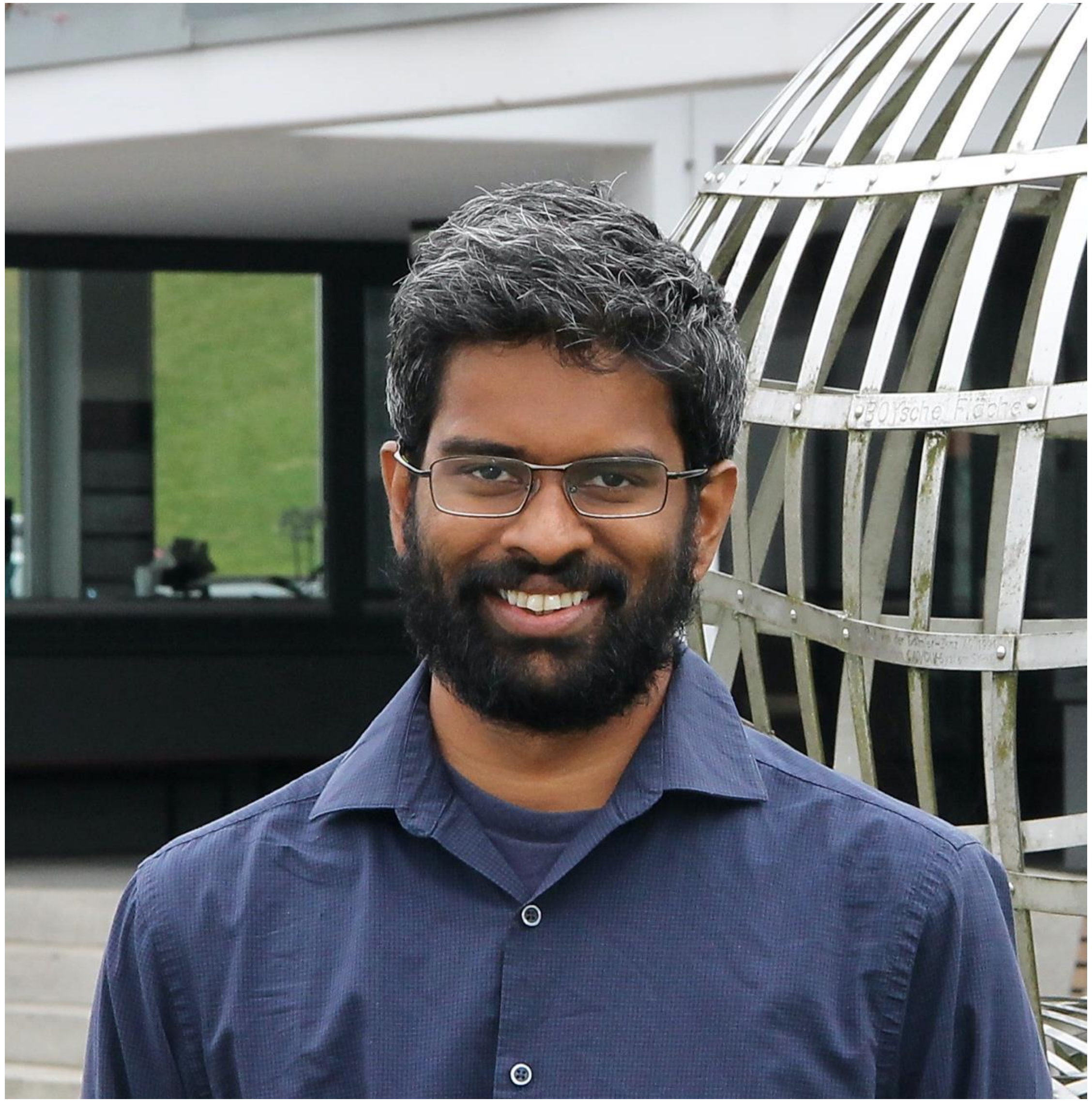
Neyman Orthogonality and Pathwise Differentiability

& Some Remarks on their Equivalence

Eric Chen
CMU Statistics

Carnegie Mellon University
Statistics & Data Science

Edward Kennedy



Sivaraman Balakrishnan

Overview

On the use of influence functions. A special case of our framework can be phrased in the language of classical semiparametric inference as follows: If the population risk functional is pathwise differentiable, and one estimates the target by minimizing an estimator for the risk based on influence functions, which will typically lead to a Neyman orthogonal loss and the resulting target estimator will have favorable second-order errors dependence on the error of the nuisance estimator; see Curth et al. (2020) for follow-up work which takes this approach explicitly.

Foster & Syrgkanis (2023)

Springer Series in Statistics

Mark J. van der Laan
James M. Robins

Unified Methods for Censored Longitudinal Data and Causality

Springer

Background

Parametric Submodels

Fix a measurable space $(\mathcal{Z}, \mathcal{A})$ and a σ-finite dominating measure ν. For $P \in \mathcal{P}$, write $p = dP/d\nu$. Fix $P_0 \in \mathcal{P}$ with density p_0.

A **regular parametric submodel** through P_0 is a family $\{P_t : t \in (-\epsilon, \epsilon)\} \subset \mathcal{P}$ with $P_{t=0} = P_0$, $P_t \ll \nu$, such that there exists $s \in L_2^0(P_0)$ with

$$\int \left(\frac{\sqrt{p_t} - \sqrt{p_0}}{t} - \tfrac{1}{2} s \sqrt{p_0} \right)^2 d\nu \;\to\; 0 \quad \text{as } t \to 0.$$

We call s the **score** of the submodel at 0.

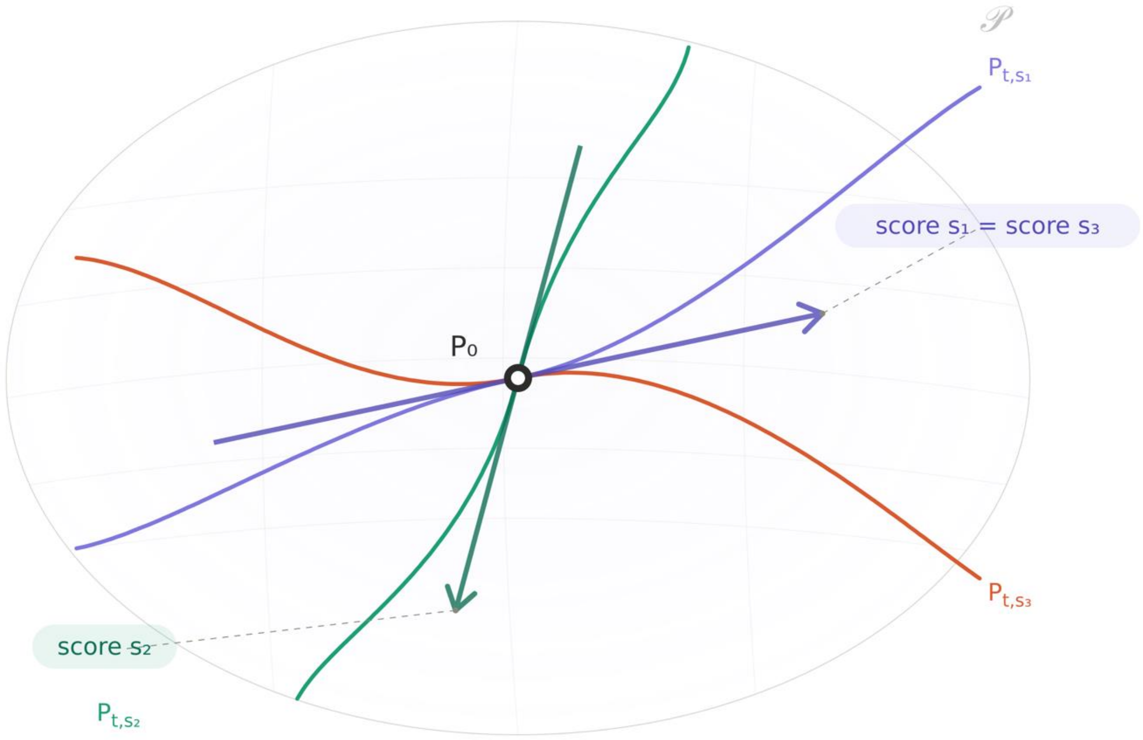
Parametric Submodels

A submodel is…

- a tool to analyze the local geometry of the model
- a first-order characterization
- a smoothness condition

We denote the submodel as $P_{t,s}$.

- two submodels can have the same score
- only first order behavior is important



Pathwise Differentiability and Influence Functions

Let $\mathcal{S} \subset L_2^0(P_0)$ be the set of scores of all regular submodels through P_0. The **tangent space** is

$$\mathcal{T} := \overline{\text{span}(\mathcal{S})}^{L_2(P_0)} \subseteq L_2^0(P_0).$$

Let $\beta : \mathcal{P} \to \mathbb{R}$ be the target parameter. We say β is **pathwise differentiable** at P_0 if there exists $\varphi \in L_2^0(P_0)$ such that for every regular submodel with score s,

$$\frac{d}{dt}\beta(P_{t,s})\bigg|_{t=0} = \mathbf{E}_0[\varphi(Z;\, P_0)\, s(Z)].$$

We call φ an **influence function** (or *gradient of the pathwise derivative*) of β at P_0.

Pathwise Differentiability and Influence Functions

Important: The pathwise derivative condition is verified by only taking inner product between the influence function and scores.

What this means: If you have an element orthogonal to the tangent space, you can add it to your existing φ to produce another influence function ☺

$$\varphi + h, \quad h \in \mathcal{T}^{\perp} \implies \mathbf{E}_0[(\varphi + h)s] = \mathbf{E}_0[\varphi s] = \frac{d}{dt}\beta(P_{t,s})\Big|_{t=0}$$

Efficiency: The projection of any influence function onto the tangent space is the efficient influence function.

Pathwise Differentiability and Influence Functions

Distinction: Verifying pathwise differentiability is (somewhat) different than finding the influence function

How this usually proceeds: Check the inner product condition holds for a dense class of submodels, then extend via a Hellinger Lipschitz argument

$$p_t(z) := p_0(z)\{1 + tg(z)\}, \quad g \in L_\infty(P_0)$$

It generally suffices to verify pathwise differentiability with **linear tilts**, since regular submodels are distinguished by first-order behavior.

Neyman Orthogonality

Let

$$\beta : \mathcal{P} \to \mathbb{R}, \qquad \eta : \mathcal{P} \to \mathcal{H}$$

be functionals on the model, with $\mathcal{H} \subset \mathcal{V}$ a subset of a normed space. Write $\beta_0 := \beta(P_0)$, $\eta_0 := \eta(P_0)$. Let

$$\dot{\mathcal{H}} := \{h \in \mathcal{V} : \exists\, \epsilon > 0 \text{ s.t. } \eta_0 + th \in \mathcal{H} \;\; \forall\, |t| < \epsilon\}$$

be the admissible perturbation directions at η_0.

An estimating function m is **Neyman orthogonal** at (β_0, η_0) if $\eta \mapsto \mathbf{E}_0[m(Z; \beta_0, \eta)]$ is Gâteaux differentiable at η_0 and

$$\frac{\partial}{\partial \eta} \mathbf{E}_0[m(Z; \beta_0, \eta)]\Big|_{\eta=\eta_0} [h] = 0 \qquad \forall\, h \in \dot{\mathcal{H}}.$$

Equivalence

Forward Direction

Theorem (Informal). Under some verifiable regularity conditions broadly amounting to…

- Being able to use chain rules
- Mean-zero & Neyman Orthogonality
- Hellinger Lipschitz condition on the functional

a Neyman orthogonal estimating function yields an influence function.

General Idea: Neyman orthogonality helps kill the nuisance term for the restricted class of submodels, then extension via Hellinger Lipschitz argument

Forward Direction

Smoothness. Exists a set of scores $\mathcal{S} \in L_\infty(P_0) \cap L_2^0(P_0)$ dense in $\mathcal{T}$ such that for each $s \in \mathcal{S}$, there exists a regular submodel $t \mapsto P_{t,s}$ along which $\beta(P_{t,s})$ and $\eta(P_{t,s})$ is differentiable at $t = 0$.

Estimating function. The estimating function m is mean-zero and Neyman orthogonal.

Nondegenerate Jacobian. $\mathbf{E}_0[\partial_\beta m(Z; \beta_0, \eta_0)] \neq 0$.

Hellinger Lipschitz. There exist $c, \delta > 0$ such that

$$|\beta(P_1) - \beta(P_2)| \leq cH(P_1, P_2) \quad \forall P_1, P_2 \in \mathcal{P} \text{ with } H(P_i, P_0) \leq \delta.$$

etc. boundedness, being able to take derivatives, and interchange with expectation ...

Forward Direction

Proof sketch.

Choose $s \in \mathcal{S}$ and submodel $t \mapsto P_{t,s}$. For sufficiently small t, we have

$$\mathbf{E}_{P_{t,s}}[m(Z;\beta_{t,s},\eta_{t,s})] = 0.$$

Let $f_{t,s}(Z) := m(Z;\beta_{t,s},\eta_{t,s})$. We can write

$$\begin{aligned}
0 &= \frac{d}{dt}\mathbf{E}_{P_{t,s}}[f_{t,s}]|_{t=0} \\
&= \mathbf{E}_0[f_0(Z)s(Z)] + \mathbf{E}_0[\dot{f}_{0,s}(Z)] \\
&= \mathbf{E}_0[m(Z;\beta_0,\eta_0)s(Z)] + \mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]\dot{\beta}_{0,s} + \mathbf{E}_0[\partial_\eta m(Z;\beta_0,\eta_0)][\dot{\eta}_{0,s}]] \\
&= \mathbf{E}_0[m(Z;\beta_0,\eta_0)s(Z)] + \mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]\dot{\beta}_{0,s} + \partial_\eta \mathbf{E}_0[m(Z;\beta_0,\eta)]|_{\eta=\eta_0}[\dot{\eta}_{0,s}]
\end{aligned}$$

Forward Direction

Master equation. We will use this equation in both directions to investigate the pathwise behavior induced by the two conditions.

$$0 = \mathbf{E}_0[m(Z;\beta_0,\eta_0)s(Z)] + \mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]\dot{\beta}_{0,s} + \partial_\eta \mathbf{E}_0[m(Z;\beta_0,\eta)]|_{\eta=\eta_0}[\dot{\eta}_{0,s}]$$

Neyman orthogonality ensures the third term is zero.

$$0 = \mathbf{E}_0[m(Z;\beta_0,\eta_0)s(Z)] + \mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]\dot{\beta}_{0,s}$$

$$\dot{\beta}_{0,s} = \mathbf{E}[\varphi(Z)s(Z)] \quad \text{where } \varphi = -\mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]^{-1} m(Z;\beta_0,\eta_0)$$

Forward Direction

What we've shown. For a dense class of submodels, the inner product condition holds with our choice of the influence function.

Extend this to all regular submodels. We adapt an argument of Luedtke and Chung (2024).

At a high-level, **Hellinger Lipschitz** controls the first order "difference" between two arbitrary regular submodels.

Forward Direction

Now let $t \mapsto P_{t,s'}$ be any arbitrary submodel. We choose another submodel $t \mapsto P_{t,g}$ with $g \in \mathcal{S}$.

$$\left| \frac{\beta(P_{t,s'}) - \beta_0}{t} - \mathbf{E}_0[\varphi s'] \right|$$
$$\leq \underbrace{\frac{|\beta(P_{t,s'}) - \beta(P_{t,g})|}{|t|}}_{\text{(I)}} + \underbrace{\left| \frac{\beta(P_{t,g}) - \beta_0}{t} - \mathbf{E}_0[\varphi g] \right|}_{\text{(II)}} + \underbrace{|\mathbf{E}_0[\varphi(s' - g)]|}_{\text{(III)}}$$

Next, since $\mathcal{S}$ is dense in $\mathcal{T}$ by assumption, we can pick g to satisfy $\|s' - g\|_{L_2(P_0)} \leq \epsilon$ for any arbitrary $\epsilon > 0$.

Forward Direction

We now take the limit.

Term I.

Lemma. For the aforementioned submodels,

$$\limsup_{t\to 0} \frac{|\beta(P_{t,s'}) - \beta(P_{t,g})|}{|t|} \leq c \cdot \limsup_{t\to 0} \frac{H(P_{t,s'}, P_{t,g})}{|t|} \leq \frac{c}{2\sqrt{2}} \|s' - g\|_{L_2(P_0)} \leq \frac{c\epsilon}{2\sqrt{2}}.$$

Term II. 0 since we showed pathwise differentiability holds for the dense class.

Term III.

$$|\mathbf{E}_0[\varphi(s' - g)]| \leq \|\varphi\|_{L_2(P_0)} \cdot \|s' - g\|_{L_2(P_0)} \leq \|\varphi\|_{L_2(P_0)} \cdot \epsilon.$$

Forward Direction

What you would typically do...

1. Verify existence of derivatives with a linear tilt submodel
2. Check Hellinger Lipschitz
3. Check remaining conditions

How does this show any equivalence?

$$\dot{\beta}_{0,s} = \mathbf{E}[\varphi(Z)s(Z)] \quad \text{where } \varphi = -\mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]^{-1} m(Z;\beta_0,\eta_0)$$

Can check that $-\mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]^{-1}$ is usually **1** for many functionals (ATE, ATT...)

Reverse Direction

Theorem (Informal). Under a **local product structure** assumption and verifiable regularity conditions , an influence function is Neyman orthogonal

These conditions broadly amount to…

- The functional and nuisance can move "independently"
- Regularity along submodels where this happens
- Pathwise differentiability of the functional with influence function *m*

Intuitively: Influence functions are Neyman orthogonal whenever a nuisance is truly a "nuisance," i.e., it doesn't tell you everything about the functional.

Reverse Direction

Recall

$$\dot{\mathcal{H}} := \{h \in \mathcal{V} : \exists\, \epsilon > 0 \text{ s.t. } \eta_0 + th \in \mathcal{H} \;\; \forall\, |t| < \epsilon\}$$

is the set of all admissible perturbation directions at η_0.

Local product structure. The following conditions hold:

1. β-**coordinate submodel.** There exists a regular (QMD) submodel $t \mapsto P_t \in \mathcal{P}$ with

$$\frac{d}{dt}\beta(P_t)\Big|_{t=0} = 1 \qquad \text{and} \qquad \frac{d}{dt}\eta(P_t)\Big|_{t=0} = 0.$$

2. η-**coordinate submodel.** For every admissible nuisance perturbation direction $h \in \dot{\mathcal{H}}$, there exists a regular (QMD) submodel $t \mapsto P_t \in \mathcal{P}$ with

$$\frac{d}{dt}\beta(P_t)\Big|_{t=0} = 0 \qquad \text{and} \qquad \frac{d}{dt}\eta(P_t)\Big|_{t=0} = h.$$

Reverse Direction

Proof sketch.

Recall the master equation (we can show it holds here too for submodels that satisfy local product structure)

$$0 = \mathbf{E}_0[m(Z;\beta_0,\eta_0)s(Z)] + \mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]\dot{\beta}_{0,s} + \partial_\eta \mathbf{E}_0[m(Z;\beta_0,\eta)]|_{\eta=\eta_0}[\dot{\eta}_{0,s}]$$

$$= \dot{\beta}_{0,s} + \mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]\dot{\beta}_{0,s} + \partial_\eta \mathbf{E}[m(Z;\beta_0,\eta)]|_{\eta=\eta_0}[\dot{\eta}_{0,s}]$$

where we assumed pathwise differentiability holds with $\varphi(Z) \equiv m(Z;\beta_0,\eta_0)$:

$$\mathbf{E}_0[m(Z;\beta_0,\eta_0)s(Z)] = \mathbf{E}_0[\varphi(Z)s(Z)] = \dot{\beta}_{0,s}.$$

Reverse Direction

$$0 = \dot{\beta}_{0,s} + \mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)]\dot{\beta}_{0,s} + \partial_\eta \mathbf{E}[m(Z;\beta_0,\eta)]|_{\eta=\eta_0}[\dot{\eta}_{0,s}]$$

Local Product Structure I

↓

Pick submodels where $\dot{\beta}_{0,s_\beta} = 1$ and $\dot{\eta}_{0,s_\beta} = 0$

↓

$$\mathbf{E}_0[\partial_\beta m(Z;\beta_0,\eta_0)] = -1$$

Local Product Structure II

↓

Pick submodels where $\dot{\beta}_{0,s_h} = 0$ and $\dot{\eta}_{0,s_h} = h$

↓

$$0 = \partial_\eta \mathbf{E}_0[m(Z;\beta_0,\eta)]|_{\eta=\eta_0}[h]$$

Reverse Direction

Checking local product structure is nontrivial, need to construct submodels that nicely balance interplay between functional and nuisance.

We can also say something interesting when local product structure **doesn't hold**, as in the case of the squared density.

$$\beta(P) = \int p^2 d\nu, \quad \eta = p$$

Conclusion

In general nonparametric settings, Neyman orthogonality and pathwise differentiability are equivalent given…

1) derivatives exists for a dense class of submodels
2) functional is Hellinger Lipschitz
3) local product structure holds
4) plus some additional mild regularity

Future work

Which assumptions can be relaxed? Are there weaker conditions that we can verify instead? Is there any equivalence in the conditional case?

Thank You!