Section 17

Lecture 6

Section 18

Dynamic regimes

Dynamic regimes

Definition (Dynamic regime)

A dynamic regime $g = (g_0, \ldots, g_k)$, where $g_k : (\bar{a}_{k-1}, \bar{l}_k) \mapsto a_k$, is a policy that assigns treatment (possibly at multiple time points) based on the measured history $(\overline{A}_{k-1}, \overline{L}_k)$.

We will restrict ourselves to settings where

$$g_k : (\bar{l}_k) \mapsto a_k$$

.

Dynamic regime SWIGs

Definition (d-SWIG from Robins and Richardson)

Given a template $\mathcal{G}(a)$ and a dynamic regime g for $\bar{a}$, the d-SWIG $\mathcal{G}(g)$ is defined by applying the following transformation:

- Replace each fixed node a_j with a random node A_j^{g+} that inherits children from a_j. Include dashed directed edges from every variable that is an input to the function g_i that determines the variable A_i^{g+}.
- Each random node V_i that is a descendant of at least one variable A_i^{g+} is relabeled as V_i^g.

Time-varying exposures (treatments) are frequent

Examples:

- Smoking status, which depends on other events in life.
- A therapeutic drug, for which the dose is adjusted according to the response over time (patients take the drug every day, every week etc)
- Cancer screening, which e.g. depends on previous diagnostic tests.
- Surgical interventions (e.g. transplants) are given at a certain time after the diagnosis.
- Expression of genes.

Running example: HIV

Consider a 5-year follow-up study of individuals infected with the human immunodeficiency virus (HIV)[33].

- A_k takes value 1 if the individual receives antiretroviral therapy in month k, and 0 otherwise. Define $A_{-1} = 0$.
- Suppose Y measures health status at 5 years of follow-up.
- So far we have considered *deterministic* treatment rules, for example "always treat", where the outcome of interest is $Y^{a=1}$ vs "never treat", where the outcome of interest is $Y^{a=0}$.
  When $\overline{A} \equiv \overline{A}_K$, we can define 2^K such static regimes...
- However, often we want to make *dynamic* treatment decisions.
- Let $L_k \in \{0, 1\}$ be an indicator of low CD4 cell count measured at month k.
- Depending on the value of L_k, we may argue that it is good or bad to start treatment at time k.

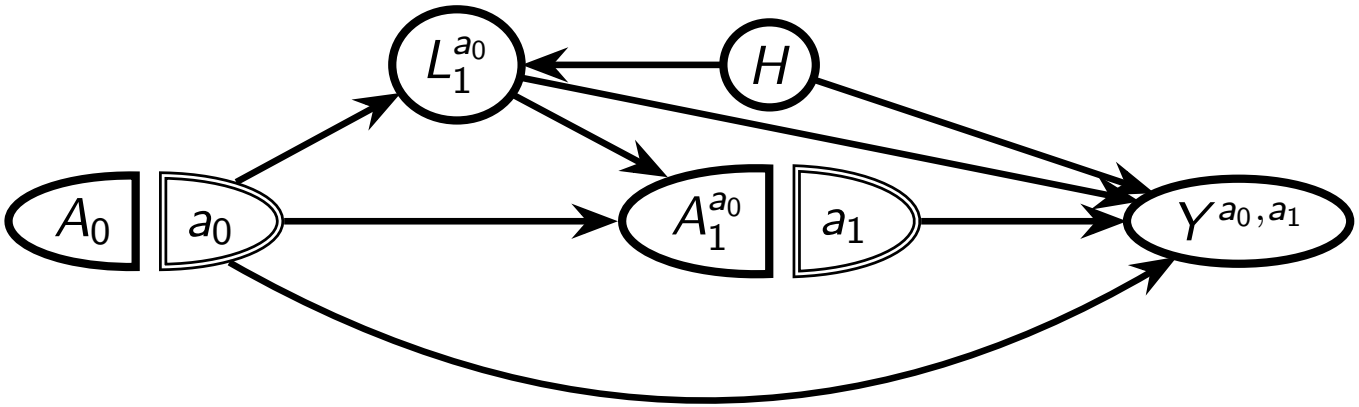
[33] Hernan and Robins, *Causal inference: What if?*

Example of Dynamic Regime

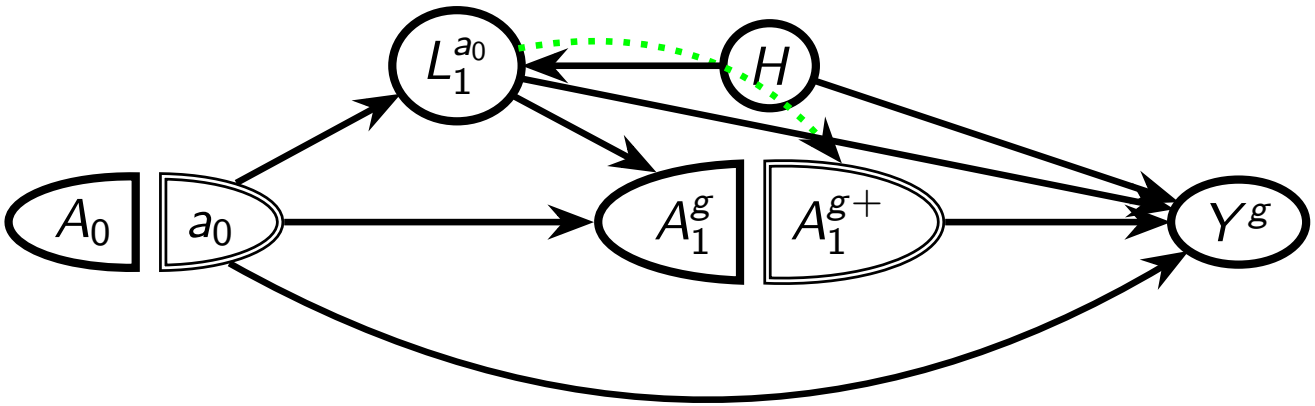
A simple example of a dynamic regime g for setting with two treatments is

- $A_0^{g+} = a_0$.
- $A_1^{g+} = L_1^{a_0}$
  In the HIV example this would mean that you are treated at time 1 if the CD4 cell count is low at that time.

Static vs dynamic



$Y^{a_0,a_1} \perp\!\!\!\perp A_0$ and $Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0$.

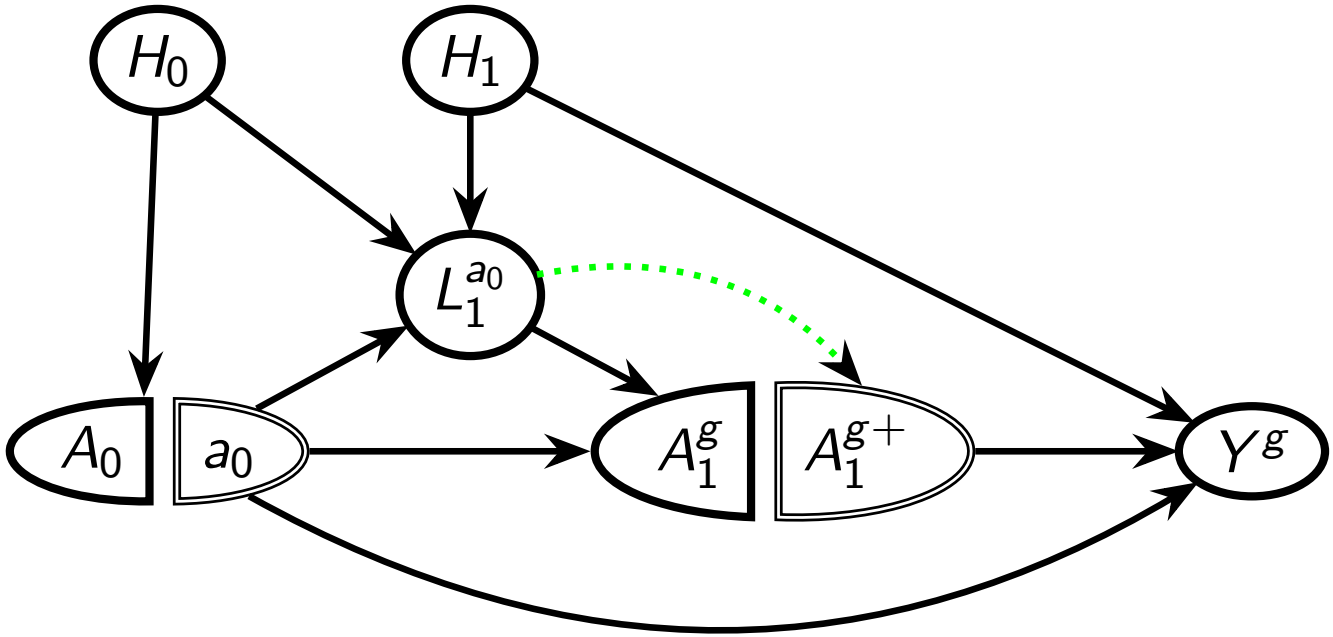


$Y^{g} \perp\!\!\!\perp A_0$ and $Y^{g} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0$.
Consistency gives: $Y^{g} \perp\!\!\!\perp A_0$ and $Y^{g} \perp\!\!\!\perp A_1 \mid L_1, A_0 = a_0$.

Identification results for dynamic regimes

- We can use the same identification conditions (independencies in Slide 163) as for static regimes, only if A_k^{g+} **is not** a function of A_j^{g+} for $j < k$; that is, A_k^{g+} cannot be written a function of only $\overline{L}_k$. However, we need to use the extended g-formula as the identification formula (as defined in Slide 184).
- if A_k^{g+} **is** a function of A_j^{g+} for any $j < k$, we need slightly stronger conditions (we are not presenting them now). This is e.g. the case in the graph in Slide 182 (due to the red arrow).

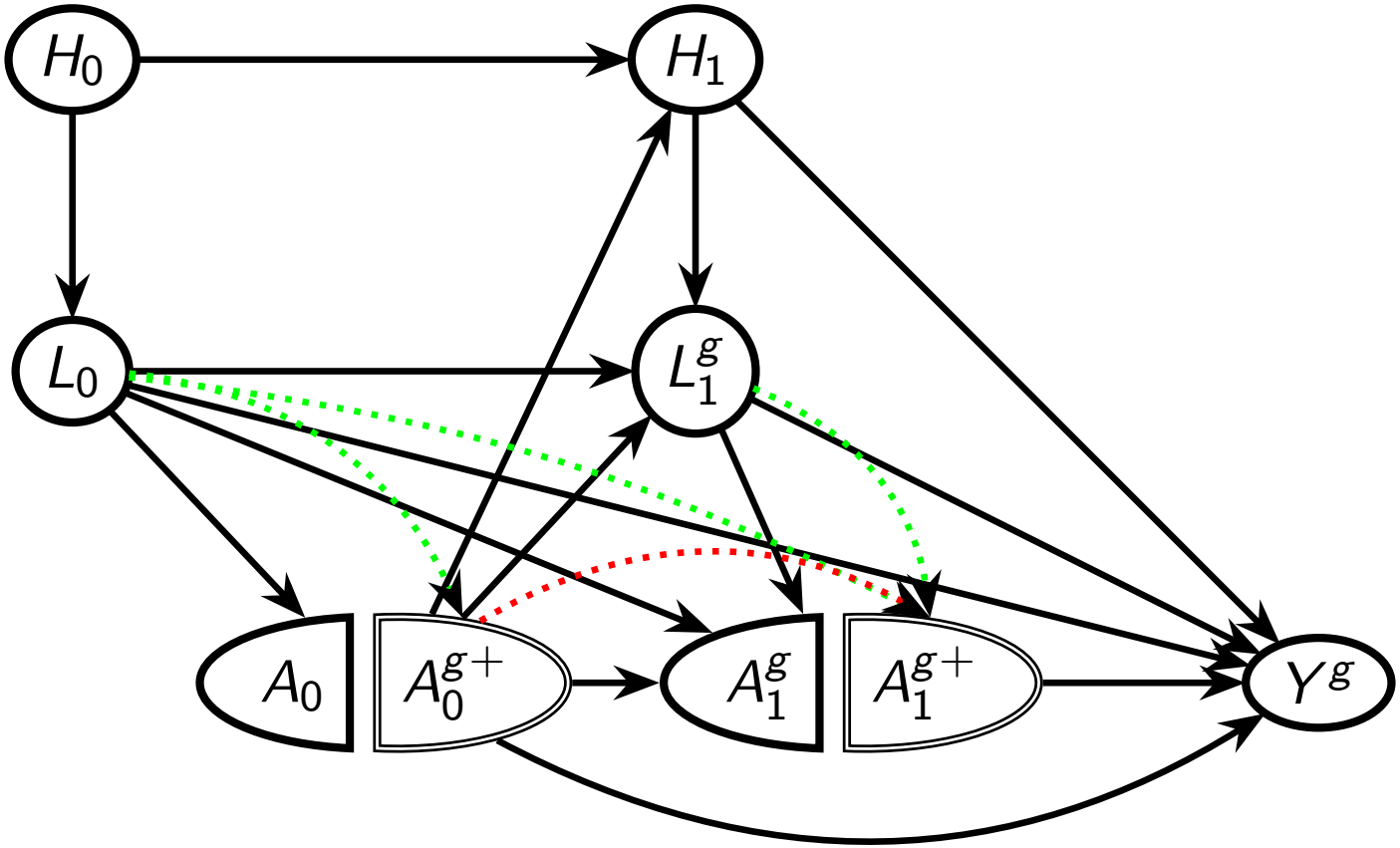
Does the identification conditions hold in the following Dynamic SWIG?



$Y^g \not\perp\!\!\!\perp A_0$ because $A_0 \leftarrow H_0 \rightarrow L_1^{a_0} \rightarrow A_1^{g+} \rightarrow Y^g$ is open. However, we would have identification in a static SWIG where $A_1^{g+} \equiv a_1$. So, in that sense, dynamic regimes require stronger conditions for identification, even though the independencies are stated in the same way.

HIV SWIG

A (busy) graph illustrating a conditional RCT, where H_0 and H_1 are hidden variables (e.g. the actual immune status of the patient).



Consistency for dynamic regimes

Now we state a more general consistency con dition, which is valid for time-varying dynamic regimes. Indeed, it can simply be expressed as

$$\text{if } \overline{A}_K = \overline{A}_K^{g+}, \text{ then } Y^g = Y.$$

A special case for static regimes is: if $\overline{A}_K = \overline{a}_K$, then $Y^{\overline{a}_K} = Y$.

Marginal extended g-formula under interventions that depend on $\overline{L}_k$

Suppose that g_k is only a function of $\overline{L}_k$. Then, the marginal extended g-formula is defined as the following function of observed random variables $\overline{L}_K, \overline{A}_K, Y$, topologically ordered $L_0, A_0, \ldots, L_K, A_K, Y$.

Definition (Marginal extended g-formula)

$$b_g(y) = \sum_{\bar{a}_K} \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) p^g(a_j \mid \bar{l}_j),$$

where $\bar{l}_k = (l_0, \ldots, l_k)$, $k \leq K$, are instantiations of **observed** variables and $p^g(a_j \mid \bar{l}_j)$ is the density of A_k^{g+} given $\overline{L}_k^g$, which is determined by g_k.

We let variables indexed by subscript -1, e.g. L_{-1} be empty.
Note that $p^g(a_k \mid \bar{l}_k)$ is a known function. It is determined by the investigator (even if it has a superscript g). If g_k is a deterministic function of $\bar{l}_k$, then

$$p^g(a'_k \mid \bar{l}_k) = \begin{cases} 1 \text{ if } a'_k = g_k(\bar{l}_k), \\ 0 \text{ if } a'_k \neq g_k(\bar{l}_k), \ k \in \{0, \ldots, K\}. \end{cases}$$

Relation to the g-formula for static regimes

The dynamic extended g-formula density generalizes the marginal g-formula from slide 163, because for a static intervention that sets $\bar{a} = (a_0, \ldots, a_K)$ we have that for $k \in \{0, \ldots, K\}$,

$$p^g(a'_k \mid \bar{l}_k) = \begin{cases} 1 \text{ if } a'_k = a_k, \\ 0 \text{ if } a'_k \neq a_k. \end{cases}$$

SWIG criterion to identify effects of dynamic regimes (you do not need to understand the extended g-formula density)

Definition (extended g-formula density)

The *dynamic extended g-formula density* for $Y \equiv Y_K$ under treatment regime g given by the functions $g_0, \ldots, g_K$ that determine $\overline{A}_K = (A_0, \ldots, A_K)$ is

$$f^g(y, \bar{l}_K, \bar{a}_K, \bar{a}_K^+) = p(y \mid \bar{l}_K, \bar{a}_K^+) \prod_{j=0}^{K} p(l_j, a_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}^+) \prod_{t=0}^{K} p^g(a_t^+ \mid pa_{A_t^{g+}}),$$

where $\bar{l}_k = (l_0, \ldots, l_k)$, $k \leq K$, are **observed** variables, $p^g(a_t^+ \mid pa_{A_t^{g+}})$ is the density of A_t^{g+} given $PA_{A_t^{g+}}$ is the input to g_t, for $t \in \{0, K\}$.

James M Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512; Thomas S Richardson and James M Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality". In: *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128.30 (2013).

Section 19

Estimation

Plan for today

- Review foundations of estimation theory that are relevant to causal inference.
  - Statistical models (Parametric and non-parametric).
  - Correctly specified models.
- Motivate why we need to study certain estimation problems.
  - Convergence of conditional means.
- Introduce some commonly used estimators: Regression estimators and inverse probability weighted estimators.
  - Brief summary of linear models.
  - Logistic regression models.
  - M-estimators.
  - Link this back to counterfactuals.

Estimation in causal inference settings (informal motivation)

- An identification formula motivates estimators.
- Estimation in causal inference settings is, in principle, identical to the "inverse" problem you have studied in previous statistics classes.
- However, the functionals we are estimating are sometimes unusual, and therefore we sometimes need new estimators. Indeed, a lot of identification results in causal inference have motivated new estimation theory.
- Broadly speaking, causal inference researchers are concerned about bias.
  - After doing the hard work of deriving an identification formula, we do not want to induce bias in the estimation step either.
- I remind you about how we divide the causal inference into different tasks: (i) Define your question of interest (estimand), (ii) Evaluate whether the estimand is identified, (iii) if the estimand is identified, we proceed with estimation.

What is bias

- *Systematic bias:* We say there is *systematic bias* if the causal estimand of interest is not identified.
  Informally, any structural association between the treatment and the outcome that does not arise from the causal effect of treatment on the outcome.
- *Bias due to model misspecification:* When we use a statistical model that is misspecified (I give a formal definition of model mis-specification in a later slide).

Estimation vs. identification

- We have considered identification assumptions that are necessary even if we had an infinite amount of data.
- The statistical modeling assumption we consider now are invoked because we do not have infinite amount of data.

PS: In this course we will mainly consider frequentist inference: probability is defined as a limiting frequency. An alternative is Bayesian inference,[34] which defines probability as a degree of belief.

[34] Again, this is not the same as a Bayesian network

Motivation for regression modelling and the curse of dimensionality

Definition (Statistical model)

A statistical model $\mathcal{P}$ is a collection of laws, $\mathcal{P} = \{P_\eta : \eta \in \Gamma\}$.

Definition (Parametric statistical model)

A statistical model $\mathcal{P}$ is parametric $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$ for a positive integer k.

So far we have been non-parametric: we have not restricted ourselves to parametric models. This is arguably desirable, because then we *do not* impose parametric restrictions on the data generating mechanism.

Motivation: Simple mean estimation

- Suppose we are interested in estimating a parameter, say, $h(L, A, Y)$ from an observed sample of n observations, (L_i, A_i, Y_i), $i = 1, \ldots, n$.
- Suppose we would like to ignore the assumptions encoded in our model $\mathcal{P}$ when we study $h(L, A, Y)$; more precisely, we will only use the fact that we have draws from i.i.d. individuals where $\mathbb{E}(Y) = \mu$ and that Y is continuous with finite variance $\sigma^2 < \infty$.
- Our statistical model is non-parametric; $\mathcal{P} = \{P(Y = y) : \int y^2 f(y) dy < \infty\}$. For $h(L, A, Y) \equiv \mathbb{E}(Y)$, we would simply do the empirical mean (sample mean) $\mathbb{E}_n(Y) = \frac{1}{n} \sum_{i=1}^{n} Y_i$. By the weak law of large numbers (WLLN),

$$\lim_{n \to \infty} P(|\mathbb{E}_n(Y) - \mu| > \epsilon) = 0.$$

  So the estimator is consistent. Indeed, the estimator is $\sqrt{n}$-consistent, and by the CLT $\sqrt{n}(\mathbb{E}_n(Y) - \mu) \sim \mathcal{N}(0, \sigma^2)$.
- Because $\mathbb{E}_n(Y)$ has variance σ^2/n, which is $O_P(1/n)$, then $\sqrt{n}(\mathbb{E}_n(Y) - \mu)$ has variance σ^2 which is $O_P(1)$, i.e. "bounded in probability" or "uniformly tight": A sequence $\{Q_n\}$ is uniformly tight if for all $\epsilon > 0$ there exists an M s.t. $\sup_n P(|Q_n| > M) < \epsilon$.

Motivation continues

- Now, suppose L is continuous and our parameter of interest is the conditional mean $h(L, A, Y) \equiv \mathbb{E}(Y \mid L)$.
- In particular, to estimate $\mathbb{E}(Y \mid L = l)$ there exists at most one individual l with $L_i = l$ and $\mathbb{E}_n(Y \mid L = l) = Y_i$, regardless of n, and clearly we do not have $\sqrt{n}$-consistency.
- Thus, we have to do something else...

Parameteric modelling

- Can we really say that the distribution that generated the data belongs to a parametric model?
- The answer is **no** in most settings. Therefore many argue that non-parametric methods are more desirable.
- However, it is often argued that studying parametric models is useful (i) because they can be good approximations, (ii) sometimes we have knowledge about the data generating mechanism and (iii) they provide the background for understanding non-parametric methods.

Reminder: Maximum Likelihood Estimation (MLE)

Consider a vector $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^\mathsf{T}$ of parameters that indexes the distribution $\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$, where Θ is a parameter space.
We evaluate the observed data sample $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, which gives us the likelihood,

$$L_n(\theta) = L_n(\theta; \boldsymbol{Y}) = f_n(\boldsymbol{Y}; \theta),$$

where $f_n(\boldsymbol{Y}; \theta)$ is a product of n density functions evaluated at $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$.
MLE maximises the likelihood, i.e.

$$\theta = \arg\max_{\theta \in \Theta} L_n(\theta\,;\boldsymbol{Y}).$$

The logarithm is a monotone function, and thus it is more convenient to maximise the log-likelihood: $\ell(\theta\,;\boldsymbol{Y}) = \log L_n(\theta\,;\boldsymbol{Y})$. If $\ell(\theta\,;\boldsymbol{Y})$ is differentiable in θ, we solve $M(\boldsymbol{Y};\theta) = \frac{\delta \ell(\theta\,;\boldsymbol{Y})}{\delta\theta}$, , i.e. the score equations (also called likelihood equations)

$$p_1 \equiv \frac{\partial \ell}{\partial \theta_1} = 0, \quad \frac{\partial \ell}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \ell}{\partial \theta_k} = 0.$$

.

Likelihood continues

We need local concavity. Thus, the Hessian matrix

$$\mathbf{H}\left(\widehat{\theta}\right) = \begin{bmatrix} \left.\frac{\partial^2 \ell}{\partial \theta_1^2}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_1 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_2^2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_2 \, \partial \theta_k}\right|_{\theta=\widehat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_1}\right|_{\theta=\widehat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_k \, \partial \theta_2}\right|_{\theta=\widehat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_k^2}\right|_{\theta=\widehat{\theta}} \end{bmatrix},$$

is negative semi-definite at $\widehat{\theta}$. The Fisher information matrix is defined as $\mathcal{I}(\theta) = \mathbb{E}\left[\mathbf{H}\left(\widehat{\theta}\right)\right]$.