\min_{\theta \in \Theta} KLIC\left(f, g\left(z; \theta\right)\right) > 0$
- MLE is still meaningful
- Pseudo-true parameter:

$$\theta^* = \arg\max_{\theta \in \Theta} E[\ell\left(\theta\right)]$$

  the minimizer of $KLIC\left(f, g\left(z; \theta\right)\right)$ in the parameter space $\Theta$
- Under standard assumption, the MLE estimator $\widehat{\theta} \xrightarrow{p} \theta^*$ and

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} N\left(0, H_*^{-1} I_* H_*^{-1}\right)$$

# Summary

- Parametric models
- Specification of distribution family
- MLE
- Score, Hessian, information matrix
- Misspecification