Influence Functions for Nonparametric Parameters

Whitney K. Newey

ESIF Economics and AI+ML Meeting
August 13-14 2024

Ongoing joint work with Victor Chernozhukov and Vasilis Syrgkanis.

INTRODUCTION

There are many nonparametric, conditional parameters of interest.

An example is the conditional average treatment effect given one or more key covariates.

An economic example is average, welfare effects of increasing taxes conditional on income.

Conditional parameters can provide useful, low dimensional summaries of high dimensional unknown functions and may have causal interpretations.

We consider influence functions for conditional parameters.

Similarly to the classic influence function of a functional of a cumulative distribution function (CDF), conditional influence functions can be used to:

–Construct Neyman orthogonal estimating equations for conditional, nonparametric parameters.

–Quantify local effects of misspecification on nonparametric parameters.

–Define conditional, local policy effects.

Here we mostly focus on Neyman orthogonal estimating equations using machine learning but also touch on other uses for conditional influence functions.

Estimators of conditional parameters have previously been developed.

Convergence rates for conditional parameters given in Foster and Syrgkanis (2019).

Linear projection estimators in Semenova and Chernozhukov (2021).

Kernel estimators with automatic debiasing in Chernozhukov, Newey, and Singh (2022).

Estimators of the Conditional Average Treatment Effect in Kennedy et al. (2024).

Purpose here is to give conditional influence function and show how it can be used in these and other settings.

We describe conditional influence functions.

Show how they can be used to construct Neyman orthogonal estimating equations.

Briefly discuss other possible uses.

Describe estimators based on Neyman orthogonality.

CONDITIONAL INFLUENCE FUNCTIONS

Recall the classical influence function.

$$\begin{aligned}
W &: \text{ One data observation.} \\
F &: \text{ CDF of one observation having true value } F_0, \\
\theta(F) &: \text{ parameter of interest having true value } \theta_0 = \theta(F_0), \\
F_\varepsilon &= (1-\varepsilon)F_0 + \varepsilon H \text{ for CDF } H \text{ represents local variation.}
\end{aligned}$$

The influence function (when it exists) is $\psi_0(W)$ satisfying.

$$\frac{\partial \theta(F_\varepsilon)}{\partial \varepsilon} = \int \psi_0(W)dH, \;\; E[\psi_0(W)] = 0.$$

$\psi_0(W)$ is Gateaux derivative of $\theta(F)$ at F_0.

To describe a conditional influence function, let $T(W)$ be some random variable, and define

$$\begin{aligned} F^v &: \text{ CDF of one observation conditional on } T(W) = v, \\ &\quad \text{having true value } F_0^v; \\ \theta(F^v) &: \text{ parameter of interest, having true value } \theta_0(v) = \theta(F_0^v); \\ F_\varepsilon^v &= (1-\varepsilon)F_0^v + \varepsilon H^v \text{ for alternate conditional CDF } H^v. \end{aligned}$$

The conditional influence function (when it exists) is $\psi_0^v(W)$ satisfying

$$\frac{\partial \theta(F_\varepsilon^v)}{\partial \varepsilon} = \int \psi_0^v(W) dH^v, \; E[\psi_0^v(W) | T(W) = v] = 0.$$

Running Example: Conditional average potential outcome.

There is an outcome variable $Y = DY(1) + (1 - D)Y(0)$ where $D \in \{0, 1\}$ is treatment, $Y(1)$, $Y(0)$ are potential outcomes with and without treatment, and are mean independent of treatment conditional on covariates Z.

Let $V = t(Z)$ some measurable function of the covariates, $X = (D, Z)$, $W = (Y, X)$, and $\gamma_0(X) = E[Y|X]$.

A parameter of interest is

$$\theta_0(v) = E[Y(1)|T(Z) = v] = E[\gamma_0(1, Z)|T(Z) = v].$$

The conditional influence function of $\theta_0(V)$ is shown here to be

$$\gamma_0(1, Z) + \alpha_0(X)\{Y - \gamma_0(X)\} - \theta_0(V), \ \alpha_0(X) = \frac{D}{\Pr(D = 1|Z)}.$$

Conditional influence functions can be used to:

–Construct Neyman orthogonal estimating equations for conditional, nonparametric parameters.

–Quantify local effects of misspecification on nonparametric parameters.

–Define conditional, local policy effects.

It seems they cannot be used to characterize asymptotic variances or make efficiency comparisons for conditional parameters, when these are nonparametric, meaning conditional on continuous variables.

For example series and kernel estimators of conditional means have different asymptotic variances because they employ different localizations.

Neyman Orthogonal Estimating Equations

These are conditional estimating equations for parameters where the effect of nuisance functions on the conditional expectation is zero, to first order.

We construct these by starting with an identifying conditional moment equation.

Let γ denote a first step function that helps identify $\theta_0(V)$, e.g. $\gamma(X)$ is a possible $E[Y|X]$ as in running example.

Let $g(W,\gamma,\theta)$ be a function of data observation W and θ a possible $\theta_0(v)$ function.

Assume that $\theta_0(v)$ solves

$$0 = E[g(W,\gamma_0,\theta)|V = v].$$

In the running example

$$g(W,\gamma,\theta) = \gamma(1,Z) - \theta(V),$$

so $\theta_0(V) = E[\gamma_0(1,Z)|V]$.

We finish construction by adding to $g(W,\gamma,\theta)$ the conditional influence function $\phi(W,\gamma,\alpha,\theta)$ of $E[g(W,\gamma(F^v),\theta)|V=v]$, where $\gamma(F^v)$ is the nonparametric object that helps identify $\theta_0(v)$ and α is a first step function additional to γ.

That is, $\phi(W,\gamma,\alpha,\theta)$ satisfies, for $F^v_\varepsilon = (1-\varepsilon)F^v_0 + \varepsilon H^v$,

$$\frac{\partial}{\partial \varepsilon} E[g(W,\gamma(F^v_\varepsilon),\theta)|V=v] = \int \phi(w,\gamma_0,\alpha_0,\theta)dH^v, \; E[\phi(W,\gamma_0,\alpha_0,\theta)|V]=0.$$

A conditionally Neyman orthogonal estimating function is

$$\psi(W,\gamma,\alpha,\theta) = g(W,\gamma,\theta) + \phi(W,\gamma,\alpha,\theta).$$

In the running example we have $\phi(W,\gamma,\alpha,\theta) = \alpha(X)[Y-\gamma(X)]$, so that

$$\psi(W,\gamma,\alpha,\theta) = \gamma(1,Z) - \theta(V) + \alpha(X)[Y-\gamma(X)].$$

Here $\alpha_0(X) = D/\Pr(D=1|Z)$.

The function $\psi(W,\gamma,\alpha,\theta)$ has two orthogonality properties:

$$(1)\ \frac{\partial}{\partial\varepsilon}E[\psi(W,\gamma(F_\varepsilon^V),\alpha_0,\theta)|V]=0,\ \ (2)\ E[\phi(W,\gamma_0,\alpha,\theta)|V]=0,$$

for any α satisfying $\alpha=\alpha(F^v)$ for some F^v and $\gamma_0=\gamma((1-\varepsilon)F^v+\varepsilon F_0^v)$ for ε small enough.

Property 1 follows from differentiating the identity

$$E_{F_\varepsilon^v}[\phi(W,\gamma(F_\varepsilon^v),\alpha_0,\theta)|V=v]\equiv 0,$$

for any F_ε^v with $\alpha_0=\alpha(F_\varepsilon^v)$.

Property 2 follows from

$$\begin{aligned}
E[\phi(W,\gamma_0,\alpha,\theta)|V] &= \int \phi(W,\gamma_0,\alpha(F^v),\theta)dF_0^v \\
&= \frac{\partial}{\partial\varepsilon}E[g(W,\gamma(\bar{F}_\varepsilon^v),\alpha_0(F^v),\theta)|V] \\
&= \frac{\partial}{\partial\varepsilon}E[g(W,\gamma_0,\alpha_0(F^v),\theta)|V]=0,
\end{aligned}$$

where $\bar{F}_\varepsilon^v=(1-\varepsilon)F^v+\varepsilon F_0^v$.

The two orthogonality properties are:

$$(1)\ \frac{\partial}{\partial\varepsilon}E[\psi(W,\gamma(F_\varepsilon^V),\alpha_0,\theta)|V]=0,\ \ (2)\ E[\phi(W,\gamma_0,\alpha,\theta)|V]=0,$$

for any α.

Under regularity conditions (1) implies

$$\frac{\partial}{\partial\delta}E[\psi(W,\gamma_0+\delta(\gamma-\gamma_0),\alpha_0,\theta)|V]=0,$$

i.e. orthogonality (zero Gateaux derivative) with respect to γ.

Also have conditional double robustness, i.e.

$$0=E[\psi(W,\gamma,\alpha_0,\theta_0)|V]=E[\psi(W,\gamma_0,\alpha,\theta_0)|V],$$

if and only if $E[\psi(W,\gamma,\alpha_0,\theta_0)|V]$ is linear affine in γ.

In the running example,

$$\psi(W,\gamma,\alpha,\theta)=\gamma(1,Z)-\theta(V)+\alpha(X)[Y-\gamma(X)].$$

By iterated expectations for $\pi_0(Z)=\Pr(D=1|Z)$,

$$\begin{aligned}
E[\psi(W,\gamma,\alpha,\theta_0)|V] &= E[\gamma(1,Z)|V]-\theta_0(V) \\
&\quad +E[\alpha(X)\{\gamma_0(X)-\gamma(X)\}|V] \\
&= E[\alpha_0(X)\{\gamma(X)-\gamma_0(X)\}|V] \\
&\quad +E[\alpha(X)\{\gamma_0(X)-\gamma(X)\}|V] \\
&= -E[\{\alpha(X)-\alpha_0(X)\}\{\gamma(X)-\gamma_0(X)\}|V].
\end{aligned}$$

Another Example: (Average potential outcome for continuous treatment): $X = (D, Z)$, D continuously distributed, $Y = Y(D)$ for D independent of $Y(d)$ conditional on Z.

For $\gamma_0(X) = E[Y|X]$, $V = D$,

$$\theta_0(d) = \int \gamma_0(d, Z) f_Z(z) dz = E[\gamma_0(D, Z) \frac{f_D(D)}{f_{D|Z}(D|Z)} | D = d]$$

is average potential outcome.

Can construct Neyman orthogonal estimating equation, like Lee (2019), that accounts for presence of pdf's.

Other Uses of Conditional Influence Functions

Can be used to quantify the effect of misspecification on conditional parameters of interest.

Gateaux derivative

$$\frac{\partial\theta(F^v_\varepsilon)}{\partial\varepsilon} = \int \psi^v_0(W)dH^v$$

gives the local effect of changing the conditional distribution of a data observations away from F^v_0 in the direction $(1-\varepsilon)F^v_0 + \varepsilon H^v$.

Could be used to evaluate conditional versions of local misspecification effects like those in Andrews, Gentzkow and Shapiro (2017).

Alternatively if changes in F^v represent policy shifts then $\partial\theta(F^v_\varepsilon)/\partial\varepsilon$ can be interpreted as a local policy effect.

Estimation of Nonparametric Parameters

Generalize running example to

$$\theta_0(V) = E[m(W, \gamma_0)|V], \ \gamma_0(X) = E[Y|X], \ V = T(X).$$

In running example, $m(W, \gamma) = \gamma(1, Z)$, for $X = (D, Z)$.

Suppose there is $\alpha_0(X)$ such that $E[m(W, \gamma)|V] = E[\alpha_0(X)\gamma(X)|V]$.

This applies to running example and others, such as

$$\theta_0(V) = E[\frac{\partial}{\partial d}\gamma_0(D, Z)|V], \ V = T(Z).$$

In this setting the conditionally Neyman orthogonal moment function is

$$m(W, \gamma) - \theta(v) + \alpha(X)[Y - \gamma(X)].$$

To estimate $\theta_0(V)$ from observations $W_1, ..., W_n$ we can run a nonparametric regression of

$$\hat{S}_i = m(W_i, \hat{\gamma}) + \hat{\alpha}(X_i)[Y_i - \hat{\gamma}(X_i)]$$

on V_i at v.

For theoretical reasons use cross fitting where partition observation indices $i = 1, ..., n$ into distinct sets $I_\ell, (\ell = 1, ..., L)$, and let $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ be computed from observations not in I_ℓ.

Let $\hat{S}_{i\ell} = m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]$.

For example let $K(u)$ be a pdf, h a bandwidth, and r the dimension of v.

A kernel regression estimator is

$$\hat{\theta}(v) = \frac{\sum_{\ell=1}^{L} \sum_{i \in I_\ell} K((v - V_i)/h^r)\hat{S}_{i\ell}}{\sum_i K((v - V_i)/h^r)}.$$

$$\hat{\theta}(v) = \frac{\sum_{\ell=1}^{L} \sum_{i \in I_\ell} K((v - V_i)/h^r)\hat{S}_{i\ell}}{\sum_i K((v - V_i)/h^r)}.$$

For small dimensional v a locally linear regression is preferable.

Could also use high dimensional regression method.

Straightforward to give conditions for this estimator to have same limiting distribution as estimator whe $\hat{S}_{i\ell}$ is replaced by $S_i = m(W_i, \gamma_0) + \alpha_0(X_i)[Y_i - \gamma_0(X_i)]$.

This is possible because of Neyman orthogonality of the conditional influence function.

The conditions are stronger the larger is r, requiring faster mean square convergence rates for $\hat{\gamma}$ and $\hat{\alpha}$ when r is larger.

$$\hat{\theta}(v) = \frac{\sum_{\ell=1}^{L}\sum_{i\in I_\ell} K((v-V_i)/h^r)\hat{S}_{i\ell}}{\sum_i K((v-V_i)/h^r)},$$
$$\hat{S}_{i\ell} = m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)].$$

The $\hat{\alpha}_\ell(X_i)$ can be obtained by minimizing

$$\sum_{i\notin I_\ell}[-2m(W_i,\alpha) + \alpha(X_i)^2],$$

over some set of approximating functions, with penalty possibly added.

For linear combinations of basis functions and L_1 penalty this is automatic Lasso from Chernozhukov, Newey, and Singh (2022).

Can minimize over neural net and other specifications for α and generalize from functionals of conditional means to other conditional location functions, like quantiles; Chernozhukov, Newey, Quintas-Martinez, and Syrgkanis (2021).

These are automatic methods for constructing $\hat{\alpha}$, requiring only $m(W,\alpha)$ and not using a specific formula for α_0.

SUMMARY

We have considered the conditional influence function.

Showed how to construct conditionally Neyman orthogonal moment functions.

Touched on other uses of conditional influence functions.

Described estimators of conditional nonparametric parameters.