Semiparametric Inference

1 Introduction

1.1 Infinite-Dimensional Space

We can roughly regard the definition of **models** as a class or probability distributions/law.[1] For instance, the class of densities can be described as

$$\mathscr{P} = \{p(z, \theta), \theta \in \Omega \subset \mathbb{R}^p\}$$

where the dimension p is some finite positive integer.

The prominent difference between parametric and semiparametric inference is the inclusion of an infinite-dimensional component. The parameter spaces that we will consider later will be subsets of linear **vector spaces**, i.e. $\Omega \subset S$

A space $\mathscr{S}$ is a linear space if, for θ_1 and θ_2 is in S, $a\theta_1 + b\theta_2$ will also be an element of S for any scalar constants a and b. Now, we illustrate a linear space S that's **infinite-dimensional**.

Consider $S = C(\mathbb{R})$, the space for all continuous functions f. We want to show that it can't be spanned by any finite set of elements in S. This can be accomplished by noting that S contains the linear subspaces made up of the class of polynomials of order m, that is $S_m = \{f(x) = \sum_{j=0}^{m} a_j x^j\}$ for all constants a_j. $dim(S_m) = m + 1$ because $\{1, \ldots, x^m\}$ is a basis of the subspace. Suppose x^j can be written as a linear combination of $\{1, \ldots, x^{j-1}\}$ for any $\jmath = 1, 2, \ldots$. If it could, then

$$x^j = \sum_{l=0}^{j-1} a_l x^l, \text{for all } x \in \mathbb{R}$$

for some constants $a_0, \ldots, a_{j-1}$. Notice that the j-th order derivative of $x^j = j!$, but the j-th order derivative of the r.h.s is 0. So here comes the contradiction, which means that the space S cannot be spanned by any finite number of elements of S. If it's possible, the space of polynomials of order greater than m could also be spanned by the m elements. Hence, S is infinite-dimensional.

Now we look at a familiar semiparametric model, the **restricted moment model**. Consider a family of probability distributions for $Z = (Y, X)$ that satisfy the regression relationship

$$E(Y|X) = \mu(X, \beta)$$

where $\mu(X, \beta)$ is a known function of X and the unknown q-dimensional parameter β. We can easily write the model as

$$Y = \mu(X, \beta) + \epsilon$$

where $E(\epsilon|X) = 0$. The data are realizations of $(Y_1, X_1), \ldots, (Y_n, X_n)$ that are iid with density $p_{Y,X}\{y, x, \beta, \eta(\cdot)\}$ where $\eta(\cdot)$ denotes the infinite dimensional nuisance parameter function characterizing the joint distribution of η and X. Knowledge of β and the joint distribution of (η, X) will induce the joint distribution of (Y, X). Since $\epsilon = Y - \mu(X, \beta)$ and $p_{Y,X}(y, x) = p_{\epsilon,X}\{y - \mu(x, \beta), x\}$.

The restricted moment model only makes the assumption that $E(\epsilon|X) = 0$. That is we will allow any joint density $p_{,X}(\epsilon, x) = p_{\epsilon|X}(\epsilon|x)p_X(x)$ such that

$$\begin{aligned}
&p_{\epsilon|X}(\epsilon|x) \geq 0, \text{ for all } \epsilon, x,\\
&\int p_{\epsilon|X}(\epsilon|x)d\epsilon = 1 \text{ for all } x,\\
&\int \epsilon p_{\epsilon|X}(\epsilon|x)d\epsilon = 0 \text{ for all x},\\
&p_X(x) \geq 0 \text{ for all x},\\
&\int p_X(x)dv_X(x) = 1
\end{aligned}$$

The class of conditional densities for ϵ given X, such that $E(\epsilon|X) = 0$, can be constructed as follows:

Step 1: Choose any arbitrary positive function of ϵ and x (subject to regularity conditions):

$$h^{(0)}(\epsilon, x) > 0$$

Step 2: Normalize this function to be a conditional density:

$$h^{(1)}(\epsilon, x) = \frac{h^{(0)}(\epsilon, x)}{\int h^{(0)}(\epsilon, x)d\epsilon}$$

Step 3: Center itA random variable ϵ^* whose conditional density given $X = x$ is $h^{(1)}(\epsilon', x) = p(\epsilon^* = \epsilon'|X = x)$, has mean

$$\mu(x) = \int \epsilon' h^{(1)}(\epsilon', x)d\epsilon'$$

We consider $\epsilon = \epsilon^* - \mu(X)$ or $\epsilon = \epsilon^* + \mu(X)$. Since the transformation from ϵ to ϵ^* has Jacobian equal to 1, the conditional density of ϵ given X is given by

$$\eta_1(\epsilon, x) = h^{(1)}\left(\epsilon + \int \epsilon h^{(1)}(\epsilon, x)d\epsilon, x\right)$$

which satisfies $E(\epsilon|X) = 0$ by construction. The class of all such conditional densities $\eta_1(\epsilon, x)$ was derived from arbitrary positive functions $h^{(0)}(\epsilon, x)$ and the space of positive functions is infinite-dimensional, then the set of such conditional densities is also infinite-dimensional. Similar ly we can construct densities for X where $p_X(x) = \eta_2(x)$ such that

$$\eta_2(x) > 0, \int \eta_2(x)dv_X(x) = 1$$

Therefore, the restricted moment model is characterized by $\{\beta, \eta_1(\epsilon, x), \eta_2(x)\}$ where $\beta \in \mathbb{R}^q$ is finite-dimensional and η_1, η_2 are infinite-dimensional. Consequently, the joint density of (Y, X) is given by

$$p_{Y,X}\{y, x; \beta, \eta_1(\cdot), \eta_2(\cdot)\} = p_{Y|X}\{y|x; \beta, \eta_1(\cdot)\}p_X\{x; \eta_2(\cdot)\} = \eta_1\{y - \mu(x, \beta), x\}\eta_2(x)$$

Definition 1. Gateaux differential
Let $f : V \to U$ be a function and let $h \neq 0$ and x be vectors in V. The Gateaux differential $d_h f$ is defined [2]

$$d_h f = \lim_{\epsilon \to 0} \frac{f(x + \epsilon h) - f(x)}{\epsilon}$$

Notice that the Gateaux differential is a generalization of the directional derivative. We also have Frechet derivative as a stronger version of derivative. Consider the following example,

Let $J : H^1(\Omega) \to \mathbb{R}$ be

$$J(u) = \int_\Omega \left(\frac{1}{2} u_x^2 + \frac{1}{2} u^2 \right) dx$$

Then

$$\begin{aligned}
d_h J &= \lim_{\epsilon \to 0} \frac{\int_\Omega \left[\frac{1}{2}(u_x + \epsilon h_x)^2 + \frac{1}{2}(u + \epsilon h)^2 - \frac{1}{2}u_x^2 - \frac{1}{2}u^2 \right] dx}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\int_\Omega \left[\frac{1}{2}u_x^2 + \frac{1}{2}u^2 + \epsilon u_x h_x + \frac{1}{2}\epsilon^2 h_x^2 + u\epsilon h + \frac{1}{2}\epsilon^2 h^2 - \frac{1}{2}u_x^2 - \frac{1}{2}u^2 \right] dx}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\int_\Omega \left[\epsilon u_x h_x + \frac{1}{2}\epsilon^2 h_x^2 + u\epsilon h + \frac{1}{2}\epsilon^2 h^2 \right] dx}{\epsilon} \\
&= \lim_{\epsilon \to 0} \int_\Omega \left[u_x h_x + \frac{1}{2}\epsilon h_x^2 + uh + \frac{1}{2}\epsilon h^2 \right] dx \\
&= \int_\Omega \left[u_x h_x + uh \right] dx
\end{aligned}$$

1.2 Hilbert Space

Suppose $Z_i \overset{i.i.d}{\sim} Z$. Z denotes a random vector for a single observation. The underlying probability space is (Ω, A, P). Consider the space consisting of q-dimensional mean zero random functions of Z,

$$h : \Omega \to \mathbb{R}^q$$

where $h(Z)$ is measurable and satisfies $E[h(Z)] = 0$ and $E[h^T(Z)h(Z)] < \infty$. Clearly the space is linear and $h(Z) = 0^{q \times 1}$ is the origin of the space.

1.2.1 Dimension of the Space of Random Functions

An element of the linear space defined above is a q-dimensional function of Z. **This should not be confused with the dimensionality of the space itself**. First we consider the space of one-dimensional random functions of Z, where Z is a discrete variable with finite support. $P(Z = z_i) = \pi_i$ for all $i \in \{1, \dots, k\}$. Any one-dimensional random function of Z can be defined as $h(Z) = a_1 I(Z = z_1) + \cdots + a_k I(Z = z_k)$ for any real-valued constants a_j. **The space of all such random functions is a linear space spanned by the k linearly independent functions** $I(Z = z_i)$.

If we assume $E[h(Z)] = 0$, that means $a_k = -(\sum_{i=1}^{k-1} a_i\pi_i)/\pi_k$. The space of one-dimensional mean-zero random functions of Z is a linear space spanned by the $k-1$ linearly independent functions $\{I(Z = z_i) - \frac{\pi_i}{\pi_k} I(Z = z_k)\}$

1.2.2 Projection Theorem

Let H be a Hilbert space and U a linear subspace that is closed. For all $h \in H$, there exists a unique $u_0 \in U$ that is closest to h;

$$||h - u_0|| \leq ||h - u|| \text{ for all } \; u \in U$$

Furthermore, $h - u_0$ is orthogonal to U, that is

$$< h - u_0, u >= 0 \text{ for all } \; u \in U$$

We denote u_0 to be $\prod(h|U)$

1.2.3 Example: q-dimensional Random Functions

Let H be a Hilbert space of mean zero q-dimensional measurable random functions with finite second moments equipped with the inner product

$$< h_1, h_2 >= E(h_1^T h_2)$$

Let $v(Z)$ be an r-dimensional random function with mean zero and $E(v^T v) < \infty$. Consider the linear subspace U spanned by $v(Z)$ that is,

$$U = \{B^{q\times r}v, \text{where B is any arbitrary q times r matrix of real numbers}\}$$

The linear subspace U defined above is a finite-dimensional linear subspace contained in the infinite-dimensional Hilbert space H. If the elements $v_1(Z), \ldots, v_r(Z)$ are linearly independent, then the dimension of U is $q \times r$ This can be seen as $U = \{Bv | B \in L(\mathbb{R}^r, \mathbb{R}^q)\}$ (Refer to Axler's LADR). We can easily get the projection of h onto U by noticing that $E[(h - B_0v)^T Bv] = 0$ for all $B \in \mathbb{R}^{q\times r}$.

2 Influence Function

Consider the statistical model where $Z_i \overset{i.i.d}{\sim} Z$. The density of Z belongs to the class $\{p_Z(z, \theta), \theta \in \Omega\}$ with respect to some dominating measure v_Z.

An estimator $\hat{\beta}_n$ of β is a q-dimensional measurable random function of $Z_1, \ldots, Z_n$. Most reasonable estimators for β are asymptotically linear, that is, there exists a random vector $\varphi^{q\times 1}(Z)$, such that $E[\varphi(Z)] = 0^{q\times 1}$,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1)$$

Notice that $E[\varphi(Z)]$ is shorthand for $E_{\theta_0}[\varphi(Z, \theta_0)]$

The random vector $\varphi(Z_i)$ is referred to as the i-th influence function of the estimator $\hat{\beta}_n$.

2.0.1 Example

Consider $Z_i \overset{i.i.d}{\sim} N(\mu, \sigma^2)$. The MLE of μ nad σ^2 are given by $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (Z_i - \hat{\mu}_n)^2$. The estimator $\hat{\mu}_n$ is asymptotically linear follows immediately because

$$n^{1/2}(\hat{\mu}_n - \mu_0) = n^{-1/2}\sum_{i=1}^n (Z_i - \mu_0)$$

Therefore, $\hat{\mu}_n$ for μ is an asymptotically linear estimator for μ whose i-th influence function is given by $(Z_i) = (Z_i - \mu_0)$. Similarly, we can show that

$$\hat{\sigma}_n^2 - \sigma_0^2 = n^{-1}\sum_{i=1}^n [(Z_i - \mu_0)^2 - \sigma_0^2] + (\hat{\mu}_n = \mu_0)^2$$

which implies that

$$n^{1/2}(\hat{\sigma}_n^2 - \sigma_0^2) = n^{-1/2}\sum_{i=1}^n [(Z_i - \mu_0)^2 - \sigma_0^2] + n^{1/2}(\hat{\mu}_n = \mu_0)^2 = n^{-1/2}\sum_{i=1}^n [(Z_i - \mu_0)^2 - \sigma_0^2] + o_p(1)$$

Therefore, $\hat{\sigma}_n^2$ is an asymptotically linear estimator for σ^2 whose i-th influence function is given by $\varphi(Z_i) = (Z_i - \mu_0)^2 - \sigma_0^2$

Recall that,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2}\sum_{i=1}^n \varphi(Z_i) + o_p(1)$$

then by the CLT,

$$n^{-1/2}\sum_{i=1}^n \varphi(Z_i) \xrightarrow{d} N(0^{q\times 1}, E(\varphi\varphi^T))$$

and, by Slutsky's theorem

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, E(\varphi\varphi^T))$$

An asymptotically linear estimator has a unique (a.s.) influence function.

Proof. Suppose not. Then there exists another influence function $\varphi^*(Z)$ such that

$$E[\varphi^*(Z)] = 0,$$

and

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2}\sum_{i=1}^n \varphi^*(Z_i) + o_p(1)$$

Since $n^{1/2}(\hat{\beta}_n - \beta_0)$ is also equal to $n^{-1/2}\sum_{i=1}^n \varphi^*(Z_i) + o_p(1)$, which implies that

$$n^{-1/2}\sum_{i=1}^n \{\varphi(Z_i) - \varphi^*(Z_i)\} = o_p(1)$$

However, by the CLT,

$$n^{-1/2}\sum_{i=1}^{n}\{\varphi(Z_i)-\varphi^*(Z_i)\} \xrightarrow{d} N(0, E[(\varphi-\varphi^*)(\varphi-\varphi^*)^T])$$

In order, for the limiting normal distribution to be $p_p(1)$, we must have,

$$E[(\varphi-\varphi^*)(\varphi-\varphi^*)^T] = 0^{q\times q}$$

which implies that $\varphi(Z)=\varphi^*(Z)$ a.s.

2.1 Super-Efficiency

Recall the Hodges example in Theoretical Statistics, that some estimators can achieve an asymptotic variance lower than the CRLB. These estimators are unnatural and have **undesirable local properties** associated with them. Therefore, in order to avoid problems associated with super-efficient estimators, we will impose some **additional regularity conditions on the class of estimators that will exclude such estimators**. Specifically, we will require that an estimator be regular, as we now define.[3]

Definition 2. Consider a local data generating process (LGDP), where, for each n, the data are distributed according to 'θ_n', where $n^{1/2}(\theta_n-\theta^*)$ converges to a constant. That is

$$Z_{1n},\dots,Z_{nn} \overset{i.i.d}{\sim} p(z,\theta_n)$$

where

$$\theta_n = (\beta_n^T, \eta_n^T)^T, \;\; \theta^* = ({\beta^*}^T, {\eta^*}^T)^T$$

An estimator $\hat{\beta}_n$, more specifically $\hat{\beta}_n(Z_{1n},\dots,Z_{nn})$ is said to be regular if, for each θ^*, $n^{1/2}(\hat{\beta}_n-\beta_n)$ has a limiting distribution that does not depend on the LDGP. For our purposes, this will ordinarily mean that if

$$n^{1/2}\{\hat{\beta}_n(Z_{1n},\dots,Z_{nn})-\beta^*\} \overset{D(\theta^*)}{\to} N(0,\Sigma^*)$$

where

$$Z_{1n},\dots,Z_{nn} \;\text{ are iid }\; p(z,\theta^*), \;\text{ for all n,}$$

then

$$n^{1/2}\{\hat{\beta}_n(Z_{1n},\dots,Z_{nn})-\beta_n\} \overset{D(\theta_n)}{\to} N(0,\Sigma^*),$$

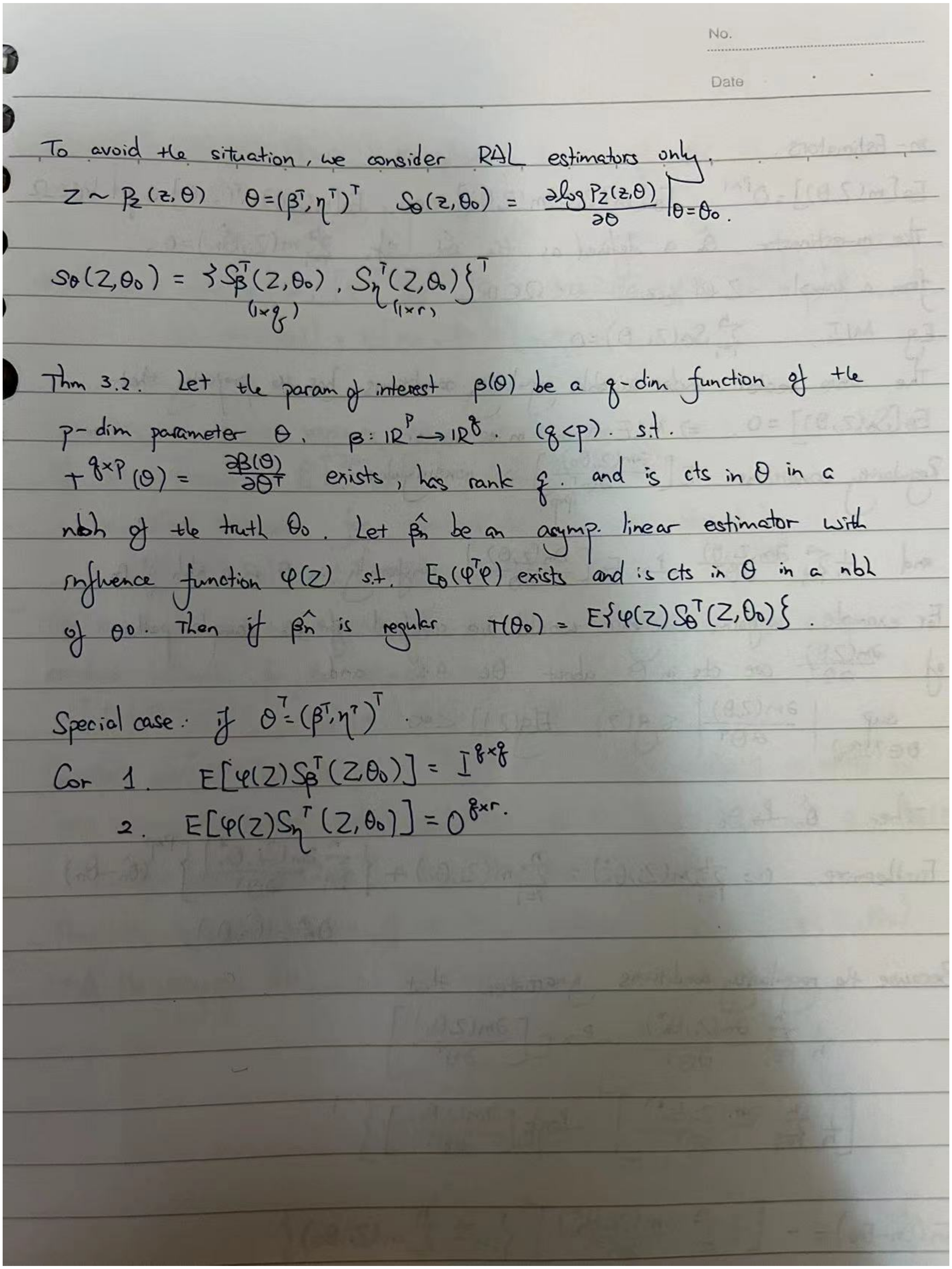
where

$$Z_{1n},\dots,Z_{nn} \;\text{ are iid }\; p(z,\theta_n), \;\text{ for all n,}$$

and $n^{1/2}(\theta_n-\theta^*)\to\tau^{p\times 1}$, where τ is any arbitrary constant vector.

We can show that the Hodes estimator is not regular.

2.2 Important theorem



No.

Date

To avoid the situation, we consider RAL estimators only.

$Z \sim P_Z(z,\theta)$ $\quad \theta = (\beta^T, \eta^T)^T$ $\quad S_\theta(z,\theta_0) = \frac{\partial \log P_Z(z,\theta)}{\partial \theta}\Big|_{\theta=\theta_0}$.

$$S_\theta(Z,\theta_0) = \{\underbrace{S_\beta^T(Z,\theta_0)}_{(1\times q)}, \underbrace{S_\eta^T(Z,\theta_0)}_{(1\times r)}\}^T$$

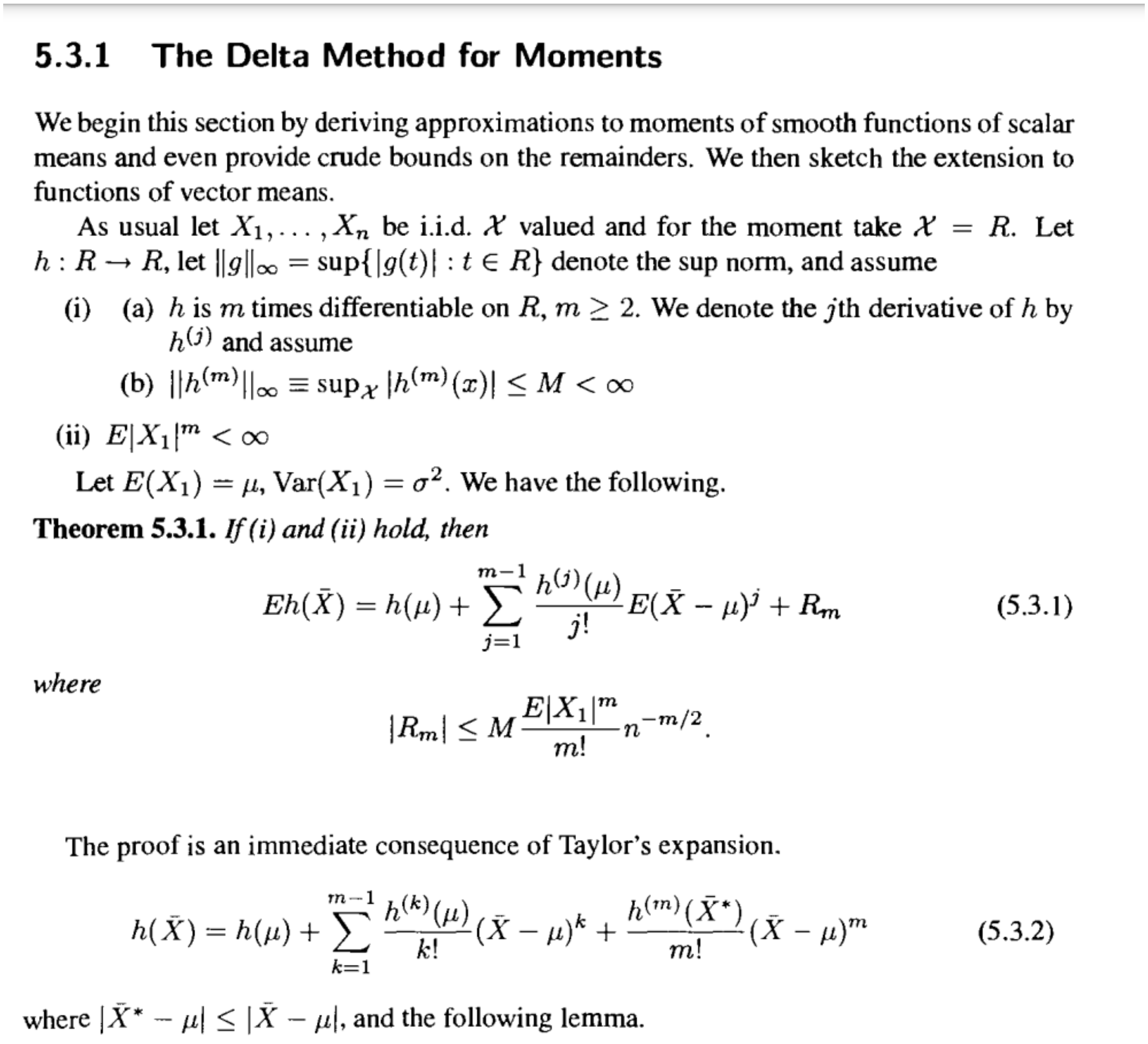
Thm 3.2. Let the param of interest $\beta(\theta)$ be a q-dim function of the p-dim parameter θ. $\beta: \mathbb{R}^p \to \mathbb{R}^q$. $(q<p)$. s.t. $\Gamma^{q\times p}(\theta) = \frac{\partial \beta(\theta)}{\partial \theta^T}$ exists, has rank q. and is cts in θ in a nbh of the truth θ_0. Let $\hat{\beta_n}$ be an asymp. linear estimator with influence function $\varphi(Z)$ s.t. $E_\theta(\varphi^T\varphi)$ exists and is cts in θ in a nbh of θ_0. Then if $\hat{\beta_n}$ is regular $\Gamma(\theta_0) = E\{\varphi(Z) S_\theta^T(Z,\theta_0)\}$.

Special case: if $\theta^T = (\beta^T, \eta^T)^T$.

Cor 1. $E[\varphi(Z) S_\beta^T(Z\theta_0)] = I^{q\times q}$

2. $E[\varphi(Z) S_\eta^T(Z,\theta_0)] = 0^{q\times r}$.

2.3 Delta Method



5.3.1 The Delta Method for Moments

We begin this section by deriving approximations to moments of smooth functions of scalar means and even provide crude bounds on the remainders. We then sketch the extension to functions of vector means.

As usual let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{X}$ valued and for the moment take $\mathcal{X} = R$. Let $h : R \to R$, let $\|g\|_\infty = \sup\{|g(t)| : t \in R\}$ denote the sup norm, and assume

(i) (a) h is m times differentiable on R, $m \geq 2$. We denote the jth derivative of h by $h^{(j)}$ and assume

(b) $\|h^{(m)}\|_\infty \equiv \sup_{\mathcal{X}} |h^{(m)}(x)| \leq M < \infty$

(ii) $E|X_1|^m < \infty$

Let $E(X_1) = \mu$, $\text{Var}(X_1) = \sigma^2$. We have the following.

Theorem 5.3.1. *If (i) and (ii) hold, then*

$$Eh(\bar{X}) = h(\mu) + \sum_{j=1}^{m-1} \frac{h^{(j)}(\mu)}{j!} E(\bar{X} - \mu)^j + R_m \tag{5.3.1}$$

where

$$|R_m| \leq M \frac{E|X_1|^m}{m!} n^{-m/2}.$$

The proof is an immediate consequence of Taylor's expansion.

$$h(\bar{X}) = h(\mu) + \sum_{k=1}^{m-1} \frac{h^{(k)}(\mu)}{k!} (\bar{X} - \mu)^k + \frac{h^{(m)}(\bar{X}^*)}{m!} (\bar{X} - \mu)^m \tag{5.3.2}$$

where $|\bar{X}^* - \mu| \leq |\bar{X} - \mu|$, and the following lemma.

Theorem 5.3.4. *Suppose $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent identically distributed d vectors with $E|\mathbf{Y}_1|^2 < \infty$, $E\mathbf{Y}_1 = \mathbf{m}$,* Var $\mathbf{Y}_1 = \boldsymbol{\Sigma}$ *and $\mathbf{h} : \mathcal{O} \to R^p$ where $\mathcal{O}$ is an open subset of R^d, $\mathbf{h} = (h_1, \ldots, h_p)$ and $\mathbf{h}$ has a total differential $\mathbf{h}^{(1)}(\mathbf{m}) = \left\| \frac{\partial h_i}{\partial x_j}(\mathbf{m}) \right\|_{p \times d}$. Then*

$$\mathbf{h}(\bar{\mathbf{Y}}) = \mathbf{h}(\mathbf{m}) + \mathbf{h}^{(1)}(\mathbf{m})(\bar{\mathbf{Y}} - \mathbf{m}) + o_p(n^{-1/2}) \tag{5.3.23}$$

$$\sqrt{n}[\mathbf{h}(\bar{\mathbf{Y}}) - \mathbf{h}(\mathbf{m})] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{h}^{(1)}(\mathbf{m})\boldsymbol{\Sigma}[\mathbf{h}^{(1)}(\mathbf{m})]^T) \tag{5.3.24}$$

Proof. Argue as before using B.8.5

(a) $$\mathbf{h}(\mathbf{y}) = \mathbf{h}(\mathbf{m}) + \mathbf{h}^{(1)}(\mathbf{m})(\mathbf{y} - \mathbf{m}) + o(|\mathbf{y} - \mathbf{m}|)$$

and

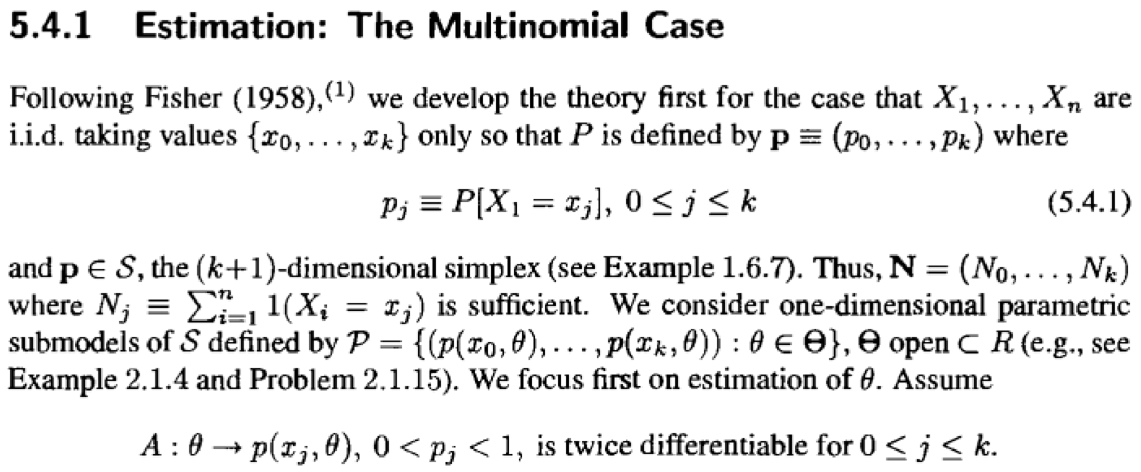
(b) $$\sqrt{n}(\bar{\mathbf{Y}} - \mathbf{m}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma})$$

so that

(c) $$\sqrt{n}(\mathbf{h}(\bar{\mathbf{Y}}) - \mathbf{h}(\mathbf{m})) = \sqrt{n}\mathbf{h}^{(1)}(\mathbf{m})(\bar{\mathbf{Y}} - \mathbf{m}) + o_p(1).$$

□

2.3.1 Multinomial Case



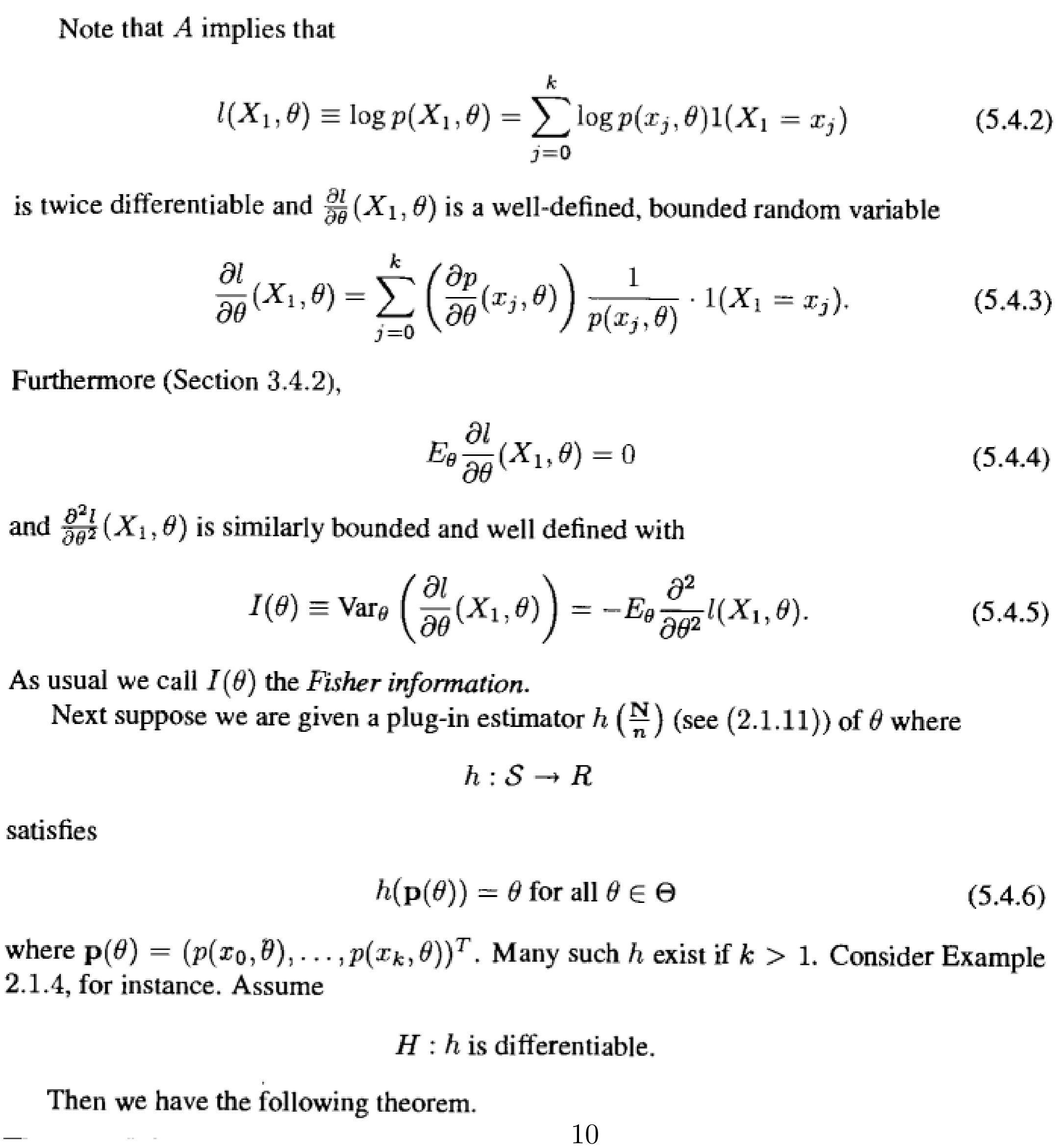
5.4.1 Estimation: The Multinomial Case

Following Fisher (1958),$^{(1)}$ we develop the theory first for the case that $X_1, \ldots, X_n$ are i.i.d. taking values $\{x_0, \ldots, x_k\}$ only so that P is defined by $\mathbf{p} \equiv (p_0, \ldots, p_k)$ where

$$p_j \equiv P[X_1 = x_j],\ 0 \leq j \leq k \tag{5.4.1}$$

and $\mathbf{p} \in \mathcal{S}$, the $(k+1)$-dimensional simplex (see Example 1.6.7). Thus, $\mathbf{N} = (N_0, \ldots, N_k)$ where $N_j \equiv \sum_{i=1}^{n} 1(X_i = x_j)$ is sufficient. We consider one-dimensional parametric submodels of $\mathcal{S}$ defined by $\mathcal{P} = \{(p(x_0, \theta), \ldots, p(x_k, \theta)) : \theta \in \Theta\}$, Θ open $\subset R$ (e.g., see Example 2.1.4 and Problem 2.1.15). We focus first on estimation of θ. Assume

$$A : \theta \to p(x_j, \theta),\ 0 < p_j < 1, \text{ is twice differentiable for } 0 \leq j \leq k.$$



Note that A implies that

$$l(X_1, \theta) \equiv \log p(X_1, \theta) = \sum_{j=0}^{k} \log p(x_j, \theta) 1(X_1 = x_j) \tag{5.4.2}$$

is twice differentiable and $\frac{\partial l}{\partial \theta}(X_1, \theta)$ is a well-defined, bounded random variable

$$\frac{\partial l}{\partial \theta}(X_1, \theta) = \sum_{j=0}^{k} \left(\frac{\partial p}{\partial \theta}(x_j, \theta) \right) \frac{1}{p(x_j, \theta)} \cdot 1(X_1 = x_j). \tag{5.4.3}$$

Furthermore (Section 3.4.2),

$$E_\theta \frac{\partial l}{\partial \theta}(X_1, \theta) = 0 \tag{5.4.4}$$

and $\frac{\partial^2 l}{\partial \theta^2}(X_1, \theta)$ is similarly bounded and well defined with

$$I(\theta) \equiv \mathrm{Var}_\theta \left(\frac{\partial l}{\partial \theta}(X_1, \theta) \right) = -E_\theta \frac{\partial^2}{\partial \theta^2} l(X_1, \theta). \tag{5.4.5}$$

As usual we call $I(\theta)$ the *Fisher information*.

Next suppose we are given a plug-in estimator $h\left(\frac{\mathbf{N}}{n}\right)$ (see (2.1.11)) of θ where

$$h : \mathcal{S} \to R$$

satisfies

$$h(\mathbf{p}(\theta)) = \theta \text{ for all } \theta \in \Theta \tag{5.4.6}$$

where $\mathbf{p}(\theta) = (p(x_0, \vartheta), \ldots, p(x_k, \theta))^T$. Many such h exist if $k > 1$. Consider Example 2.1.4, for instance. Assume

$$H : h \text{ is differentiable.}$$

Then we have the following theorem.

Theorem 5.4.1. *Under H, for all θ,*

$$\mathcal{L}_\theta\left(\sqrt{n}\left(h\left(\frac{\mathbf{N}}{n}\right)-\theta\right)\right)\rightarrow\mathcal{N}(0,\sigma^2(\theta,h)) \tag{5.4.7}$$

where $\sigma^2(\theta,h)$ is given by (5.4.11). *Moreover, if A also holds,*

$$\sigma^2(\theta,h)\geq I^{-1}(\theta) \tag{5.4.8}$$

with equality if and only if,

$$\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\bigg|_{\mathbf{p}(\theta)}=I^{-1}(\theta)\frac{\partial l}{\partial\theta}(x_j,\theta),\ 0\leq j\leq k. \tag{5.4.9}$$

Proof. Apply Theorem 5.3.2 noting that

$$\sqrt{n}\left(h\left(\frac{\mathbf{N}}{n}\right)-h(\mathbf{p}(\theta))\right)=\sqrt{n}\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\left(\frac{N_j}{n}-p(x_j,\theta)\right)+o_p(1).$$

Note that, using the definition of N_j,

$$\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\left(\frac{N_j}{n}-p(x_j,\theta)\right)=n^{-1}\sum_{i=1}^{n}\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))(1(X_i=x_j)-p(x_j,\theta)). \tag{5.4.10}$$

Thus, by (5.4.10), not only is $\sqrt{n}\left\{h\left(\frac{\mathbf{N}}{n}\right)-h(\mathbf{p}(\theta))\right\}$ asymptotically normal with mean 0, but also its asymptotic variance is

$$\begin{aligned}\sigma^2(\theta,h) &= \mathrm{Var}_\theta\left(\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))1(X_1=x_j)\right)\\ &= \sum_{j=0}^{k}\left(\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\right)^2p(x_j,\theta)-\left(\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))p(x_j,\theta)\right)^2.\end{aligned} \tag{5.4.11}$$

Note that by differentiating (5.4.6), we obtain

$$\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\frac{\partial p}{\partial\theta}(x_j,\theta)=1 \tag{5.4.12}$$

or equivalently, by noting $\frac{\partial p}{\partial\theta}(x_j,\theta)=\left[\frac{\partial}{\partial\theta}l(x_j,\theta)\right]p(x_j,\theta)$,

$$\mathrm{Cov}_\theta\left(\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))1(X_1=x_j),\frac{\partial l}{\partial\theta}(X_1,\theta)\right)=1. \tag{5.4.13}$$

By (5.4.13), using the correlation inequality (A.11.16) as in the proof of the information inequality (3.4.12), we obtain

$$1 \leq \sigma^2(\theta, h) \mathrm{Var}_\theta \frac{\partial l}{\partial \theta}(X_1, \theta) = \sigma^2(\theta, h) I(\theta) \tag{5.4.14}$$

with equality iff,

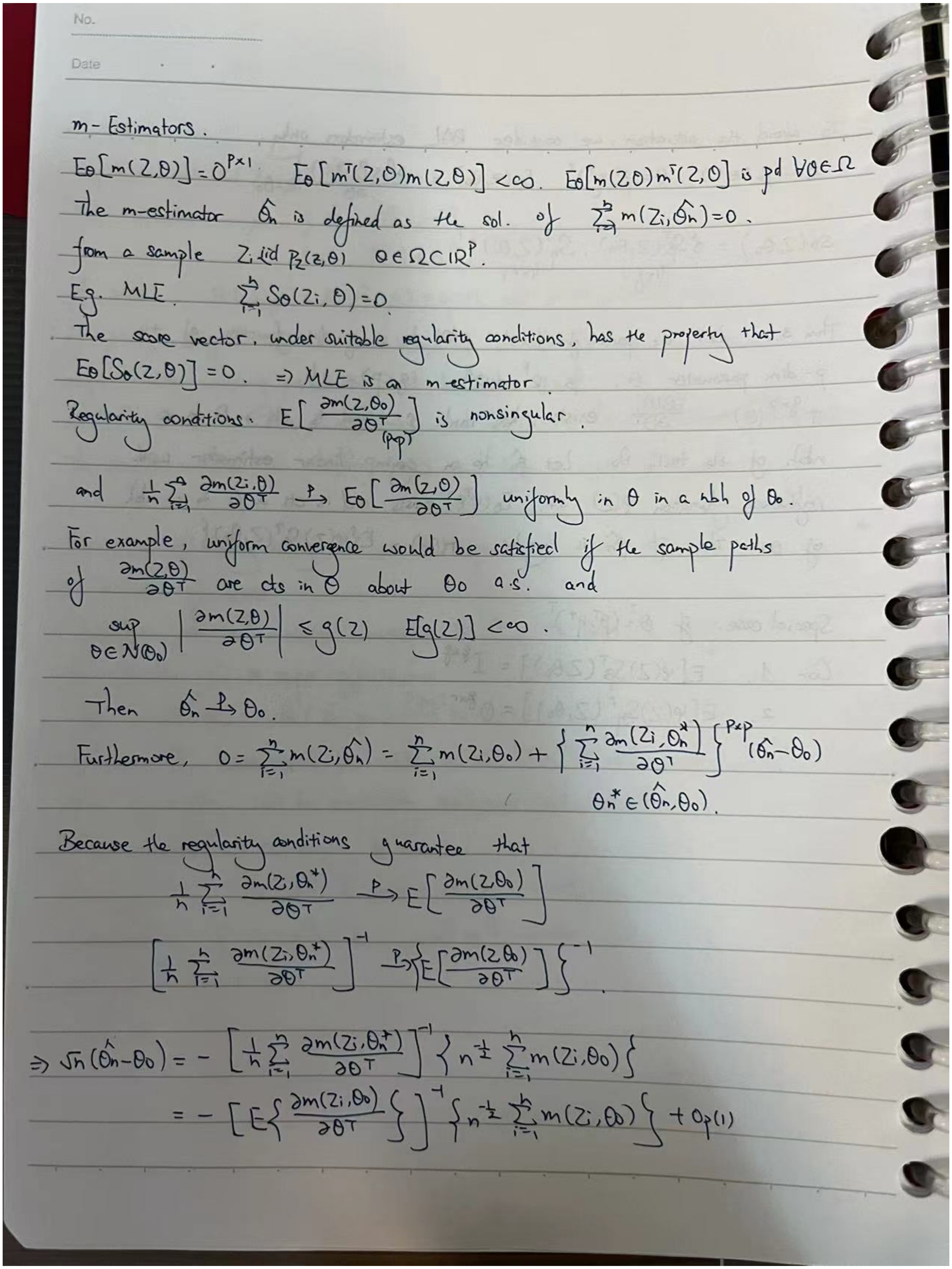
$$\sum_{j=0}^{k} \frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))(1(X_1 = x_j) - p(x_j, \theta)) = a(\theta) \frac{\partial l}{\partial \theta}(X_1, \theta) + b(\theta) \tag{5.4.15}$$

for some $a(\theta) \neq 0$ and some $b(\theta)$ with probability 1. Taking expectations we get $b(\theta) = 0$. Noting that the covariance of the right- and left-hand sides is $a(\theta)$, while their common variance is $a^2(\theta)I(\theta) = \sigma^2(\theta, h)$, we see that equality in (5.4.8) gives

$$a^2(\theta) I^2(\theta) = 1, \tag{5.4.16}$$

which implies (5.4.9). □

2.4 m-Estimator



m- Estimators.

$E_\theta[m(Z,\theta)] = 0^{p\times 1}$ $\quad E_\theta[m^T(Z,\theta)m(Z,\theta)] < \infty$. $\quad E_\theta[m(Z,\theta)m^T(Z,\theta)]$ is pd $\forall \theta\in\Omega$

The m-estimator $\hat{\theta}_n$ is defined as the sol. of $\sum_{i=1}^{n} m(Z_i,\hat{\theta}_n)=0$.

from a sample $Z_i \overset{iid}{\sim} P_Z(z,\theta)$ $\quad \theta\in\Omega\subset\mathbb{R}^p$.

Eg. MLE. $\quad \sum_{i=1}^{n} S_\theta(Z_i,\theta)=0$

The score vector, under suitable regularity conditions, has the property that $E_\theta[S_\theta(Z,\theta)]=0$. $\Rightarrow$ MLE is an m-estimator.

Regularity conditions. $E\Big[\underset{(p\times p)}{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}}\Big]$ is nonsingular.

and $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta)}{\partial\theta^T} \xrightarrow{P} E_\theta\Big[\frac{\partial m(Z,\theta)}{\partial\theta^T}\Big]$ uniformly in θ in a nbh of θ_0.

For example, uniform convergence would be satisfied if the sample paths of $\frac{\partial m(Z,\theta)}{\partial\theta^T}$ are cts in θ about θ_0 a.s. and

$$\sup_{\theta\in N(\theta_0)}\left|\frac{\partial m(Z,\theta)}{\partial\theta^T}\right| \le g(Z) \quad E[g(Z)]<\infty .$$

Then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Furthermore, $0=\sum_{i=1}^{n} m(Z_i,\hat{\theta}_n) = \sum_{i=1}^{n} m(Z_i,\theta_0) + \Big\{\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T}\Big\}^{p\times p}(\hat{\theta}_n-\theta_0)$

$\theta_n^*\in(\hat{\theta}_n,\theta_0)$.

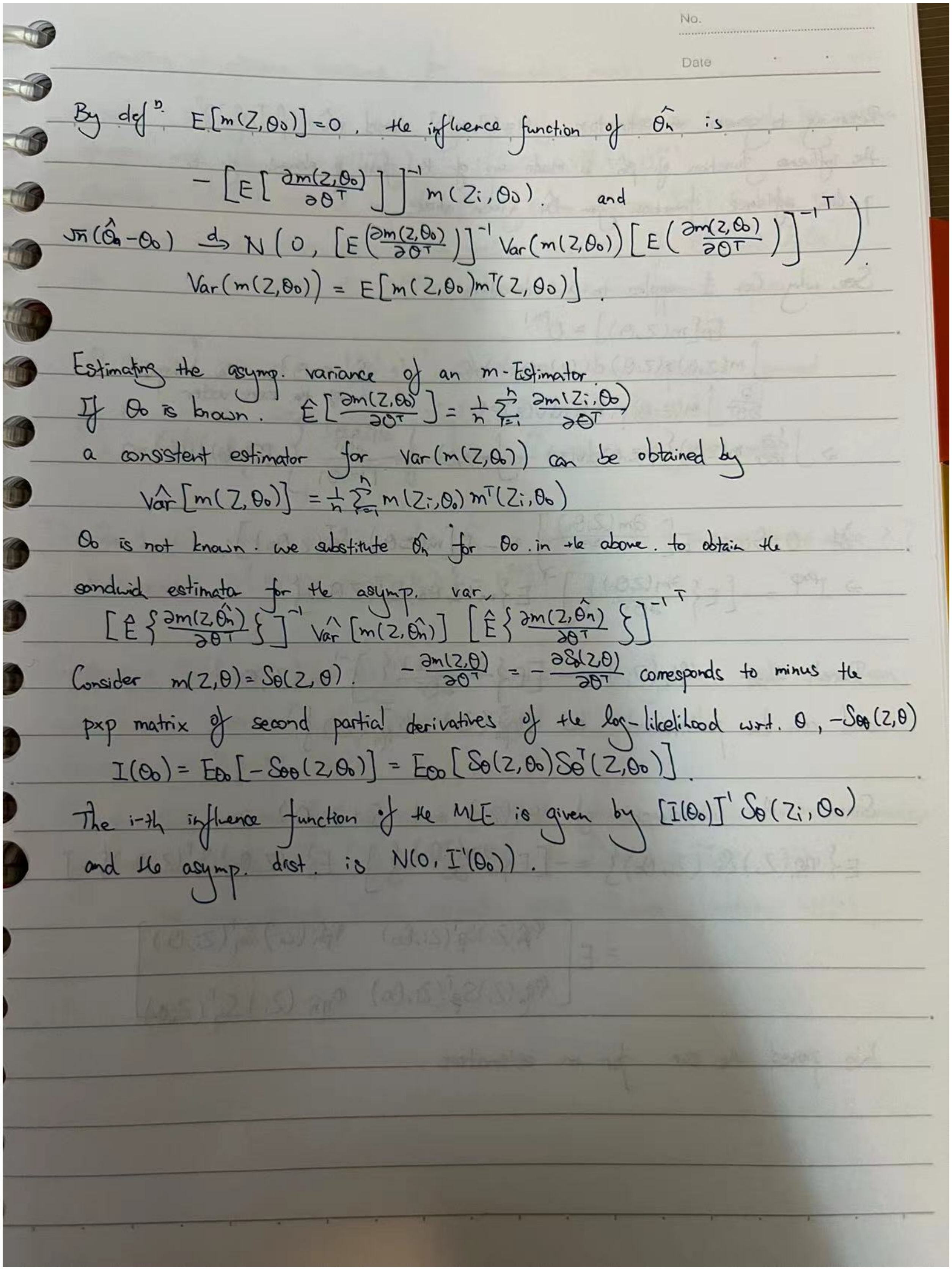
Because the regularity conditions guarantee that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T} \xrightarrow{P} E\Big[\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\Big]$$

$$\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T}\Big]^{-1} \xrightarrow{P} \Big\{E\Big[\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\Big]\Big\}^{-1}.$$

$$\Rightarrow \sqrt{n}(\hat{\theta}_n-\theta_0) = -\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^T}\Big]^{-1}\Big\{n^{-\frac{1}{2}}\sum_{i=1}^{n} m(Z_i,\theta_0)\Big\}$$

$$= -\Big[E\Big\{\frac{\partial m(Z_i,\theta_0)}{\partial\theta^T}\Big\}\Big]^{-1}\Big\{n^{-\frac{1}{2}}\sum_{i=1}^{n} m(Z_i,\theta_0)\Big\} + o_p(1)$$



No.

Date

By def^{n}: $E[m(Z,\theta_0)]=0$, the influence function of $\hat{\theta_n}$ is

$$-\left[E\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right]\right]^{-1} m(Z_i,\theta_0).$$ and

$$\sqrt{n}(\hat{\theta}_n-\theta_0) \xrightarrow{d} N\left(0, \left[E\left(\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right)\right]^{-1} Var(m(Z,\theta_0)) \left[E\left(\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right)\right]^{-1,T}\right).$$

$$Var(m(Z,\theta_0)) = E[m(Z,\theta_0)m^T(Z,\theta_0)].$$

Estimating the asymp. variance of an m-Estimator

If θ_0 is known. $\hat{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\theta_0)}{\partial\theta^T}$)

a consistent estimator for $Var(m(Z,\theta_0))$ can be obtained by

$$\hat{Var}[m(Z,\theta_0)] = \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\theta_0)\, m^T(Z_i,\theta_0)$$

θ_0 is not known. we substitute $\hat{\theta_n}$ for θ_0. in the above. to obtain the sandwich estimator for the asymp. var,

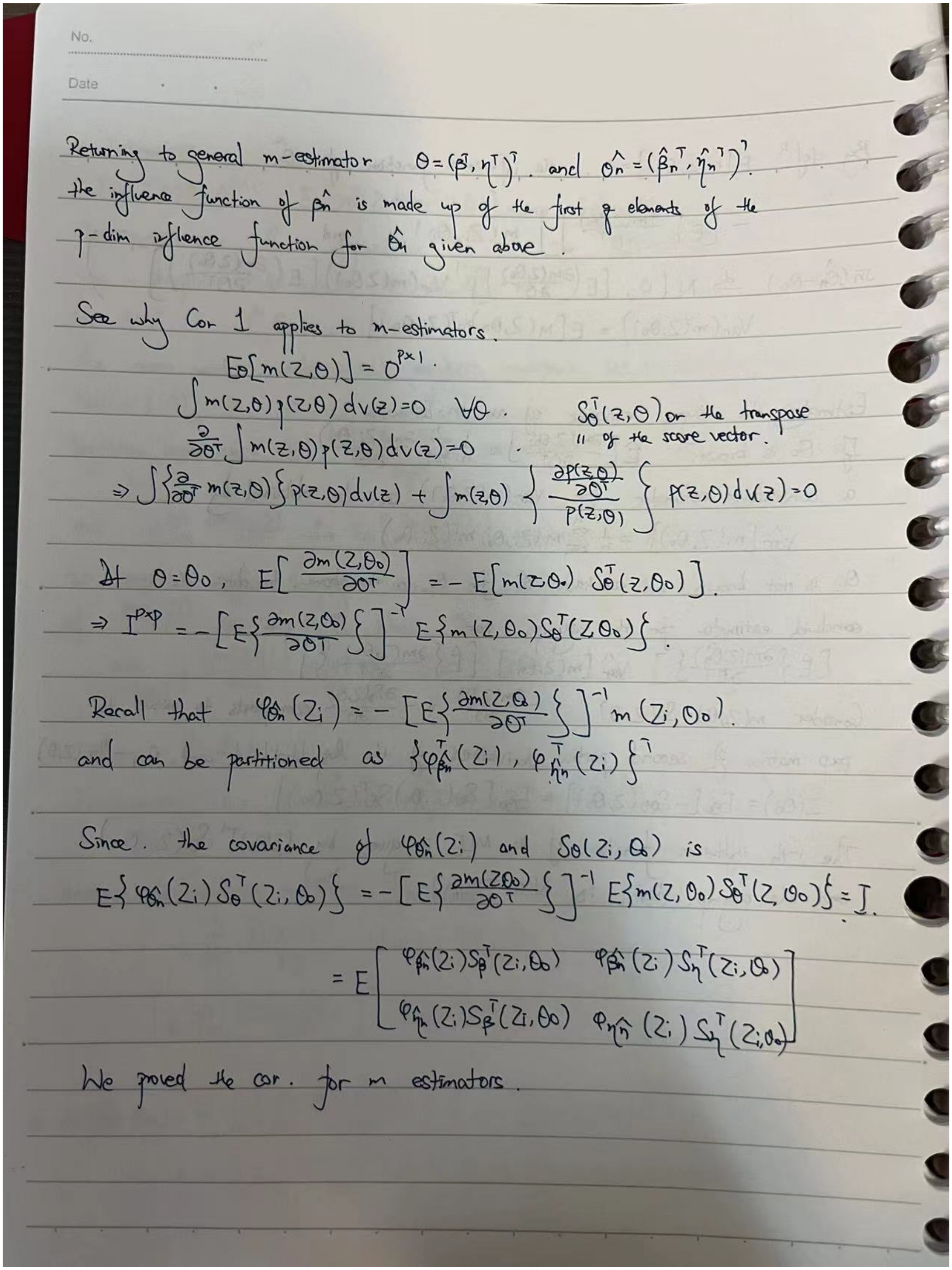
$$\left[\hat{E}\left\{\frac{\partial m(Z,\hat{\theta_n})}{\partial\theta^T}\right\}\right]^{-1} \hat{Var}[m(Z,\hat{\theta_n})] \left[\hat{E}\left\{\frac{\partial m(Z,\hat{\theta_n})}{\partial\theta^T}\right\}\right]^{-1\,T}$$

Consider $m(Z,\theta)=S_\theta(Z,\theta)$. $-\frac{\partial m(Z,\theta)}{\partial\theta^T} = -\frac{\partial S_\theta(Z,\theta)}{\partial\theta^T}$ corresponds to minus the $p\times p$ matrix of second partial derivatives of the log-likelihood w.r.t. θ, $-S_{\theta\theta}(Z,\theta)$

$$I(\theta_0) = E_{\theta_0}[-S_{\theta\theta}(Z,\theta_0)] = E_{\theta_0}[S_\theta(Z,\theta_0)S_\theta^T(Z,\theta_0)].$$

The i-th influence function of the MLE is given by $[I(\theta_0)]^{-1} S_\theta(Z_i,\theta_0)$

and the asymp. dist. is $N(0, I^{-1}(\theta_0))$.



No.

Date

Returning to general m-estimator $\theta = (\beta^T, \eta^T)^T$. and $\hat{\theta_n} = (\hat{\beta}_n^T, \hat{\eta}_n^T)^T$.
the influence function of $\hat{\beta_n}$ is made up of the first q elements of the p-dim inflence function for $\hat{\theta_n}$ given above.

See why Cor 1 applies to m-estimators.

$$E_\theta[m(Z,\theta)] = 0^{p\times 1}.$$

$$\int m(Z,\theta)p(Z,\theta)\,dv(z) = 0 \quad \forall\theta.$$

$S_\theta^T(z,\theta)$ or the transpose
" of the score vector.

$$\frac{\partial}{\partial\theta^T}\int m(z,\theta)p(z,\theta)\,dv(z) = 0$$

$$\Rightarrow \int\left\{\frac{\partial}{\partial\theta^T} m(z,\theta)\right\}p(z,\theta)\,dv(z) + \int m(z,\theta)\left\{\frac{\frac{\partial p(z,\theta)}{\partial\theta^T}}{p(z,\theta)}\right\}p(z,\theta)\,dv(z) = 0$$

At $\theta = \theta_0$, $E\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right] = -E\left[m(Z,\theta_0)\,S_\theta^T(z,\theta_0)\right]$.

$$\Rightarrow I^{p\times p} = -\left[E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}\right]^{-1} E\{m(Z,\theta_0)S_\theta^T(Z,\theta_0)\}.$$

Recall that $\varphi_{\hat{\theta}_n}(Z_i) = -\left[E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}\right]^{-1} m(Z_i,\theta_0)$.
and can be partitioned as $\{\varphi^T_{\hat{\beta}_n}(Z_i), \varphi^T_{\hat{\eta}_n}(Z_i)\}^T$

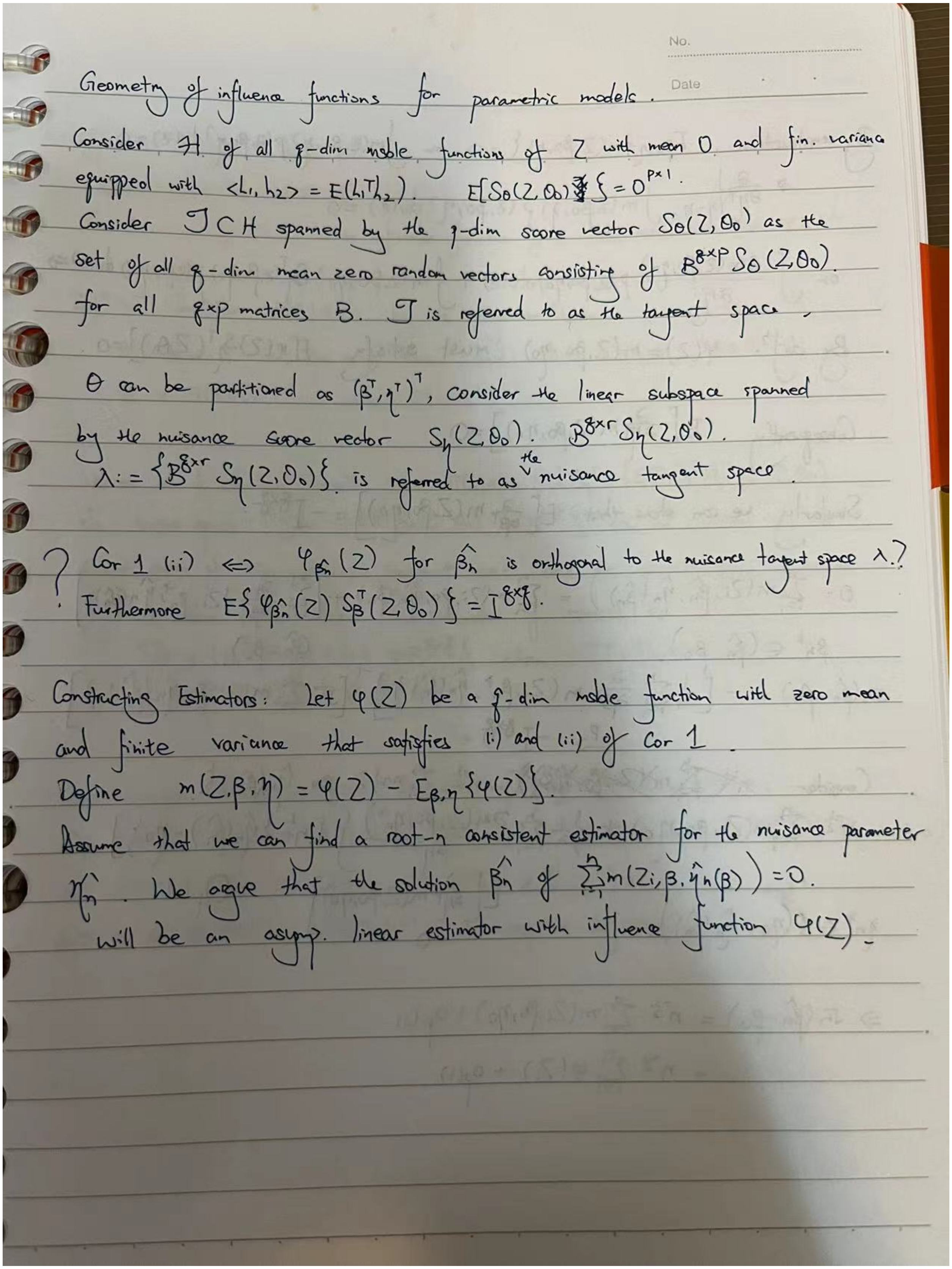
Since. the covariance of $\varphi_{\hat{\theta}_n}(Z_i)$ and $S_\theta(Z_i,\theta_0)$ is

$$E\{\varphi_{\hat{\theta}_n}(Z_i)S_\theta^T(Z_i,\theta_0)\} = -\left[E\left\{\frac{\partial m(Z,\theta_0)}{\partial\theta^T}\right\}\right]^{-1} E\{m(Z,\theta_0)S_\theta^T(Z,\theta_0)\} = I.$$

$$= E\begin{bmatrix} \varphi_{\hat{\beta}_n}(Z_i)S_\beta^T(Z_i,\theta_0) & \varphi_{\hat{\beta}_n}(Z_i)S_\eta^T(Z_i,\theta_0) \\ \varphi_{\hat{\eta}_n}(Z_i)S_\beta^T(Z_i,\theta_0) & \varphi_{\eta\hat{\eta}}(Z_i)S_\eta^T(Z_i,\theta_0) \end{bmatrix}$$

We proved the cor. for m estimators.

2.5 Geometry of Influence Functions for Parametric Models



No.

Date

Geometry of influence functions for parametric models.

Consider $\mathcal{H}$ of all q-dim msble functions of Z with mean 0 and fin. variance equipped with $<h_1, h_2> = E(h_1^T h_2)$. $E[S_\theta(Z,\theta_0)\} = 0^{p\times 1}$.

Consider $\mathcal{J} \subset \mathcal{H}$ spanned by the q-dim score vector $S_\theta(Z,\theta_0)$ as the set of all q-dim mean zero random vectors consisting of $B^{q\times p} S_\theta(Z,\theta_0)$ for all $q\times p$ matrices B. $\mathcal{J}$ is refered to as the tangent space.

θ can be partitioned as $(\beta^T,\eta^T)^T$, consider the linear subspace spanned by the nuisance score vector $S_\eta(Z,\theta_0)$. $B^{q\times r} S_\eta(Z,\theta_0)$.

$\Lambda := \{B^{q\times r} S_\eta(Z,\theta_0)\}$ is referred to as the nuisance tangent space.

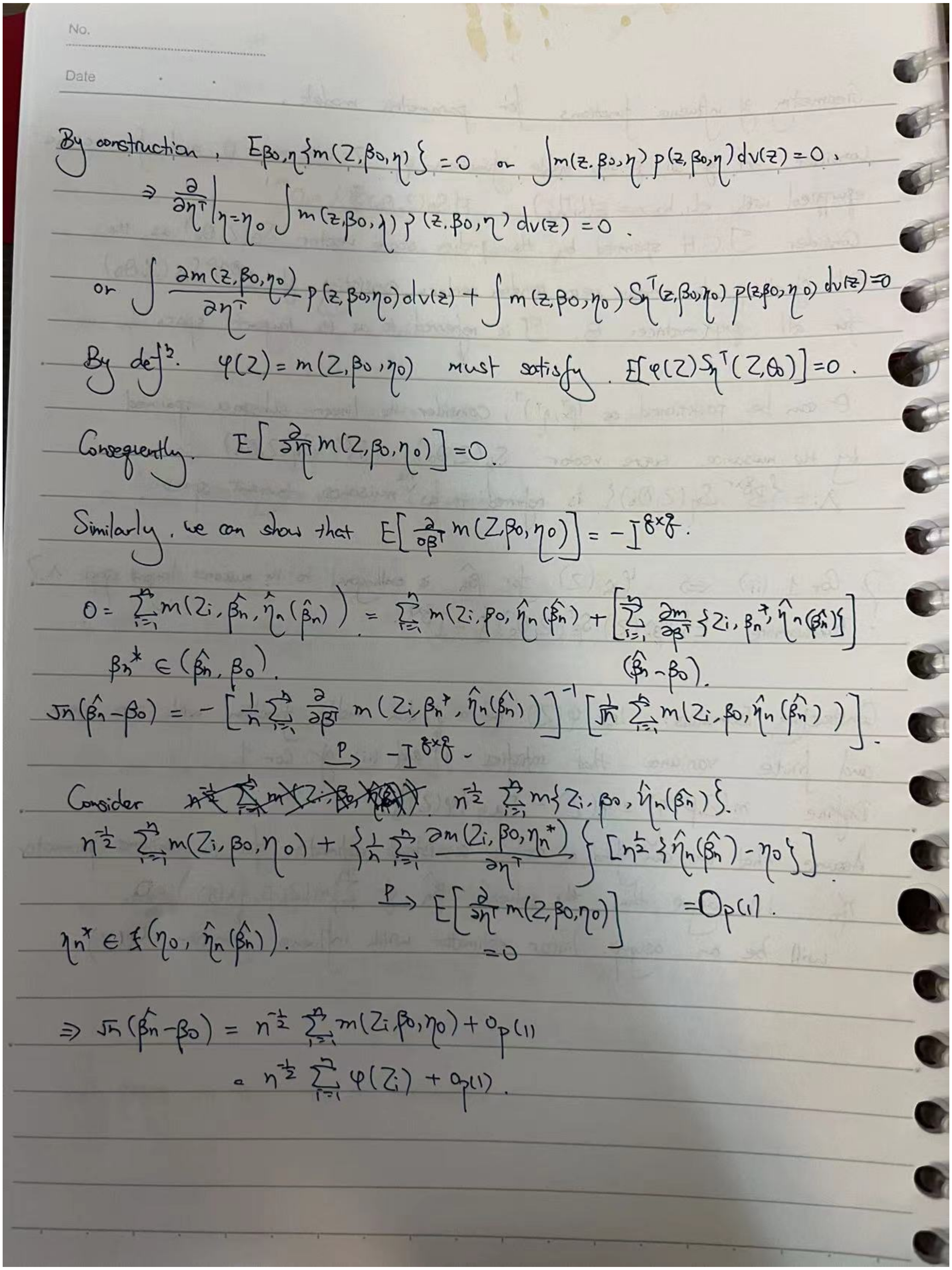
? Cor 1 (ii) $\Leftrightarrow$ $\varphi_{\hat\beta_n}(Z)$ for $\hat\beta_n$ is orthogonal to the nuisance tangent space Λ.?

Furthermore $E\{\varphi_{\hat\beta_n}(Z) S_\beta^T(Z,\theta_0)\} = I^{q\times q}$.

Constructing Estimators: Let $\varphi(Z)$ be a q-dim msble function with zero mean and finite variance that satisfies (i) and (ii) of Cor 1.

Define $m(Z,\beta,\eta) = \varphi(Z) - E_{\beta,\eta}\{\varphi(Z)\}$.

Assume that we can find a root-n consistent estimator for the nuisance parameter $\hat\eta_n$. We argue that the solution $\hat\beta_n$ of $\sum_{i=1}^n m(Z_i,\beta,\hat\eta_n(\beta))) = 0$. will be an asymp. linear estimator with influence function $\varphi(Z)$.



No.

Date

By construction, $E_{\beta_0,\eta}\{m(Z,\beta_0,\eta)\} = 0$ or $\int m(z.\beta_0,\eta)\,p(z,\beta_0,\eta)\,d\nu(z) = 0$.

$$\Rightarrow \frac{\partial}{\partial\eta^T}\Big|_{\eta=\eta_0} \int m(z,\beta_0,\lambda)\,p(z.\beta_0,\eta)\,d\nu(z) = 0.$$

or $$\int \frac{\partial m(z,\beta_0,\eta_0)}{\partial\eta^T}\,p(z,\beta_0,\eta_0)\,d\nu(z) + \int m(z,\beta_0,\eta_0)\,S_\eta^T(z,\beta_0,\eta_0)\,p(z,\beta_0,\eta_0)\,d\nu(z) = 0$$

By def^n. $\varphi(Z) = m(Z,\beta_0,\eta_0)$ must satisfy. $E[\varphi(Z)S_\eta^T(Z,\theta_0)] = 0$.

Consequently. $E\left[\frac{\partial}{\partial\eta^T} m(Z,\beta_0,\eta_0)\right] = 0$.

Similarly. we can show that $E\left[\frac{\partial}{\partial\beta^T} m(Z,\beta_0,\eta_0)\right] = -I^{q\times q}$.

$$0 = \sum_{i=1}^n m(Z_i,\hat\beta_n,\hat\eta_n(\hat\beta_n)) = \sum_{i=1}^n m(Z_i,\beta_0,\hat\eta_n(\hat\beta_n)) + \left[\sum_{i=1}^n \frac{\partial m}{\partial\beta^T}\{Z_i,\beta_n^*,\hat\eta_n(\hat\beta_n)\}\right](\hat\beta_n-\beta_0).$$

$\beta_n^* \in (\hat\beta_n,\beta_0)$.

$$\sqrt{n}(\hat\beta_n-\beta_0) = -\underbrace{\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta^T} m(Z_i,\beta_n^*,\hat\eta_n(\hat\beta_n))\right]}_{\xrightarrow{P}\ -I^{q\times q}}{}^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n m(Z_i,\beta_0,\hat\eta_n(\hat\beta_n))\right].$$

Consider ~~$n^{-\frac{1}{2}}\sum_{i=1}^n m(Z_i,\beta_0,\hat\eta(\hat\beta))$~~. $n^{-\frac{1}{2}}\sum_{i=1}^n m\{Z_i,\beta_0,\hat\eta_n(\hat\beta_n)\}$.

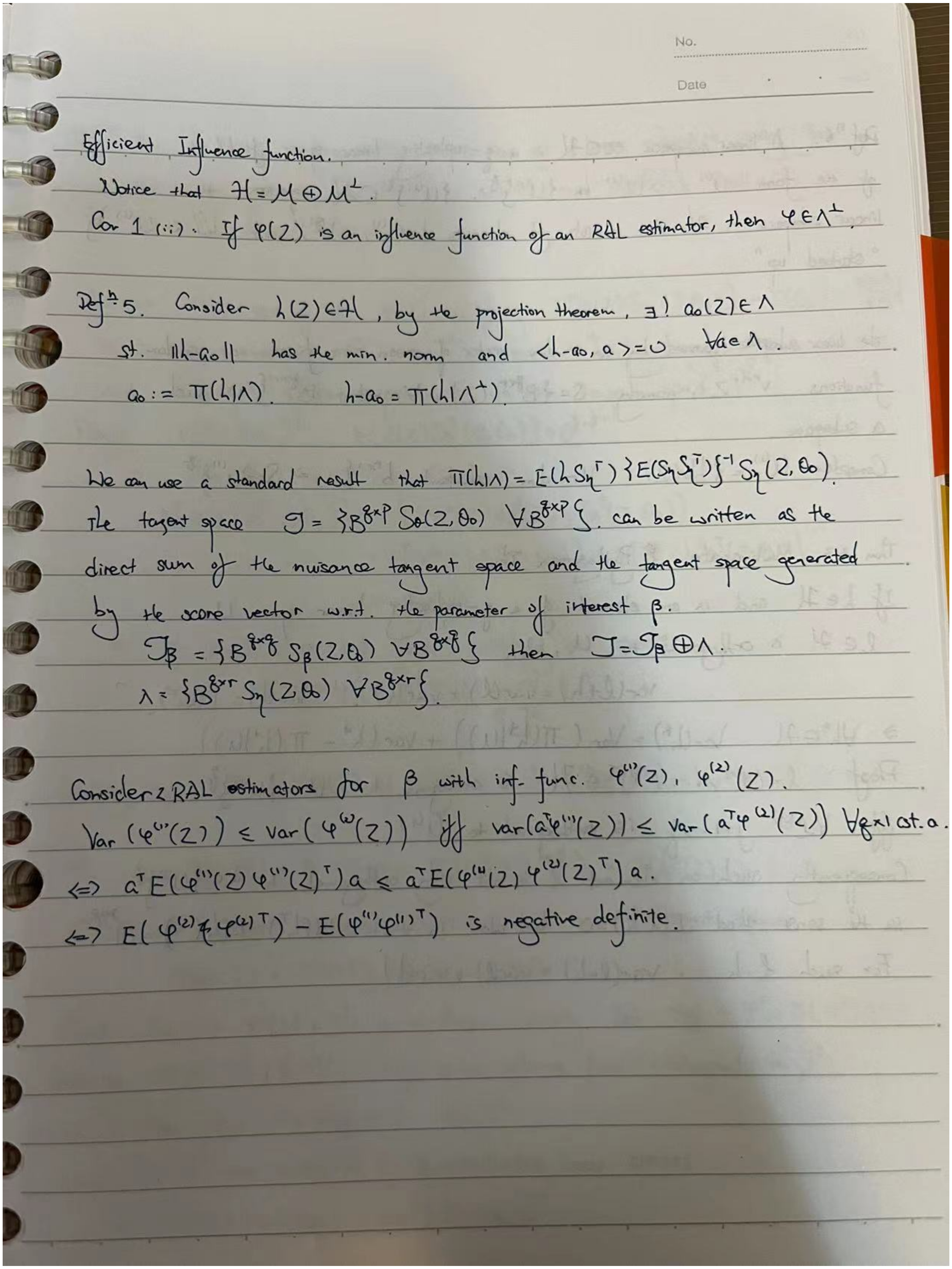
$$n^{-\frac{1}{2}}\sum_{i=1}^n m(Z_i,\beta_0,\eta_0) + \underbrace{\left\{\frac{1}{n}\sum_{i=1}^n \frac{\partial m(Z_i,\beta_0,\eta_n^*)}{\partial\eta^T}\right\}}_{\xrightarrow{P}\ \underbrace{E\left[\frac{\partial}{\partial\eta^T} m(Z,\beta_0,\eta_0)\right]}_{=0}} \underbrace{\left[n^{\frac{1}{2}}\{\hat\eta_n(\hat\beta_n)-\eta_0\}\right]}_{=O_p(1)}.$$

$\eta_n^* \in (\eta_0,\hat\eta_n(\hat\beta_n))$.

$$\Rightarrow \sqrt{n}(\hat\beta_n-\beta_0) = n^{-\frac{1}{2}}\sum_{i=1}^n m(Z_i,\beta_0,\eta_0) + o_p(1)$$

$$= n^{-\frac{1}{2}}\sum_{i=1}^n \varphi(Z_i) + o_p(1).$$

2.6 Efficient Influence Function



Efficient Influence function.

Notice that $\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^{\perp}$.

Cor 1 (ii). If $\varphi(Z)$ is an influence function of an RAL estimator, then $\varphi \in \Lambda^{\perp}$.

Def$^{\underline{n}}$ 5. Consider $h(Z) \in \mathcal{H}$, by the projection theorem, $\exists!\ a_0(Z) \in \Lambda$ s.t. $\|h - a_0\|$ has the min. norm and $\langle h - a_0, a \rangle = 0 \quad \forall a \in \lambda$.

$a_0 := \Pi(h|\Lambda)$. $\quad h - a_0 = \Pi(h|\Lambda^{\perp})$.

We can use a standard result that $\Pi(h|\Lambda) = E(h S_\eta^T)\}E(S_\eta S_\eta^T)\{^{-1} S_\eta(Z, \theta_0)$.

The tagent space $\mathcal{J} = \}B^{q\times p} S_\theta(Z, \theta_0) \ \forall B^{q \times p}\{$. can be written as the direct sum of the nuisance tangent space and the tangent space generated by the score vector w.r.t. the parameter of interest β.

$\mathcal{J}_\beta = \{B^{q\times q} S_\beta(Z, \theta_0) \ \forall B^{q\times q}\{$ then $\mathcal{J} = \mathcal{J}_\beta \oplus \Lambda$.

$\lambda = \{B^{q\times r} S_\eta(Z\theta_0) \ \forall B^{q\times r}\{$.

Consider 2 RAL estimators for β with inf. func. $\varphi^{(1)}(Z)$, $\varphi^{(2)}(Z)$.

$\mathrm{Var}(\varphi^{(1)}(Z)) \le \mathrm{var}(\varphi^{(2)}(Z))$ iff $\mathrm{var}(a^T\varphi^{(1)}(Z)) \le \mathrm{Var}(a^T\varphi^{(2)}(Z)) \ \forall_{q\times 1}$ cst. a.

$\Leftrightarrow a^T E(\varphi^{(1)}(Z)\varphi^{(1)}(Z)^T) a \le a^T E(\varphi^{(2)}(Z)\varphi^{(2)}(Z)^T) a$.

$\Leftrightarrow E(\varphi^{(2)}\varphi^{(2)T}) - E(\varphi^{(1)}\varphi^{(1)T})$ is negative definite.

Def^{n} 6. A linear subspace $\mathcal{U} \subset \mathcal{H}$ is a q-replicating linear space if $\mathcal{U}$ is of the form $\mathcal{U}^{(1)} \times \cdots \times \mathcal{U}^{(1)}$ or $\{\mathcal{U}^{(1)}\}^{q}$. $\{\mathcal{U}^{(1)}\}^{q} \subset \mathcal{H}$ represents the linear subspace in $\mathcal{H}$ that consists of elements $h=(h^{(1)},\dots,h^{(q)})^{T}$ s.t. $h^{(j)} \in \mathcal{U}^{(1)}\ \forall j$ "stacked up"

The linear subspace spanned by an r-dim vector of mean zero finite var random functions $v^{r\times 1}(Z)$, namely $S=\{B^{q\times r}v(Z) : \forall$ constant $B^{q\times r}\}$ is such a subspace.

Consider $\mathcal{U}^{(1)}=\{b^{T}v(Z) : \forall r$-dim constant $b^{r\times 1}\}$. $\Rightarrow S=\{\mathcal{U}^{(1)}\}^{q}$.

Thm 3.3. Multivariate ~~B~~ Pythagorean Thm

If $h\in\mathcal{H}$ and is an element of a q-replicating linear space $\mathcal{U}$, $\ell\in\mathcal{H}$ is orthogonal to $\mathcal{U}$, then

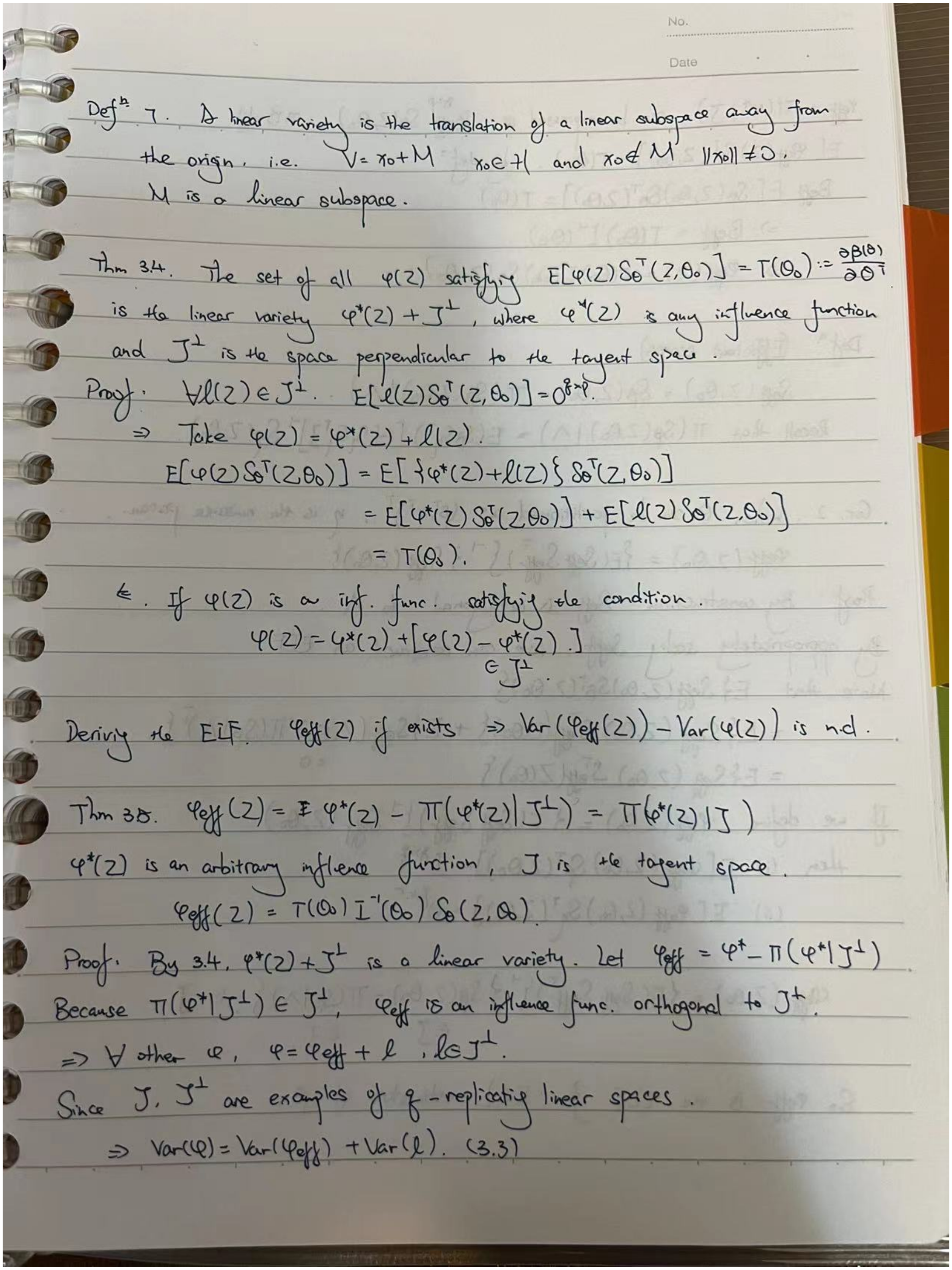
$$\mathrm{var}(\ell+h)=\mathrm{var}(\ell)+\mathrm{var}(h). \quad \mathrm{Var}(h)=E(hh^{T})$$

$\Rightarrow \forall h^{*}\in\mathcal{H}\quad \mathrm{Var}(h^{*})=\mathrm{Var}(\Pi(h^{*}|\mathcal{U}))+\mathrm{var}(h^{*}-\Pi(h^{*}|\mathcal{U}))$.

Proof: $\ell=(\ell^{(1)},\dots,\ell^{(q)})^{T}\in\mathcal{H}$ is orthogonal to $\mathcal{U}=\{\mathcal{U}^{(1)}\}^{q}$ iff $\forall \ell^{(j)}$ $j=1,\dots,q$ is orthogonal to $\mathcal{U}^{(1)}$.

Consequently, such an element ℓ is not only orthogonal to $h\in\{\mathcal{U}^{(1)}\}^{q}$ in the sense that $E(\ell^{T}h)=0$ but also in that $E(\ell h^{T})=E(h\ell^{T})=0^{q\times q}$.

For such ℓ, h, $\mathrm{var}(\ell+h)=\mathrm{var}(\ell)+\mathrm{var}(h)$.



Def^{n} 7. A linear variety is the translation of a linear subspace away from the orign, i.e. $V = x_0 + M$ $\quad x_0 \in \mathcal{H}$ and $x_0 \notin M$ $\|x_0\| \neq 0$, M is a linear subspace.

Thm 3.4. The set of all $\varphi(Z)$ satisfying $E[\varphi(Z) S_\theta^T(Z,\theta_0)] = T(\theta_0) := \frac{\partial \beta(\theta)}{\partial \theta^T}$ is the linear variety $\varphi^*(Z) + J^\perp$, where $\varphi^*(Z)$ is any influence function and $J^\perp$ is the space perpendicular to the tagent space.

Proof. $\forall \ell(Z) \in J^\perp$. $E[\ell(Z) S_\theta^T(Z,\theta_0)] = 0^{q \times p}$.

$\Rightarrow$ Take $\varphi(Z) = \varphi^*(Z) + \ell(Z)$.

$$E[\varphi(Z) S_\theta^T(Z,\theta_0)] = E[\{\varphi^*(Z) + \ell(Z)\} S_\theta^T(Z,\theta_0)]$$
$$= E[\varphi^*(Z) S_\theta^T(Z,\theta_0)] + E[\ell(Z) S_\theta^T(Z,\theta_0)]$$
$$= T(\theta_0).$$

$\Leftarrow$. If $\varphi(Z)$ is an inf. func. satisfying the condition.

$$\varphi(Z) = \varphi^*(Z) + \underbrace{[\varphi(Z) - \varphi^*(Z)]}_{\in J^\perp}.$$

Deriving the EIF. $\varphi_{eff}(Z)$ if exists $\Rightarrow \mathrm{Var}(\varphi_{eff}(Z)) - \mathrm{Var}(\varphi(Z))$ is n.d.

Thm 3.5. $\varphi_{eff}(Z) = \cancel{E}\, \varphi^*(Z) - \Pi(\varphi^*(Z) | J^\perp) = \Pi(\varphi^*(Z) | J)$

$\varphi^*(Z)$ is an arbitrary influence function, J is the tagent space.

$$\varphi_{eff}(Z) = T(\theta_0) I^{-1}(\theta_0) S_\theta(Z,\theta_0).$$

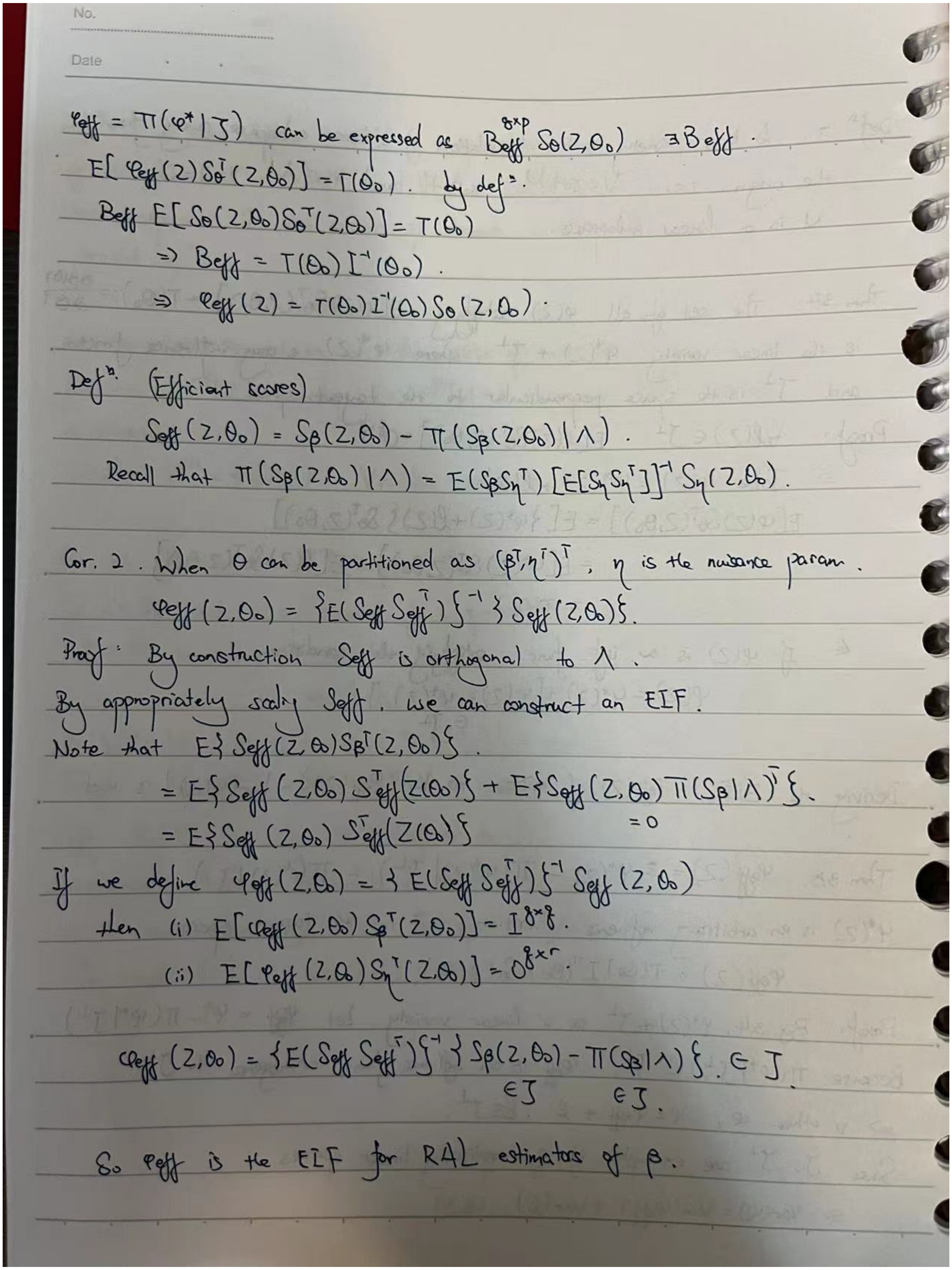
Proof. By 3.4, $\varphi^*(Z) + J^\perp$ is a linear variety. Let $\varphi_{eff} = \varphi^* - \Pi(\varphi^* | J^\perp)$

Because $\Pi(\varphi^* | J^\perp) \in J^\perp$, φ_{eff} is an influence func. orthogonal to $J^\perp$.

$\Rightarrow \forall$ other φ, $\varphi = \varphi_{eff} + \ell$, $\ell \in J^\perp$.

Since J, $J^\perp$ are examples of q-replicating linear spaces.

$\Rightarrow \mathrm{Var}(\varphi) = \mathrm{Var}(\varphi_{eff}) + \mathrm{Var}(\ell)$. (3.3)



No.

Date

$\varphi_{eff} = \Pi(\varphi^* | \mathcal{J})$ can be expressed as $\overset{q\times p}{B_{eff}} S_\theta(Z,\theta_0)$ $\exists B_{eff}$.

$E[\varphi_{eff}(Z) S_\theta^T(Z,\theta_0)] = \Gamma(\theta_0)$. by def^n.

$B_{eff} E[S_\theta(Z,\theta_0) S_\theta^T(Z,\theta_0)] = \Gamma(\theta_0)$

$\Rightarrow B_{eff} = \Gamma(\theta_0) I^{-1}(\theta_0)$.

$\Rightarrow \varphi_{eff}(Z) = \Gamma(\theta_0) I^{-1}(\theta_0) S_\theta(Z,\theta_0)$.

Def^n. (Efficient scores)

$$S_{eff}(Z,\theta_0) = S_\beta(Z,\theta_0) - \Pi(S_\beta(Z,\theta_0) | \Lambda).$$

Recall that $\Pi(S_\beta(Z,\theta_0)|\Lambda) = E(S_\beta S_\eta^T)[E[S_\eta S_\eta^T]]^{-1} S_\eta(Z,\theta_0)$.

Cor. 2. When θ can be partitioned as $(\beta^T, \eta^T)^T$, η is the nuisance param.

$$\varphi_{eff}(Z,\theta_0) = \{E(S_{eff} S_{eff}^T)\}^{-1} \{ S_{eff}(Z,\theta_0)\}.$$

Proof: By construction S_{eff} is orthogonal to Λ.

By appropriately scaling S_{eff}, we can construct an EIF.

Note that $E\{S_{eff}(Z,\theta_0) S_\beta^T(Z,\theta_0)\}$.

$= E\{S_{eff}(Z,\theta_0) S_{eff}^T(Z(\theta_0)\} + \underbrace{E\{S_{eff}(Z,\theta_0)\Pi(S_\beta|\Lambda)^T\}}_{=0}.$

$= E\{S_{eff}(Z,\theta_0) S_{eff}^T(Z(\theta_0)\}$

If we define $\varphi_{eff}(Z,\theta_0) = \{E(S_{eff} S_{eff}^T)\}^{-1} S_{eff}(Z,\theta_0)$

then (i) $E[\varphi_{eff}(Z,\theta_0) S_\beta^T(Z,\theta_0)] = I^{q\times q}$.

(ii) $E[\varphi_{eff}(Z,\theta_0) S_\eta^T(Z,\theta_0)] = 0^{q\times r}$.

$$\varphi_{eff}(Z,\theta_0) = \{E(S_{eff} S_{eff}^T)\}^{-1} \{ \underbrace{S_\beta(Z,\theta_0)}_{\in \mathcal{J}} - \underbrace{\Pi(S_\beta|\Lambda)}_{\in \mathcal{J}.} \}. \in \mathcal{J}.$$

So φ_{eff} is the EIF for RAL estimators of β.

2.7 Another Treatment

2.7.1 Semiparametric linear models with stochastic covariates

Suppose $\{(Z_i, Y_i)\} \overset{iid}{\sim} (Z, Y)$, $Z \in \mathbb{R}^p$, $Y \in \mathbb{R}$, with [4]

$$Y = \alpha + Z^T \beta + \sigma \epsilon$$

There are many semiparametric versions of this model. (a) Z and ϵ are independent, $\epsilon \sim F$, $Z \sim H$, $\Sigma \equiv Var_H(Z) \equiv E_H(Z - E_H(Z))(Z - E_H(Z))^T$ is nonsingular.
(b) The restricted moment model from above.
Model (a) is parametrized by $(\alpha, \beta, \sigma, F, H)$, and is not σ identifiable. A symmetry density f is required for identifiability of α, β is $\sqrt{n}$ consistently estimable in the sense that $\hat{\beta} = \beta + O_p(n^{-1/2})$ for model (a) using the estimating equations.

2.7.2 Biased Sampling. Stratification

We want to obtain information about a population df F, or at least features of F such as the mean and variance, we are only able to observe a sample from $P_{(F,G)}$, G is finite or infinite dimensional parameter. We begin by considering the nonparametric model of biased sampling from a single population. Here, if the distribution F of Z has density f, observations are not on Z but on X from

$$p_F(x) = \frac{w(x) f(x)}{W(F)}$$

where $W(F) = \int w(x) dF(x)$, and $w(\cdot)$ is assumed known.

An important special case is "length biased" sampling where $X \in \mathbb{R}^+$, $X \sim F$, and $w(x) = x$. This arises classically if we are interested in, the proportion $f(2)$ of two-child families in a city with a known total number of households N. If we could sample households with at least one child, and Z is the number of children in a sampled household we could estimate $f(2) = P(Z = 2)$ directly. Suppose instead we sample n children at random, and consider $p_F(2)$ of children coming from two-child families in this group. If the city is large, $f(j)$ is the proportion of j child families in the city, for all j, and X is the number of children in a sampled household, then,

$$p_F(2) = P(X = 2) = 2f(2) / \sum_{j=1}^{\infty} j f(j)$$

Based on $p_F(\cdot)$, is F identifiable? This is true iff $w(x) > 0$ whenever $f(x) > 0$. We claim that

$$f(x) = \frac{p_f(x)/w(x)}{\int \frac{1}{w(x)} dP_F(x)}$$

Now we consider stratified populations where biased sampling may occur within each strata. A stratified population S is made up of subpopulations $S_1, \ldots, S_k$ called strata. We are interested in the density $f(z)$ of a random draw Z from S, but sample by first randomly selecting an index I with corresponding strata S_I. This situation arises if S is the set of all patients in a country and $S_1, \ldots, S_k$ are the patients in the k hospitals and clinics in the country. Length biased sampling may occur within each strata. In this case, we observe $X = (I, Y)$, where $I = 1, \ldots, k$, $P(I = j) = \lambda_j$, $1 \leq j \leq k$, are assumed known, and given $I = j$, Y has density

$$p_F(y|j) = \frac{w_j(y) f(y)}{W_j(F)}$$

where $W_j(F) = \int w_j(x) dF(x)$. Again, $w_1, \ldots, W_k$ are assumed known. In stratified sampling λ_j is the probability that Y will be from the j-th strata S_j and $w_j(x) = 1(x \in S_j)$. We claim that F is identifiable iff $\sum_{j=1}^k \lambda_j w_j(x) f(x) > 0$.

3 Empirical Processes

3.1 Empirical Distribution Functions

Let $X_1, \ldots, X_n$ be a random sample from a distribution function F on the real line. The empirical distribution function is defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n 1\{X_i \leq t\}$$

Notice that $nF_n(t) \sim Bin(n, F(t))$, by the SLLN,

$$F_n(t) \overset{\text{a.s.}}{\to} F(t), \quad \text{for all t}$$

By the CLT,

$$\sqrt{n}(F_n(t) - F(t)) \overset{d}{\to} N\left(0, F(t)(1 - F(t))\right)$$

Consider the uniform distance

$$||F_n - F||_\infty = \underset{t}{sup} |F_n(t) - F(t)|$$

is known as the Kolmogorov-Smirnov statistic.

Theorem 1. Glivenko-Cantelli
If $X_1, \ldots$ are i.i.d. random variables with distribution function F, then $||F_n - F||_\infty \overset{a.s.}{\to} 0$

Proof. By the SLLN, $F_n(t) \overset{\text{a.s.}}{\to} F(t)$ and $F_n(t-) \overset{\text{a.s.}}{\to} F(t-)$ for every t. Given a fixed $\epsilon > 0$, there exists a partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that $F(t_i-) - F(t_{i-1}) < \epsilon$ for every i. Now for $t_{i-1} \leq t < t_i$,

$$F_n(t) - F(t) \leq F_n(t_i-) - F(t_i-) + \epsilon$$

$$F_n(t) - F(t) \geq F_n(t_{i-1}) - F(t_{i-1}) + \epsilon$$

The convergence of $F_n(t)$ and $F_n(t-)$ for every fixed t is certainly uniform for t in the finite set $\{t_1, \ldots, t_{k-1}\}$. Conclude that $\limsup ||F_n - F||_\infty \leq \epsilon$, almost surely. This is true for all $\epsilon > 0$, so the limit superior is zero.

References

[1] A. A. Tsiatis, “Semiparametric theory and missing data,” 2006.

[2] K. Long. (2009) Math 5311 – gateaux differentials and frechet derivatives.

[3] B. Sen. (2018) Semiparametric statistics. [Online]. Available: http://www.stat.columbia.edu/ bodhi/Talks/SPThNotes.pdf

[4] P. J. Bickel and K. A. Doksum, *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.